PSTAT 160B LECTURE NOTES - SPRING 2022 (LAST UPDATED - August 2, 2022)

ADAM WATERBURY

CONTENTS

1. Introduction

Recall that a stochastic process is a collection of random variables $\{X_t\}_{t\in\mathscr{I}}$, where $\mathscr{I}$ is some index set. We typically use stochastic processes to model situations where a quantity changes over time, so we generally encounter index sets of the form $\mathscr{I} = \mathbb{R} = (-\infty,\infty)$, $\mathscr{I} = \mathbb{R}_+ = [0,\infty)$, $\mathscr{I} = [0,T]$, or $\mathscr{I} = \mathbb{N}_0 = \{0,1,2,\dots\}$.

In PSTAT 160A we studied discrete-time stochastic processes; of particular interest were discrete-time Markov chains (DTMC) on finite or countable state spaces and random walks. Recall that a collection of random variables $\{X_n\}_{n=0}^{\infty}$ taking values in a countable (or finite) state space $\mathscr{S}$ is said to be a DTMC if there is some stochastic matrix P such that for all $x_0, x_1, \dots, x_n, x \in \mathscr{S}$,

$$\mathbb{P}[X_{n+1} = x | X_0 = x_0, X_1 = x_1, \dots, X_n = x_n] = \mathbb{P}[X_{n+1} = x | X_n = x_n] = P_{x_n,x}.$$

DTMCs are useful for modeling a wide range of random processes; for example, suppose that we are interested in modeling the price of a stock. One approach is to only consider the value of the stock at the beginning of each day, namely to define a DTMC $\{X_n\}_{n\in\mathbb{N}_0}$ by

$$X_n \doteq \text{"value of the stock at the start of the } n^{\text{th}} \text{ day."}$$

While this approach might be useful for some situations, it fails to fully describe the behavior of the stock's price over the course of each day, as it only captures a snapshot of the price at particular time instants. If we are also interested in modeling the evolution of the stock price throughout a day, then we would instead need to consider a continuous-time stochastic process. In this case we might model the stock's price as a stochastic process $\{X_t\}_{0\le t\le 24}$, where

$$X_t \doteq \text{"value of the stock at } t \text{ hours."}$$

Note that in this case the stochastic process $\{X_t\}_{0\le t\le 24} = \{X_t : t \in [0,24]\}$ is an *uncountable* collection of random variables. Additionally, it (conceptually) makes sense to model a stock's price at a given time instant as a continuous random variable, meaning that for each $t \in [0,24]$, X_t is a continuous random variable with state space $\mathscr{S} = [0,\infty)$. Therefore, to accurately model the evolution of a stock's price, we will need to understand continuous-time stochastic processes on uncountable (continuous) state spaces.

For another example of a continuous-time stochastic process, suppose that we are interested in modeling the number of customers that visit a store over the course of a day. Customers can arrive at any time instant, so it would make sense to model the process as a continuous-time stochastic process $\{N_t\}_{0\le t\le 24}$, where time is measured in hours, and

$$N_t \doteq \text{"number of customers that have visited the store by } t \text{ hours."}$$

Note that while time is indexed by a continuous parameter, for each $t \in [0,24]$, N_t is a discrete random variable with state space $\mathscr{S} = \mathbb{N}_0 = \{1,2,3,\dots\}$. Accordingly, to accurately model the number of customers visiting a store over the course of a day, we will need to understand continuous-time stochastic processes on countable (discrete) state spaces.

In the first part of this course we will study continuous-time stochastic processes on discrete state spaces such as the process $\{N_t\}$ described above. These processes will fall largely into two main categories; Poisson processes (see Section 2 and continuous-time Markov chains (see Section 3). We will then move our attention towards continuous-time stochastic processes on continuous state spaces. The first – and most important – example we will see of such a process is Brownian motion (see Section 4), which is the continuous-time analogue of the simple symmetric random walk. After studying some fundamental properties of Brownian motion, we will briefly introduce martingales (see Section 5), before finishing with an introduction to stochastic calculus, which will give us the tools to model processes such as the stock price $\{X_t\}$ described above.

2. Poisson Processes

Consider a sequence of events that occur one after another at random times, starting at time $t = 0$. For example, the first event might be that the first customer arrives at a store, the second event might be that the second customer arrives at a store, and so on. Consider a stochastic process $\{N_t\}_{t\geq 0}$ that describes the number of events that have occurred up to time t, namely, suppose that

$$N_t \doteq \text{"the number of customers that have arrived up to time } t\text{"}, \quad t \geq 0.$$

The stochastic process $\{N_t\}_{t\geq 0}$ is known as a counting process.

Definition 2.1. *A **counting process** $\{N_t\}_{t\geq 0}$ is a collection of non-negative, integer-valued random variables such that if $0 \leq s \leq t$, then $N_s \leq N_t$.*

The counting processes that will be of particular interest to us are known as Poisson processes, but before we introduce Poisson processes, it will be helpful to review some of the important properties of the exponential distribution.

2.1. Exponential Distribution.

We begin by reviewing some of the important properties of the exponential distribution because exponentially-distributed random variables are the foundational building blocks for a wide range of continuous-time stochastic processes.

Definition 2.2. *A non-negative random variable X is said to follow an exponential distribution with rate $\lambda > 0$ if its CDF is given by*

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}.$$

We write $X \sim Exp(\lambda)$.

Recall that the density (i.e., probability density function) of a continuous random variable can be obtained by differentiating its CDF. If $X \sim \text{Exp}(\lambda)$, this tells us that the density of X is given by

$$f_X(x) = \frac{d}{dx}F_X(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0. \end{cases}$$

Additionally, the moment generating function – or MGF – of X is given by

$$m_X(t) \doteq \mathbb{E}\left[e^{tX}\right] = \int_0^\infty e^{tx} f_X(x)dx = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

Using the MGF, we can calculate the first and second moments of X by noting that

$$\mathbb{E}[X] = \frac{d}{dt}m_X(t)\Big|_{t=0} = \frac{1}{\lambda},$$

and

$$\mathbb{E}[X^2] = \frac{d^2}{dt^2}m_X(t)\Big|_{t=0} = \frac{2}{\lambda^2}.$$

2.1.1. *Memoryless Property.*

The following property will be useful when studying continuous-time stochastic processes on discrete state spaces.

Definition 2.3. *A non-negative random variable X is* ***memoryless*** *if*

$$\mathbb{P}(X > s+t | X > s) = \mathbb{P}(X > t), \quad s, t \geq 0.$$

If a memoryless random variable X represents the time a lightbulb will last before failing, the memoryless property says that if the lightbulb has already lasted for s years, then the probability that it lasts an additional t years is the same as the probability that a brand new lightbulb would last for t years. The following tells us that exponentially distributed random variables are the only continuous random variables with the memoryless property.

Proposition 2.4. *Let X be a continuous non-negative random variable. Then X has the memoryless property if and only if $X \sim Exp(\lambda)$ for some $\lambda > 0$.*

Proof. Suppose that X has the memoryless property. Let

$$F^c(t) \doteq \mathbb{P}(X > t), \quad t \geq 0.$$

denote the *complementary* CDF of X, and observe that for all $s, t \geq 0$,

$$\mathbb{P}(X > t) = \mathbb{P}(X > s+t | X > s) = \frac{\mathbb{P}(X > s+t \text{ and } X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)},$$

which tells us that

$$F^c(t)F^c(s) = \mathbb{P}(X > t)\mathbb{P}(X > s) = \mathbb{P}(X > s+t) = F^c(s+t). \tag{1}$$

If $\frac{m}{n} \in \mathbb{Q}_+$ is a non-negative rational number, then (1) tells us that

$$F^c\left(\frac{m}{n}\right) = F^c\left(\sum_{i=1}^{m} \frac{1}{n}\right) = \left(F^c\left(\frac{1}{n}\right)\right)^m, \tag{2}$$

and

$$F^c(1) = F^c\left(\sum_{i=1}^{n} \frac{1}{n}\right) = \left(F^c\left(\frac{1}{n}\right)\right)^n, \tag{3}$$

so

$$F^c\left(\frac{1}{n}\right) = \left(F^c(1)\right)^{\frac{1}{n}}. \tag{4}$$

Combining (2), (3), and (4), we see that for all rational numbers $q \in \mathbb{Q}_+$,

$$F^c(q) = \left(F^c(1)\right)^q = e^{\log(F^c(1))q}.$$

Since F^c is the complementary CDF of X, it is a non-increasing function. Now, fix $t \in \mathbb{R}_+$, and let $\{t_n\}_{n=1}^{\infty}$ and $\{s_n\}_{n=1}^{\infty}$ be two sequence of rational numbers such that $t_n \downarrow t$ and $s_n \uparrow t$. Then, using the monotonicity of F^c, we can see that for each $n \in \mathbb{N}$, if we let $\lambda \doteq -\log(F^c(1))$, then

$$e^{-\lambda t_n} \leq F^c(t) \leq e^{-\lambda s_n}.$$

Additionally, as $n \to \infty$, $e^{-\lambda t_n} \to e^{-\lambda t}$ and $e^{-\lambda s_n} \to e^{-\lambda t}$, so it follows that

$$F^c(t) = e^{-\lambda t},$$

which says that $X \sim \text{Exp}(\lambda)$.

Now suppose that $X \sim \text{Exp}(\lambda)$ for some $\lambda > 0$. Then

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

so the memoryless property holds. □

We just saw that the only non-negative continuous random variables with the memoryless property are those which are exponentially distributed. A similar result holds for discrete random variables as well, and proving it is a good exercise.

Proposition 2.5. *Let X be a discrete non-negative random variable. Then X has the memoryless property if and only if $X \sim$ Geometric(p) for some $p \in [0,1]$.*

The following example illustrates the memoryless property.

Example 2.6. *Suppose that lifespan of a machine component follows is exponentially distributed with a rate of $\lambda = \frac{1}{2}$. If the machine has already lasted for 3 years, what is the probability that it lasts for at least 4 years?*

Denote the lifetime of the component by X. Then $X \sim Exp\left(\frac{1}{2}\right)$, and the memoryless property tells us that

$$\mathbb{P}(X > 4 | X > 3) = \mathbb{P}(X > 1) = e^{-\frac{1}{2}}.$$

2.1.2. *Minimum of Exponentials.* The following example illustrates an interesting property of the exponential distribution.

Example 2.7. *Helen the cat and Teddy the cat are each served a bowl of food. If we let T denote the time that Teddy takes to finish eating his bowl of food, then $T \sim Exp(\lambda)$, and if we let H denote the time that Helen takes to finish eating her bowl of food, then $H \sim Exp(\mu)$, where $\lambda = \frac{1}{2}$ and $\mu = 1$. What is the probability that Teddy finishes eating his bowl of food first?*

The probability that Teddy finishes first is given by

$$\begin{aligned}
\mathbb{P}(T < H) &= \int_0^\infty \mathbb{P}(T < H | T = t) \lambda e^{-\lambda t} dt \\
&= \lambda \int_0^\infty \mathbb{P}(t < H) e^{-\lambda t} dt \\
&= \lambda \int_0^\infty e^{-\mu t} e^{-\lambda t} dt \\
&= \lambda \int_0^\infty e^{-(\mu+\lambda)t} dt \\
&= \frac{\lambda}{\mu + \lambda} \\
&= \frac{1}{3}.
\end{aligned}$$

Now suppose that a machine has n components, and that we are interested in understanding the probability distribution of the time that it takes for any of the components to fail. The following result is in fact slightly stronger, as it characterizes the joint probability distribution of time that it takes for any of the components to fail and which component fails. Recall that for an event A, the indicator function of A is the function 1_A given by

$$1_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Theorem 2.8. *For $i \in \{1,2,\dots,n\}$, let $X_i \sim Exp(\lambda_i)$, where $\lambda_i > 0$. Let*

$$M = \min\{X_1, X_2, \dots, X_n\},$$

and let

$$N \doteq \sum_{i=1}^{n} i 1_{\{M=X_i\}},$$

so that $N = i$ if $M = X_i$. Then

$$\mathbb{P}(M > x, N = i) = \frac{\lambda_i}{\lambda} e^{-\lambda x}, \quad x \geq 0, i \in \{1,2,\dots,n\},$$

where

$$\lambda \doteq \sum_{i=1}^{n} \lambda_i.$$

Proof. For notational simplicity, we prove the result for the case when $i = 1$, as the proof for the other cases is similar. Fix $x \geq 0$ and note that

$$\begin{aligned}
\mathbb{P}(M > x, N = 1) &= \mathbb{P}(X_j > X_1 > x \text{ for all } j \in \{2,\dots,n\}) \\
&= \int_0^\infty \mathbb{P}(X_j > X_1 > x \text{ for all } j \in \{2,\dots,n\} | X_1 = y) \lambda_1 e^{-\lambda_1 y} dy \\
&= \lambda_1 \int_x^\infty \mathbb{P}(X_j > X_1 \text{ for all } j \in \{2,\dots,n\} | X_1 = y) e^{-\lambda_1 y} dy \\
&= \lambda_1 \int_x^\infty \mathbb{P}(X_j > y \text{ for all } j \in \{2,\dots,n\}) e^{-\lambda_1 y} dy \\
&= \lambda_1 \int_x^\infty \left(\prod_{j=2}^{n} \mathbb{P}(X_j > y) \right) e^{-\lambda_1 y} dy \\
&= \lambda_1 \int_x^\infty e^{-\sum_{j=2}^{n} \lambda_j y} e^{-\lambda_1 y} dy \\
&= \lambda_1 \int_x^\infty e^{-\lambda y} dy \\
&= \frac{\lambda_1}{\lambda} e^{-\lambda x}.
\end{aligned}$$

□

The following corollary shows that M and N are independent and identifies their marginal probability distributions.

Corollary 2.9. *Let M and N be as in Theorem 2.8. Then M and N are independent, and $M \sim Exp(\lambda)$, and*

$$\mathbb{P}(N = i) = \frac{\lambda_i}{\lambda}, \quad i \in \{1,2,\dots,n\}.$$

Proof. The CDF of M is given by

$$F_M(x) = \mathbb{P}(M \leq x) = 1 - \mathbb{P}(M > x) = 1 - \sum_{i=1}^{n} \mathbb{P}(M > x, N = i) = 1 - \sum_{i=1}^{n} \frac{\lambda_1}{\lambda} e^{-\lambda x} = 1 - e^{-\lambda x},$$

so $M \sim \text{Exp}(\lambda)$, as claimed. Similarly,

$$\mathbb{P}(N = i) = \mathbb{P}(M > 0, N = i) = \frac{\lambda_i}{\lambda},$$

and the independence of M and N follows from the fact that for all $x \geq 0$ and $i \in \{1, 2, \ldots, n\}$,

$$\mathbb{P}(M > x, N = i) = \mathbb{P}(M > x)\mathbb{P}(N = i).$$

□

The following example illustrates how we can apply Theorem 2.8.

Example 2.10. *A petri dish contains 100 bacteria. Each bacteria has an Exp*$\left(\frac{1}{300}\right)$ *lifespan, where the rate is measure in minutes, and each of their lifespans is independent of the others'. What is the probability that none of the bacteria die within the first 10 minutes? How long do we expect it to take for the first bacteria to die?*

If we let X_i denote the lifespan of bacteria i for $i \in \{1, 2, \ldots, 100\}$, and let $M = \min\{X_1, X_2, \ldots, X_{100}\}$, then Corollary 2.9 tells us that $M \sim$ Exp(λ), where

$$\lambda \doteq \sum_{i=1}^{100} \frac{1}{300} = \frac{1}{3},$$

and the probability that all of the bacteria survive for more than 10 minutes is given by

$$\mathbb{P}(M > 10) = e^{-\frac{1}{3}\cdot 10} = e^{-\frac{10}{3}} \approx 0.03567.$$

Additionally, we expect the first bacteria to die after

$$\mathbb{E}(M) = \frac{1}{\lambda} = 3$$

minutes.

Example 2.11. *Consider the setting from Example 2.7, where the time that it took Teddy to finish eating followed an Exp(λ_T) distribution, and the time that it took Helen to finish eating followed an Exp(λ_H) distribution, where $\lambda_T = \frac{1}{2}$ and $\lambda_H = 1$. If we write T and H to denote the time that it took Teddy and Helen to finish eating, respectively, then through direct calculation we saw that probability that Teddy finished eating first was*

$$\mathbb{P}(\textit{Teddy finishes first}) = \mathbb{P}(T < H) = \frac{1}{3}.$$

We can recover this result using Corollary 2.9 by letting $M = \min\{T, H\}$ and

$$N = \begin{cases} t & \text{if } M = T \\ h & \text{if } M = H, \end{cases}$$

Then with $\lambda = \lambda_T + \lambda_H$, we have

$$\mathbb{P}(\textit{Teddy finishes first}) = \mathbb{P}(N = t) = \frac{\lambda_T}{\lambda} = \frac{1}{3}.$$

2.1.3. *Sum of Exponentials.* An artist wants to model how long it will take her to complete 7 paintings. She believes that she can model the time that it takes to finish each painting as an exponential distribution with a rate of $\lambda = \frac{1}{10}$. What is the probability distribution of the total time that it will take her to complete the 7 paintings?

Definition 2.12. *For $\beta > 0$ and $n \in \mathbb{N}$, a random variable Y is said to follow a Gamma(n, β) distribution if the density of Y is given by*

$$f_Y(x) = \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}, \quad x \geq 0.$$

The CDF of Y is given by

$$F_Y(x) = 1 - e^{-\lambda x} \sum_{r=0}^{n-1} \frac{(\lambda x)^r}{r!}, \quad z \geq 0.$$

The following proposition gives the MGF of the gamma distribution.

Proposition 2.13. *Let $Y \sim$ Gamma(n, β) for some $n \in \mathbb{N}$ and $\beta > 0$. The MGF of Y is given by*

$$m_Y(t) = \left(\frac{\beta}{\beta - t}\right)^n, \quad t < \beta.$$

Proof. For $t < \beta$, we have, by taking $u = (\beta - t)x$ and using the definition of the gamma function

$$\begin{aligned} M_Y(t) &= \int_0^\infty e^{tx} \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x} dx \\ &= \frac{\beta^n}{(n-1)!} \int_0^\infty e^{(t-\beta)x} x^{n-1} dx \\ &= \frac{\beta^n}{(n-1)!(\beta - t)^n} \int_0^\infty e^{-u} u^{n-1} du \\ &= \frac{\beta^n}{(\beta - t)^n}, \end{aligned}$$

so the result follows. □

Using Proposition 2.13 we obtain the following result.

Proposition 2.14. *Suppose $X_1, X_2, \ldots, X_n$ are iid Exp(λ) random variables and let $S \doteq \sum_{i=1}^n X_i$. Then $S \sim$ Gamma(n, λ).*

Proof. Recall that the MGF of each X_i is given by

$$m_{X_i}(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda,$$

so the MGF of S is given by

$$m_S(t) = \prod_{i=1}^n m_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n.$$

so the result follows on noting that MGFs uniquely characterize probability distributions. □

We now return to the example of the artist.

Example 2.15. *The time that it takes for an artist to complete a painting follows an exponential distribution with a rate of* $\lambda = \frac{1}{10}$. *If the time that it takes to complete each painting is independent of the others, what is the probability that it will take her at least 100 hours to complete 7 paintings?*

Let X_i *denote the time to complete painting* i, *for* $i \in \{1, \dots, 7\}$, *and note that the time to complete all paintings is given by* $S = \sum_{i=1}^{7} X_i$. *Using Proposition* 2.14, *we see that* $S \sim Gamma\left(7, \frac{1}{10}\right)$, *and*

$$\mathbb{P}(S > 100) = \int_{100}^{\infty} \frac{1}{10^7 6!} x^6 e^{-\frac{x}{10}} \, dx \approx 0.13014.$$

2.2. **Introduction to Poisson Processes.** We now introduce the notion of a Poisson process. These processes are useful for modeling the number of events that have occurred up to a point, such as hospital admissions, earthquakes, or phone calls received. We begin by introducing the Poisson process in terms of the number of events that have occurred, but we will see several other equivalent (and equally insightful) definitions as well.

Definition 2.16. *Let* $\{X_n\}$ *be a sequence of iid Exp*(λ) *random variables, and define*

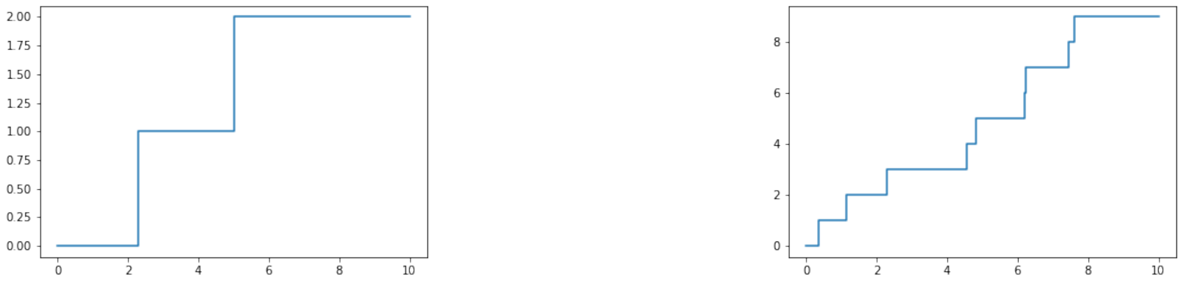
$$S_0 \doteq 0, \quad S_n \doteq \sum_{i=1}^{n} X_i.$$

The counting process $\{N_t\}_{t \geq 0}$ *defined by*

$$N_t \doteq \max\{n \geq 0 : S_n \leq t\},$$

is called a ***Poisson process*** *with rate* λ. *For convenience, we often write* $\{N_t\}_{t \geq 0} \sim PP(\lambda)$. *We often refer to* $\{X_n\}$ *as the* ***inter-arrival times***, *and* $\{S_n\}$ *as the* ***arrival times***.

In the definition of a Poisson process above, note that if each X_i represents the time that it takes for an event to occur, so if $N_t = k$, that means that up to time t, k events have occurred; in particular it tells us that over half-over time interval $(0, t]$, k events occurred. In Figure 4.6 we plot a realization of two Poisson processes over the time interval $[0, 10]$; the first has a rate of $\lambda_1 = 0.5$, and the second has a rate of $\lambda_2 = 1$.



(A) A realization of a Poisson process with rate $\lambda_1 = 0.5$ on the time interval $[0, 10]$.

(B) A realization of a Poisson process with rate $\lambda_2 = 1$ on the time interval $[0, 10]$.

FIGURE 2.1. Two realizations of the process $\{Y_t\}$ on different time intervals. Note the difference in the values on the y-axis.

The code used to generate these simulations is provided below.

```
import numpy as np
import matplotlib.pyplot as plt

def PoissonProcess(lam,T):
    timeCount = 0
    num = 0
    Nt = [0]
    times = [0]
    while(timeCount <= T):
        x = np.random.exponential(1/lam)
        timeCount = timeCount + x
        if(timeCount <= T):
            num = num + 1
            Nt.append(num)
            times.append(timeCount)
        if(timeCount > T):
            Nt.append(num)
            times.append(T)
    plt.plot(times, Nt, drawstyle='steps-post')
```

The following result described the marginal distribution of a Poisson process at each time instant and explains why they are known as such.

Proposition 2.17. *Let $\{N_t\}_{t\geq 0} \sim PP(\lambda)$ for some $\lambda > 0$. Then for each $t > 0$, $N_t \sim Poisson(\lambda t)$. In particular,*

$$\underline{\mathbb{P}(N_t = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad k \in \mathbb{N}_0.}$$

Proof. Using the notation from Definition 2.16, note that $N_t \geq k$ if and only if the k^{th} event has occurred by time t, meaning that

$$\mathbb{P}(N_t \geq k) = \mathbb{P}(\max\{n \geq 0 : S_n \leq t\} \geq k) = \mathbb{P}(S_k \leq t),$$

from which we can see that

$$\mathbb{P}(N_t = k) = \mathbb{P}(N_t \geq k) - \mathbb{P}(N_t \geq k+1) = \mathbb{P}(S_k \leq t) - \mathbb{P}(S_{k+1} \leq t).$$

Using Definition 2.12 and Proposition 2.14, we know that

$$\mathbb{P}(S_k \leq t) = 1 - e^{-\lambda t}\sum_{r=0}^{k-1}\frac{(\lambda t)^r}{r!}$$

and

$$\mathbb{P}(S_{k+1} \leq t) = 1 - e^{-\lambda t}\sum_{r=0}^{k}\frac{(\lambda t)^r}{r!},$$

so it follows that

$$\mathbb{P}(N_t = k) = \left(1 - e^{-\lambda t}\sum_{r=0}^{k-1}\frac{(\lambda t)^r}{r!}\right) - \left(1 - e^{-\lambda t}\sum_{r=0}^{k}\frac{(\lambda t)^r}{r!}\right) = e^{-\lambda t}\frac{(\lambda t)^k}{k!},$$

so $N_t \sim \text{Poisson}(\lambda t)$. □

Example 2.18. *Suppose that customers arrive at the bank according to a Poisson process with a rate of 9 per hour. Compute the distribution and expected value of the number of customers who enter the bank over the course of an 10-hour day.*

Denote the number of customers who enter the bank by time t by N_t. Then $\{N_t\}_{t\geq 0} \sim PP(9)$, and the number of customers who enter the bank over the course of the day is given by N_{10}. We know from Proposition 2.17 *that $N_{10} \sim$ Poisson$(10\cdot 9)$, so the expected number of customers who enter the bank over the course of the day is $\mathbb{E}[N_t] = 10\cdot 9 = 90$.*

Proposition 2.17 tells us the marginal distribution of the number of events in a Poisson process that occur by each time instant, but it does not tell us the joint distribution of the number of arrivals up to several different time instants. For example, consider the setting from Example 2.18. suppose that we wanted to calculate the probability that exactly 5 customers had shown up within the first 4 hours of being open and that exactly 7 customers had shown up by the end of the 10-hour day. That is to say, we are interested in calculating

$$\mathbb{P}(N_4 = 5, N_{10} = 7).$$

In order to calculate this quantity, we begin by introducing the notion of a shifted Poisson process.

Definition 2.19. *Let $\{N_t\}_{t\geq 0} \sim PP(\lambda)$. For a fixed $s \geq 0$, define the process $\{N_t^s\}_{t\geq 0}$ by*

$$N_t^s \doteq N_{t+s} - N_s, \quad t \geq 0.$$

*The process $\{N_t^s\}_{t\geq 0}$ is called a **shifted Poisson process**. Note that N_t^s denotes the number of events that occur in the t time-units after time s.*

Since N_s counts the number of events that occurred in the interval $(0, s]$, and N_{t+s} counts the number of events that occurred in the interval $(0, t+s]$, we can see that for a fixed $s \geq 0$, for each $t \geq 0$, N_t^s counts the number of events that occurred in the interval $(s, t+s]$. Observe that the time that it takes for each event to occur is independent of the time that it takes for the other events to occur and that the rate at which events occur does not vary over time. Together, these observations lead us to the following result.

Proposition 2.20. *Let $\{N_t\}_{t\geq 0}$ be a Poisson process with rate $\lambda > 0$. Then for each $s > 0$, the shifted Poisson process $\{N_t^s\}_{t\geq 0}$ is also a Poisson process with rate λ, and the process $\{N_t\}_{0\leq t\leq s}$ is independent of the process $\{N_t^s\}_{t\geq 0}$. Namely, for each $0 \leq t_1 \leq s$ and $t_2 \geq 0$, the random variables N_{t_1} and $N_{t_2}^s$ are independent.*

Proof. Exercise □

The following definitions will lead us to a second characterization of Poisson processes. Recall that for two random variables X and Y, we write $X \overset{d}{=} Y$ if X and Y have the same probability distribution.

Definition 2.21. *Let $\{X_t\}_{t\geq 0}$ be a stochastic process. For $t_1 < t_2$, the quantity $X_{t_2} - X_{t_1}$ is referred to as the increment of $\{X_t\}$ over the interval $(t_1, t_2]$. We say that $\{X_t\}$ has **stationary increments** if for all $s_1 < s_2$ and $t_1 < t_2$ such that $s_2 - s_1 = t_2 - t_1$, we have*

$$X_{t_2} - X_{t_1} \overset{d}{=} X_{s_2} - X_{s_1}.$$

We saw that $\{X_t\}$ has ***independent increments*** *if for all $t_1 < t_2 \leq t_3 < t_4$, the random variables $X_{t_2} - X_{t_1}$ and $X_{t_4} - X_{t_3}$ are independent.*

Intuitively, a process has stationary increments if the distribution of the change that it undergoes over a time interval depends only on the length of the interval. So, the total increase/decrease of the process over an interval $[0, t]$ has the same distribution as its increase/decrease over the interval $[s, s+t]$, since both intervals have a length of t. A process has independent increments if the amount that it increases/decreases over a time interval is independent of the amount the it increases/decreases on any other disjoint time interval.

It is very important to note that a process having stationary increments **does not** mean that it is identically distributed at all time instants; for example, we have seen that if $\{N_t\}_{t\geq 0} \sim PP(\lambda)$, then for each $t > 0$, $N_t \sim \text{Poisson}(\lambda t)$, so the distribution of the process at each time instant does vary over time. Similarly, the independent increment property **does not** imply that the process is independent at any two different time instants, only that the *change* that it undergoes over some time interval is independent of the change it undergoes over any other disjoint interval. For example, we have

$$\mathbb{P}(N_9 = 4 | N_5 = 7) = 0 \neq \mathbb{P}(N_9 = 4)\mathbb{P}(N_5 = 7),$$

so N_9 and N_5 are not independent.

We verify in Theorem 2.22 that Poisson processes have stationary and independent increments.

Theorem 2.22. *Let $\{N_t\}_{t\geq 0} \sim PP(\lambda)$. Then $\{N_t\}_{t\geq 0}$ has stationary and independent increments.*

Proof. From Proposition 2.17 and Proposition 2.20 we can see that if $t_1 < t_2$ and $s_1 < s_2$ are such that $u = t_2 - t_1 = s_2 - s_1$, then

$$\mathbb{P}(N_{t_2} - N_{t_1} = k) = \mathbb{P}(N_{u+t_1} - N_{t_1} = k) = \mathbb{P}(N_u^{t_1} = k) = e^{-\lambda u}\frac{(\lambda u)^k}{k!},$$

and, similarly,

$$\mathbb{P}(N_{s_2} - N_{s_1} = k) = e^{\lambda u}\frac{(\lambda u)^k}{k!},$$

so $N_{t_2} - N_{t_1} \overset{d}{=} N_{s_2} - N_{s_1}$, meaning that $\{N_t\}_{t\geq 0}$ has stationary increments. The independent increment property follows from Proposition 2.20. □

Using Theorem 2.22, we can derive a result which allows us to calculate quantities of the form

$$\mathbb{P}(N_s = k_1, N_t = k_2).$$

Proposition 2.23. *Let $\{N_t\}_{t\geq 0} \sim PP(\lambda)$. Then for $0 = t_0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$ and integers $0 = k_0 \leq k_1 \leq k_2 \leq \cdots \leq k_n$, we have*

$$\mathbb{P}(N_{t_1} = k_1, N_{t_2} = k_2, \ldots, N_{t_n} = k_n) = e^{-\lambda t_n}\prod_{i=1}^{n}\frac{(\lambda(t_i - t_{i-1}))^{k_i - k_{i-1}}}{(k_i - k_{i-1})!}$$

Proof. We prove the result inductively. First, note that when $n = 1$, we have

$$\mathbb{P}(N_{t_1} = k_1) = e^{-\lambda t_1}\frac{(\lambda t_1)^{k_1}}{k_1!} = e^{-\lambda t_1}\frac{(\lambda(t_1 - t_0))^{k_1 - k_0}}{(k_1 - k_0)!},$$

so the base case holds. Now, assuming that

$$\mathbb{P}(N_{t_1}=k_1,N_{t_2}=k_2,\dots,N_{t_{n-1}}=k_{n-1})=e^{-\lambda t_{n-1}}\prod_{i=1}^{n-1}\frac{(\lambda(t_i-t_{i-1}))^{k_i-k_{i-1}}}{(k_i-k_{i-1})!},$$

we have, from Theorem 2.22,

$$\begin{aligned}\mathbb{P}(N_{t_1}=k_1,\dots,N_{t_n}=k_n)&=\mathbb{P}(N_{t_n}=k_n|N_{t_1}=k_1,\dots,N_{t_{n-1}}=k_{n-1})\mathbb{P}(N_{t_1}=k_1,\dots,N_{t_{n-1}}=k_{n-1})\\&=\mathbb{P}(N_{t_n}-N_{t_{n-1}}=k_n-k_{n-1}|N_{t_1}=k_1,\dots,N_{t_{n-1}}=k_{n-1})\mathbb{P}(N_{t_1}=k_1,\dots,N_{t_{n-1}}=k_{n-1})\\&=e^{-\lambda(t_n-t_{n-1})}\frac{(\lambda(t_n-t_{n-1}))^{k_n-k_{n-1}}}{(k_n-k_{n-1})!}e^{-\lambda t_{n-1}}\prod_{i=1}^{n-1}\frac{(\lambda(t_i-t_{i-1}))^{k_i-k_{i-1}}}{(k_i-k_{i-1})!}\\&=e^{-\lambda t_n}\prod_{i=1}^{n}\frac{(\lambda(t_i-t_{i-1}))^{k_i-k_{i-1}}}{(k_i-k_{i-1})!},\end{aligned}$$

as claimed. □

The following example illustrates how we can apply the ideas from the proof of Proposition 2.23.

Example 2.24. *The number of retweets you get over the course of a day, starting at 8:00am, can be modeled as a Poisson process with a rate of $\lambda=2$ retweets per hour. What is the probability that at 9:00am you have had 3 retweets, at 10:30am you have had 4 retweets, and then between 11:00am and 12:00pm you get an additional 3 retweets?*

Denote the Poisson process by $\{N_t\}_{t\geq 0}$. Using the independent increment property and then the stationary increment property, we have

$$\begin{aligned}\mathbb{P}(N_1=3,N_{2.5}=4,N_4-N_3=3)&=\mathbb{P}(N_4-N_3=3,N_{2.5}-N_1=1,N_1-N_0=3)\\&=\mathbb{P}(N_4-N_3=3)\mathbb{P}(N_{2.5}-N_1=1)\mathbb{P}(N_1-N_0=3)\\&=\mathbb{P}(N_1=3)\mathbb{P}(N_{1.5}=1)\mathbb{P}(N_1=3)\\&=e^{-2\cdot 1}\frac{(2\cdot 1)^3}{3!}e^{-2\cdot 1.5}\frac{(2\cdot 1.5)^1}{1!}e^{-2\cdot 1}\frac{(2\cdot 1)^3}{3!}\\&=e^{-7}\frac{16}{3}\end{aligned}$$

The following example may be helpful as well.

Example 2.25. *Let $\{N_t\}$ be a Poisson process with rate $\lambda=5$. Calculate $\mathbb{P}(N_7=2,N_9=5,N_{11}-N_{9.5}=4)$.*

We have, once more using the independent and stationary increments properties,

$$\begin{aligned}\mathbb{P}(N_7=2,N_9=5,N_{11}-N_{9.5}=4)&=\mathbb{P}(N_7-N_0=2,N_9-N_7=3,N_{11}-N_{9.5}=4)\\&=\mathbb{P}(N_7-N_0=2)\mathbb{P}(N_9-N_7=3)\mathbb{P}(N_{11}-N_{9.5}=4)\\&=\mathbb{P}(N_7-N_0=2)\mathbb{P}(N_2-N_0=3)\mathbb{P}(N_{1.5}-N_0=4)\\&=\mathbb{P}(N_7=2)\mathbb{P}(N_2=3)\mathbb{P}(N_{1.5}=4)\\&=e^{-5\cdot 7}\frac{(5\cdot 7)^2}{2!}e^{-5\cdot 2}\frac{(5\cdot 2)^3}{3!}e^{-5\cdot 1.5}\frac{(5\cdot 1.5)^4}{4!}\end{aligned}$$

2.3. **Other Characterizations of Poisson Processes.**

In this section we introduce several other characterizations of Poisson processes.

2.3.1. *The Second Characterization.*

The following result provides a second characterization of the Poisson process.

Theorem 2.26. *A stochastic process $\{N_t\}_{t\geq 0}$ is a Poisson process with rate $\lambda > 0$ if and only if it has stationary and independent increments and, for each $t \geq 0$, $N_t \sim$ Poisson(λt).*

Proof. Suppose that $\{N_t\}_{t\geq 0}$ has stationary and independent increments and that, for each $t \geq 0$, $N_t \sim$ Poisson(λt). Consequently, $\mathbb{P}(N_0 = 0) = 1$. Since $\{N_t\}_{t\geq 0}$ has stationary increments, we know that for each $s, t \geq 0$, $N_{t+s} - N_s \sim$ Poisson(λt). Let $S_1 \doteq \inf\{t \geq 0 : N_t = 1\}$, and note from the Proof of Proposition 2.17 that

$$\mathbb{P}(S_1 \leq t) = \mathbb{P}(N_t \geq 1) = 1 - \mathbb{P}(N_t = 0) = 1 - e^{-\lambda t}\frac{(\lambda t)^0}{0!} = 1 - e^{-\lambda t},$$

so $S_1 \sim$ Exp(λ). Similarly, if we let $S_n \doteq \inf\{t \geq 0 : N_t = n\}$, then $S_n - S_{n-1} = \inf\{t \geq 0 : N_{S_{n-1}+t} - N_{S_{n-1}} = 1\}$, and so, using the independent increments property, for $s, t \geq 0$, with $\bar{t} \doteq t_1 + t_2 + \cdots + t_{n-1}$,

$$\begin{aligned}\mathbb{P}(S_n - S_{n-1} > t | S_1 - S_0 = t_1, \ldots, S_{n-1} - S_{n-2} = t_{n-1}) &= \mathbb{P}(N_{\bar{t}+t} - N_{\bar{t}} = 0 | S_1 - S_0 = t_1, \ldots, S_{n-1} - S_{n-2} = t_{n-1})\\ &= \mathbb{P}(N_{\bar{t}+t} - N_{\bar{t}} = 0)\\ &= \mathbb{P}(N_t) = 0\\ &= e^{-\lambda t},\end{aligned}$$

from which we obtain that $\{S_i - S_{i-1}\}_{i=1}^n$ is a sequence of iid Exp(λ) random variables, which tells us that $\{N_t\}_{t\geq 0}$ is a Poisson process with rate λ, as defined in Definition 2.16. Since we have already shown that a Poisson process has stationary and independent increments and the stated distributional property, the result follows. □

2.3.2. *The Third Characterization.*

Before introducing a third characterization of Poisson processes, we recall some notation regarding the rate at which functions grow or decay.

Definition 2.27. *Consider two functions $f, g : \mathbb{R} \to \mathbb{R}$. We say that $f(x) = o(g(x))$ (as $x \to 0$) if*

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = 0.$$

In particular, we saw that $f = o(x)$ (as $x \to 0$) if

$$\lim_{x\to 0} \frac{f(x)}{x} = 0.$$

We now recall Taylor's theorem from calculus.

Theorem 2.28. *Let $k \in \mathbb{N}$ and suppose that $f : \mathbb{R} \to \mathbb{R}$ is k times differentiable at some $a \in \mathbb{R}$. Then there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ such that*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k$$

and

$$\lim_{x\to a} h_k(x) = 0.$$

Example 2.29. *Consider the function $f(x) = x + x^2$. Note that*

$$\frac{f(x)}{x} = \frac{x}{x} + \frac{x^2}{x} \overset{x\to 0}{\longrightarrow} 0,$$

so $f = o(x)$.

We now recall a result from calculus regarding a class of ordinary differential equations (ODE).

Proposition 2.30. *Consider constant $\alpha, x_0 \in \mathbb{R}$ and a function $g : \mathbb{R} \to \mathbb{R}$. The unique solution to the ODE*

$$\begin{cases} f'(t) \doteq \frac{d}{dt} f(t) = \alpha f(t) + g(t) \\ f(0) = 0 \end{cases} \tag{5}$$

is given by

$$f(t) = e^{\alpha t} \int_0^t e^{-\alpha s} g(s) ds,$$

Proof. Define

$$F(t) \doteq e^{\alpha t}, \quad G(t) \doteq \int_0^t e^{-\alpha s} g(s) ds,$$

and observe that

$$F'(t) = \alpha e^{\alpha t},$$

and

$$G'(t) = e^{-\alpha t} g(t).$$

Since $f(t) = F(t)G(t)$, so the chain rule tells us that

$$\begin{aligned} f'(t) &= F'(t)G(t) + F(t)G'(t) \\ &= \alpha e^{\alpha t} \int_0^t e^{-\alpha s} g(s) ds + e^{\alpha t} e^{-\alpha t} g(t) \\ &= \alpha \left(e^{\alpha t} \int_0^t e^{-\alpha s} g(s) ds \right) + g(t) \\ &= \alpha f(t) + g(t), \end{aligned}$$

and clearly

$$f(0) = 0,$$

so we have shown that f solves (5).

We omit the proof of uniqueness. □

The next theorem characterizes the Poisson process in terms of infinitesimal probabilities.

Theorem 2.31. *A counting process $\{N_t\}_{t\geq 0}$ is a Poisson process with rate $\lambda > 0$ if and only if the following properties hold:*

(1) $N_0 = 0$
(2) The process has stationary and independent increments.
(3) $\mathbb{P}(N_t = 0) = 1 - \lambda t + o(t)$ (as $t \to 0$).
(4) $\mathbb{P}(N_t = 1) = \lambda t + o(t)$ as (as $t \to 0$).

(5) $\mathbb{P}(N_t > 1) = o(t)$ *(as $t \to 0$).*

Proof. Suppose that $\{N_t\}_{t\geq 0}$ is a Poisson process with rate $\lambda > 0$. Then, by definition, (1) holds, and (2) was shown in Theorem 2.22. Additionally, since $N_t \sim \text{Poisson}(\lambda t)$, we can apply Theorem 2.28 with $k = 1$ to expand the function $t \mapsto e^{-\lambda t}$ around $a = 0$ to obtain

$$\mathbb{P}(N_t = 0) = e^{-\lambda t} = 1 - \lambda t + h_1(t)t,$$

where $\lim_{t\to 0} h_1(t) = 0$. Note that as $t \to 0$,

$$\frac{h_1(t)t}{t} = h_1(t) \to 0,$$

so $h_1(t)t = o(t)$, meaning that

$$\mathbb{P}(N_t = 0) = e^{-\lambda t} = 1 - \lambda t + o(t), \tag{6}$$

showing that (3) holds. Applying 6, we see that

$$\mathbb{P}(N_t = 1) = e^{-\lambda t}\lambda t = (1 - \lambda t + o(t))\lambda t = \lambda t - \lambda^2 t^2 + \lambda t o(t) = \lambda t + o(t),$$

where we have used the fact that

$$-\lambda^2 t^2 + \lambda t o(t) = o(t),$$

since, as $t \to 0$,

$$\frac{-\lambda^2 t^2 + \lambda t o(t)}{t} = -\lambda^2 t + \lambda o(t) \to 0.$$

This shows that (4) holds. Finally, we note that

$$\begin{aligned}\mathbb{P}(N_t > 1) = 1 - \mathbb{P}(N_t \leq 1) &= 1 - \mathbb{P}(N_t = 0) - \mathbb{P}(N_t = 1) = 1 - (1 - \lambda t + o(t)) - (\lambda t + o(t)) \\ &= o(t) - o(t),\end{aligned}$$

so, clearly, $\mathbb{P}(N_t > 1) = o(t)$.

We now assume that (1) – (5) hold and show that $\{N_t\}_{t\geq 0}$ is a Poisson process with rate λ. As a consequence of Theorem 2.26, it suffices to show that $N_t \sim \text{Poisson}(\lambda t)$ for each $t \geq 0$. We begin by fixing $t \geq 0$ and letting

$$p_k(t) \doteq \mathbb{P}(N_t = k), \quad k \in \mathbb{N}_0.$$

For $k \in \mathbb{N}$,

$$\begin{aligned}
p_k(t+h) &= \mathbb{P}(N_{t+h} = k) \\
&\overset{1}{=} \sum_{j=0}^{k} \mathbb{P}(N_{t+h} = k | N_t = j)\mathbb{P}(N_t = j) \\
&= \sum_{j=0}^{k} \mathbb{P}(N_{t+h} - N_t = k - j | N_t = j) p_j(t) \\
&\overset{2}{=} \sum_{j=0}^{k} \mathbb{P}(N_{t+h} - N_t = k - j) p_j(t) \\
&\overset{3}{=} \sum_{j=0}^{k} \mathbb{P}(N_h - N_0 = k - j) p_j(t) \\
&\overset{4}{=} \sum_{j=0}^{k} \mathbb{P}(N_h = k - j) p_j(t) \\
&= \mathbb{P}(N_h = k - k) p_k(t) + \mathbb{P}(N_h = k - (k-1)) p_{k-1}(t) + \sum_{j=0}^{k-2} \mathbb{P}(N_h = k - j) p_j(t) \\
&= \mathbb{P}(N_h = 0) p_k(t) + \mathbb{P}(N_h = 1) p_{k-1}(t) + \sum_{j=0}^{k-2} \mathbb{P}(N_h = k - j) p_j(t)
\end{aligned}$$

In the display above, $\overset{1}{=}$ follows from the fact that counting processes are nondecreasing, $\overset{2}{=}$ follows from the independent increments property, $\overset{3}{=}$ follows from the stationary increments property, and $\overset{4}{=}$ follows from the fact that counting processes, by definition, begin at 0. Using assumptions (3) – (5), it follows that

$$\begin{aligned}
p_k(t+h) &= (1 - \lambda h + o(h)) p_k(t) + (\lambda h + o(h)) p_{k-1}(t) + \sum_{j=0}^{k=2} o(h) p_j(t) \\
&= (1 - \lambda h) p_k(t) + \lambda h p_{k-1}(t) + \sum_{j=0}^{k} o(h) p_j(t).
\end{aligned}$$

Therefore,

$$p_k(t+h) - p_k(t) = -\lambda h p_k(t) + \lambda h p_{k-1}(t) + \sum_{j=0}^{k} o(h) p_j(t),$$

and, dividing both sides of the previous display by h,

$$\frac{p_k(t+h) - p_k(t)}{h} = -\lambda p_k(t) + \lambda p_{k-1}(t) + \sum_{j=0}^{k} \frac{o(h)}{h} p_j(t),$$

so it follows that

$$p_k'(t) = \lim_{h \to 0} \frac{p_k(t+h) - p_k(t)}{h} = \lambda(p_{k-1}(t) - p_k(t)) + \sum_{j=0}^{k} \lim_{h \to 0} \frac{o(h)}{h} p_j(t) = \lambda(p_{k-1}(t) - p_k(t)). \tag{7}$$

Using a similar argument as above, we can show that when $k \doteq 0$, we have

$$p_0'(t) = -\lambda p_0(t), \tag{8}$$

and, once more noting that, since $\{N_t\}$ is a counting process, we have the initial condition

$$p_0(0) = \mathbb{P}(N_0 = 0) = 1. \tag{9}$$

Combining (8) and (9), we see that $p_0(t)$ satisfies the ordinary differential equation (ODE)

$$\begin{cases} p_0(0) = 1 \\ p_0'(t) = -\lambda p_0(t) \quad t \geq 0, \end{cases} \tag{10}$$

so if we take

$$\alpha = -\lambda, \quad f(t) = p_0(t), \quad g(t) \doteq 0,$$

in (5), then Proposition 2.30 tells us that the unique solution to the ODE in (10) is given by

$$p_0(t) \doteq e^{-\lambda t}, \quad t \geq 0. \tag{11}$$

Taking $k = 1$ in (7) and plugging in (11), we see that

$$p_1'(t) = \lambda p_0(t) - p_1(t) = \lambda e^{-\lambda t} - \lambda p_1(t),$$

so, using the fact that $\mathbb{P}(N_0 = 1) = 0$, we see that $p_1(t)$ solves the ODE

$$\begin{cases} p_1(0) = 0 \\ p_1'(t) = \lambda e^{-\lambda t} - \lambda p_1(t). \end{cases} \tag{12}$$

From Proposition 2.30 it follows from taking

$$\alpha \doteq -\lambda, \quad f(t) = p_1(t), \quad g(t) \doteq \lambda e^{-\lambda t},$$

that the solution to the ODE in (12) is given by

$$p_1(t) = e^{-\lambda t} \int_0^t e^{\lambda s} \lambda e^{-\lambda s} ds = e^{-\lambda t} \lambda t = e^{-\lambda t} \frac{(\lambda t)^1}{1!}.$$

Repeating this process and arguing by induction, we see that for $k \in \mathbb{N}$,

$$p_k(t) = e^{-\lambda t} \frac{(\lambda t)^k}{k!},$$

which proves that $N_t \sim \text{Poisson}(\lambda t)$, thereby completing the proof. □

2.4. **Superposition and Splitting of Poisson Processes.** In this section we discuss how someone can combine various Poisson processes into another Poisson process, and how someone can decompose a single Poisson process into multiple Poisson processes.

2.4.1. *Superposition.* Customers visit a cafe for three possible reasons: to order food to go, to order coffee to go, or to eat at the cafe. The cafe's owner knows that he can model the arrival of customers who are only interested in ordering food to go as a Poisson process with a rate of $\lambda_1 = 15$ per hour, the arrival of customers who are only interested in ordering coffee to go as a Poisson process with a rate of $\lambda_2 = 25$ per hour, and the arrivals of customers who are only interested in eating at the cafe as a Poisson process with a rate of $\lambda_3 = 10$ per hour. Additionally, the owner thinks that it is reasonable to that the arrival of each type of customer is independent of the arrival of any other type of customer. What is the probability that over the course of 2 hours, the cafe has exactly 100 customers in total?

If we let $\{N_t^1\}_{t\geq 0} \sim \text{PP}(\lambda_1), \{N_t^2\}_{t\geq 0} \sim \text{PP}(\lambda_2)$, and $\{N_t^3\}_{t\geq 0} \sim \text{PP}(\lambda_3)$ model the arrivals of the three types of customers. Then the total number of customers who have visited the cafe by time $t \geq 0$ is given by

$$N_t \doteq N_t^1 + N_t^2 + N_t^3,$$

so we are interested in determining the probability distribution of N_t and also in determining whether it is a Poisson process.

Definition 2.32. *Let $\{N_t^1\}_{t\geq 0} \sim PP(\lambda_1), \{N_t^2\}_{t\geq 0} \sim PP(\lambda_2), \dots, \{N_t^n\}_{t\geq 0} \sim PP(\lambda_n)$ be independent Poisson processes. The stochastic process $\{N_t\}_{t\geq 0}$ defined by*

$$N_t \doteq \sum_{i=1}^{n} N_t^i,$$

is called the superposition of the n processes.

Before proving that the superposition is a Poisson process, we begin with the following proposition.

Proposition 2.33. *Let $\{X_t\}$ and $\{Y_t\}$ be stochastic processes with stationary and independent increments. Then the process $\{Z_t\}$ defined by $Z_t \doteq X_t + Y_t$ has stationary and independent increments as well.*

Proof. The proof is left as an exercise. □

The following result says that the superposition of independent Poisson processes is itself a Poisson process.

Theorem 2.34. *Let*

$$N_t \doteq \sum_{i=1}^{n} N_t^i,$$

be the superposition of independent Poisson process with rates $\lambda_1, \lambda_2, \dots, \lambda_n$. Then $\{N_t\}_{t\geq 0}$ is a Poisson process with rate λ, where

$$\lambda \doteq \sum_{i=1}^{n} \lambda_i.$$

Proof. We consider the case when $n = 2$. The result for a general $n \in \mathbb{N}$ follows from an inductive argument. We verify that $\{N_t\}_{t\geq 0}$ satisfies the conditions of Theorem 2.26. Note that $\{N_t^1\}$ and $\{N_t^2\}$ have independent increments, so if we let $t_1 < t_2 \leq t_3 < t_4$, then, for $i, j \in \mathbb{N}_0$, if we define the set

$$\mathscr{A}_{i,j} \doteq \{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \mathbb{N}_0^4 : \alpha_1, \alpha_2 \leq i, \alpha_1 + \alpha_2 = i, \alpha_3, \alpha_4 \leq j, \alpha_3 + \alpha_4 = j\}$$

then we have

$$\begin{aligned}
\mathbb{P}(N_{t_2} - N_{t_1} = i, N_{t_4} - N_{t_3} = j) &= \mathbb{P}(N_{t_2}^1 + N_{t_2}^2 - (N_{t_1}^1 + N_{t_1}^2) = i, N_{t_4}^1 + N_{t_4}^2 - (N_{t_3}^1 + N_{t_3}^2) = j) \\
&= \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathscr{A}_{i,j}} \mathbb{P}(N_{t_2}^1 - N_{t_1}^1 = \alpha_1, N_{t_2}^2 - N_{t_1}^2 = \alpha_2, N_{t_4}^1 - N_{t_3}^1 = \alpha_3, N_{t_4}^2 - N_{t_3}^2 = \alpha_4) \\
&= \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathscr{A}_{i,j}} \mathbb{P}(N_{t_2}^1 - N_{t_1}^1 = \alpha_1, N_{t_4}^1 - N_{t_3}^1 = \alpha_3)\mathbb{P}(N_{t_2}^2 - N_{t_1}^2 = \alpha_2, N_{t_4}^2 - N_{t_3}^2 = \alpha_4) \\
&= \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathscr{A}_{i,j}} \mathbb{P}(N_{t_2}^1 - N_{t_1}^1 = \alpha_1)\mathbb{P}(N_{t_4}^1 - N_{t_3}^1 = \alpha_3)\mathbb{P}(N_{t_2}^2 - N_{t_1}^2 = \alpha_2)\mathbb{P}(N_{t_4}^2 - N_{t_3}^2 = \alpha_4) \\
&= \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathscr{A}_{i,j}} \mathbb{P}(N_{t_2}^1 - N_{t_1}^1 = \alpha_1, N_{t_2}^2 - N_{t_1}^2 = \alpha_2)\mathbb{P}(N_{t_4}^1 - N_{t_3}^1 = \alpha_3, N_{t_4}^2 - N_{t_3}^2 = \alpha_4) \\
&= \sum_{(\alpha_1,\alpha_2,\alpha_3,\alpha_4)\in\mathscr{A}_{i,j}} \mathbb{P}(N_{t_2}^1 + N_{t_2}^2 - (N_{t_1}^1 + N_{t_2}^2) = \alpha_1 + \alpha_2)\mathbb{P}(N_{t_4}^1 + N_{t_4}^2 - (N_{t_3}^1 + N_{t_3}^2) = \alpha_3 + \alpha_4) \\
&= \mathbb{P}(N_{t_2} - N_{t_1} = i)\mathbb{P}(N_{t_4} - N_{t_3} = j)
\end{aligned}$$

which shows that the independent increment property holds. The proof that $\{N_t\}$ has stationary increments is similar and is omitted. From Theorem 2.26, it suffices to show that there is some $\lambda > 0$ such that for each $t \geq 0$, $N_t \sim \text{Poisson}(\lambda t)$. Let $\lambda \doteq \lambda_1 + \lambda_2$, and recall that $N_t^1 \sim \text{Poisson}(\lambda_1 t)$ and $N_t^2 \sim \text{Poisson}(\lambda_2 t)$ are independent, so it follows that $N_t = N_t^1 + N_t^2 \sim \text{Poisson}(\lambda t)$ □

Example 2.35. *Recall the cafe example from the beginning of this section. There we modeled the arrival of three types of customers as* $\{N_t^1\}_{t\geq 0} \sim PP(\lambda_1), \{N_t^2\}_{t\geq 0} \sim PP(\lambda_2)$, *and* $\{N_t^3\}_{t\geq 0} \sim PP(\lambda_3)$, *where* $\lambda_1 = 15, \lambda_2 = 25$, *and* $\lambda_3 = 10$. *If we let* $N_t \doteq N_t^1 + N_t^2 + N_t^3$ *denote the total number of customers who have arrived by time* $t \geq 0$, *then Theorem* 2.34 *tells us that* $\{N_t\} \sim PP(50)$, *and we can see that the probability that the probability that the cafe has exactly 100 customers over the course of 2 hours is given by*

$$\mathbb{P}(N_{t+2} - N_t = 100) = \mathbb{P}(N_2 - N_0 = 100) = \mathbb{P}(N_2 = 100) = e^{-2\cdot 50}\frac{(2\cdot 50)^{100}}{100!} \approx 0.0399,$$

or approximately 4%.

In the definition of the superposition of Poisson processes, why do we require independence? Let $\{N_t\} \sim \text{PP}(\lambda)$, and for each $t \geq 0$, let $N_t^1 = N_t^2 = N_t$ be the same as N_t. Then, for each $t \geq 0$,

$$\mathbb{P}(N_t^1 + N_t^2 \text{ is even}) = 1$$

which tells us that $N_t^1 + N_t^2$ does not follow a Poisson distribution.

In the next section we discuss splitting, which can be considered the opposite of the superposition. Splitting allows us to take a single Poisson processes and to break it up into several independent Poisson processes.

2.4.2. *Splitting.* Suppose that a geologist is interested in modeling the number of earthquakes that occur in a particular location each year. She knows that the arrival of earthquakes of any magnitude can be modeled as a Poisson process with rate $\lambda_1 = 7$ per day. Additionally, she knows that roughly 85% of earthquakes have a magnitude of 3 or less, 14.99% have a magnitude between 3 and 6, and 0.01% have a magnitude of 6 or greater. Under what conditions does the arrival of earthquakes of magnitude 6 or greater also follow a Poisson process? What is the rate of the arrivals? The key assumption is that the arrival of the various types of earthquakes are all independent of one another.

Definition 2.36. *Let* $\{N_t\} \sim PP(\lambda)$, *and suppose that each event can be classified as an event of a different type, where there are* r *types of events. Suppose that the probability that each event is of type* i *is* p_i, *where* $p_1, \ldots, p_r \in (0,1]$ *are such that*

$$\sum_{i=1}^{r} p_i = 1.$$

Additionally, assume that the type of each event is independent of the type of all previous events. If we let N_t^i *denote the number of events of type* i *that have occurred by time* $t \geq 0$, *then* $\{N_t^1\}, \ldots, \{N_t^r\}$ *are* split processes, *as we can write*

$$N_t \doteq \sum_{i=1}^{r} N_t^i.$$

The following result says that if we split a Poisson processes, then each of the resulting split processes are independent of one another, and each of them are a Poisson process.

Theorem 2.37. *Suppose that $\{N_t\} \sim PP(\lambda)$ can be written in terms of split processes $\{N_t^1\}, \dots, \{N_t^r\}$, where*

$$N_t = \sum_{i=1}^{r} N_t^i.$$

Then $\{N_t^i\} \sim PP(\lambda p_i)$, and the r processes $\{N_t^1\}, \dots, \{N_t^r\}$ are independent.

Proof. We consider the case when $r = 2$, as the general result is similar. We begin by noting that for $0 \le j \le k$,

$$N_t^1 | N_t = k \sim \text{Binomial}(k, p_1),$$

meaning that

$$\mathbb{P}(N_t^1 = j | N_t = k) = \binom{k}{j} p_1^j (1-p_1)^{k-j} = \binom{k}{j} p_1^j p_2^{k-j} = \frac{k!}{j!(k-j)!} p_1^j p_2^{k-j}.$$

If follows that for $t \ge 0$,

$$\begin{aligned}
\mathbb{P}(N_t^1 = k_1, N_t^2 = k_2) &= \mathbb{P}(N_t^1 = k_1, N_t^2 = k_2, N_t = k_1 + k_2) \\
&= \mathbb{P}(N_t^1 = k_1, N_t^2 = k_2 | N_t = k_1 + k_2)\mathbb{P}(N_t = k_1 + k_2) \\
&= \mathbb{P}(N_t^1 = k_1 | N_t = k_1 + k_2)\mathbb{P}(N_t = k_1 + k_2) \\
&= \binom{k_1 + k_2}{k_1} p_1^{k_1} p_2^{k_2} e^{-\lambda t} \frac{(\lambda t)^{k_1 + k_2}}{(k_1 + k_2)!} \\
&= \frac{(k_1 + k_2)!}{k_1! k_2!} p_1^{k_1} p_2^{k_2} e^{-\lambda p_1 t} e^{-\lambda p_2 t} \frac{(\lambda t)^{k_1} (\lambda t)^{k_2}}{(k_1 + k_2)!} \\
&= \frac{(\lambda p_1 t)^{k_1}}{k_1!} \frac{(\lambda p_2 t)^{k_2}}{k_2!},
\end{aligned}$$

which tells us that $\{N_t^1\}$ and $\{N_t^2\}$ are independent, and that for each $t \ge 0$, $N_t^1 \sim \text{Poisson}(\lambda p_1 t)$ and $N_t^2 \sim \text{Poisson}(\lambda p_2 t)$. In order to complete the proof, it suffices to show that $\{N_t^1\}$ and $\{N_t^2\}$ both have independent and stationary increments. This proof is omitted. □

We now return to the earthquake example from the beginning of this section.

Example 2.38. *Suppose that a geologist is interested in modeling the number of earthquakes that occur in a particular location each year. She knows that the arrival of earthquakes of any magnitude can be modeled as a Poisson process with rate $\lambda = 7$ per day. Additionally, she knows that roughly 85% of earthquakes have a magnitude of 3 or less, 14.99% have a magnitude between 3 and 6, and 0.01% have a magnitude of 6 or greater. Additionally, assume that the magnitude of each earthquake is independent of the magnitude of the other earthquakes. What is the probability that over the course of one year, what is the probability that there is at least one earthquake of magnitude* 6 *or greater?*

Let $\{N_t\}$ model the arrival of earthquakes of any type. If we let $\{N_t^a\}$, $\{N_t^b\}$, and $\{N_t^c\}$ model the arrival of earthquakes of magnitude 3 or less, between 3 and 6, and more than 6, respectively, then Theorem 2.37 *tells us that $\{N_t^c\} \sim$ Poisson$(7 \cdot 0.0001)$, so the probability that there is at least one earthquake of magnitude* 6 *or greater over the course of a year is given by*

$$\mathbb{P}(N_{365}^c \ge 1) = 1 - \mathbb{P}(N_{365}^c = 0) = 1 - e^{-0.0007 \cdot 365} \frac{(0.0007 \cdot 365)^0}{0!}.$$

Additionally, the expected number of earthquakes of magnitude 6 *or greater each year is given by*

$$\mathbb{E}(N^c_{365}) = 0.0007 \cdot 365 = 0.2555.$$

2.5. **Non-Homogeneous Poisson Processes.**

Recall that Poisson processes have stationary increments, meaning that the number of events that occur over some time interval $[s,t]$ depends only on the length of the interval, namely $t - s$. However, for many situations, the requirement that the process has stationary increments may be too restrictive for modeling purposes. For example, suppose that we are interested in modeling the number of customers who visit a bank over some period of time. If the bank is only open from 8am until 5pm, then a realistic model would need to ensure that no events (i.e., arrivals of customers) take place after 5pm and before 8am.

For example, suppose that we may still assume that during business hours (i.e., between 8am and 5pm), customers arrive at the bank at a rate of $\lambda = 10$ per hour. If we start counting arrivals at 8am on the first day, then the rate at which customers arrive at each time instant $t \in [0,9)$, is $\lambda(t) = 10$. Then, for each $t \in [9,24)$ is $\lambda(t) = 0$, since the bank is closed then. Similarly, for each $t \in [24,33)$, the rate of arrivals is $\lambda(t) = 10$. Continuing this processes, we can see that for each $t \geq 0$, the rate of arrivals is given by the function

$$\lambda(t) = \begin{cases} 10 & t \in [24k, 24k+9), \text{ for some } k \in \mathbb{N}_0 \\ 0 & \text{otherwise.} \end{cases}$$

We now introduce the notion of a non-homogeneous Poisson process.

Definition 2.39. *Let* $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ *be a function such that for all* $t \geq 0$,

$$\Lambda(t) \doteq \int_0^t \lambda(s)ds < \infty.$$

A counting processes $\{N_t\}$ *is said to be a non-homogeneous Poisson process with rate function* $\lambda(\cdot)$ *if*

(1) $\{N_t\}$ *has independent increments.*
(2) For each $t \geq 0$, $N_t \sim$ *Poisson*$(\Lambda(t))$.

We write $\{N_t\} \sim$ *NPP*$(\lambda(\cdot))$

The next result states some important properties of non-homogeneous Poisson processes. The proof is omitted.

Proposition 2.40. *Suppose that* $\{N_t\} \sim$ *NPP*$(\lambda(\cdot))$ *for some* $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$. *Then*

(1) $\mathbb{P}(N_0 = 0) = 1$.
(2) For each $s, t \geq 0$, $N_{s+t} - N_s \sim$ *Poisson*$(\Lambda(s+t) - \Lambda(s))$.

We can see that the example from the beginning of this section describes a non-homogeneous Poisson process with rate function

$$\lambda(t) = \begin{cases} 10 & t \in [24k, 24k+9), \text{ for some } k \in \mathbb{N}_0 \\ 0 & \text{otherwise.} \end{cases}$$

Below we consider another example of a non-homogeneous Poisson process.

Example 2.41. *Suppose that* $\{N_t\} \sim NPP(\lambda(\cdot))$, *where*

$$\lambda(t) = t^2.$$

Find the pdf and mean of N_t.

Observe that

$$\Lambda(t) = \int_0^t \lambda(s)ds = \int_0^t s^2 ds = \frac{t^3}{3},$$

so $N_t \sim Poisson\left(\frac{t^3}{3}\right)$. *It follows that the pmf of* N_t *is given by*

$$p_t(x) \doteq \mathbb{P}(N_t = x) = e^{-\frac{t^3}{3}} \frac{\left(\frac{t^3}{3} \cdot t\right)^x}{x!} = e^{-\frac{t^3}{3}} \frac{t^{4x}}{x!3^x},$$

and

$$\mathbb{E}(N_t) = \frac{t^3}{3}.$$

In this section we introduced non-homogeneous Poisson processes, which can be used to model situations where the rate of arrivals varies over time. In the next section we introduce another generalization of Poisson processes in which more than one event can occur at each time instant.

2.6. **Compound Poisson Processes.** Customers arrive at and depart from a theme park in groups ranging from 2 to 4 people. Suppose that the size of each group is independent of the size of the other groups and the number of groups that have arrived up to that point. Suppose that the group sizes all follow a common probability distribution

$$p(k) = \begin{cases} \frac{1}{2} & k = 2 \\ \frac{1}{4} & k = 3 \\ \frac{1}{4} & k = 4. \end{cases}$$

Additionally, suppose that we can model the arrival of groups of customers as a Poisson process with a rate of $\lambda = 120$ groups per hour. If we let Z_t denote the number of customers that have arrived up to time $t \geq 0$, then $\{Z_t\}$ is clearly not a Poisson process, as it can increase by more than one unit in a single time instant. However, $\{Z_t\}$ is closely related to the Poisson process, and indeed can be represented in terms of it. To do so, let N_t denote the number of groups that have arrived by time $t \geq 0$, and let X_i denote the size of the i^{th} group. Then we have, for each $t \geq 0$,

$$Z_t = \sum_{i=1}^{N_t} X_i,$$

where we use the convention that

$$\sum_{i=1}^{0} x_i \doteq 0.$$

The process $\{Z_t\}$ is referred to as a compound Poisson process. We introduce the general definition below.

Definition 2.42. *Let* $\{N_t\} \sim PP(\lambda)$, *and let* $\{X_n\}$ *be a collection of iid random variables that are also independent of* $\{N_t\}$. *Define the process* $\{Z_t\}$ *by*

$$Z_t \doteq \sum_{n=1}^{N_t} X_n, \quad t \geq 0.$$

We refer to $\{Z_t\}$ *as a* ***compound Poisson process*** *or* ***CPP****.*

Note: *in this definition we do not require that* $\{X_n\}$ *are integer valued or even non-negative.*

While it is difficult, in general, to describe the probability distribution of a compound Poisson process, following result illustrates how to calculate the mean and variance of a compound Poisson process.

Theorem 2.43. *Suppose that* $\{Z_t\}$ *is a CPP of the form*

$$Z_t = \sum_{n=1}^{N_t} X_n,$$

where $\{N_t\} \sim PP(\lambda)$, $\mathbb{E}(X_n) = \mu$, *and* $\mathbb{E}(X_n^2) = \theta^2$. *Then, for each* $t \geq 0$,

$$\underline{\mathbb{E}(Z_t) = \lambda \mu t, \quad Var(Z_t) = \lambda \theta^2 t.}$$

Proof. Using Wald's Identity (see e.g., Proposition B.30), we see that

$$\mathbb{E}(Z_t) = \mathbb{E}\left[\sum_{n=1}^{N_t} X_n\right] = \mathbb{E}(N_t)\mathbb{E}(X_n) = \lambda \mu t,$$

and, using the fact that $\text{Var}(N_t) = \lambda$ and $\mathbb{E}(X_n^2) = \text{Var}(X_n) + (\mathbb{E}(X_n))^2$,

$$\text{Var}\left(\sum_{i=1}^{N_t} X_n\right) = \mathbb{E}(N_t)\text{Var}(X_n) + (\mathbb{E}(X_n))^2\text{Var}(N_t) = \text{Var}(N_t)(\text{Var}(X_n) + (\mathbb{E}(X_n))^2) = \text{Var}(N_t)\mathbb{E}(X_n^2) = \lambda\theta^2 t.$$

□

We now return to the example from the beginning of this section.

Example 2.44. *Customers arrive at and depart from a theme park in groups ranging from 2 to 4 people. Suppose that the size of each group is independent of the size of the other groups and the number of groups that have arrived up to that point. Suppose that the group sizes all follow a common probability distribution*

$$p(k) = \begin{cases} \frac{1}{2} & k = 2 \\ \frac{1}{4} & k = 3 \\ \frac{1}{4} & k = 4. \end{cases}$$

Additionally, suppose that we can model the arrival of groups of customers as a Poisson process with a rate of $\lambda = 120$ *groups per hour. Compute the expected number and the variance of the number of customers that we expect to arrive within two hours of the park opening.*

Let X_n *be iid random variables with pmf* p, *and let* $\{N_t\} \sim PP(\lambda)$. *Then the number of customers who have arrived by time* $t \geq 0$ *can be described using the CPP*

$$Z_t = \sum_{n=1}^{N_t} X_n,$$

where

$$\mathbb{E}(X_n) = 2.75, \quad \mathbb{E}(X_n^2) = 8.25.$$

Then we have

$$\mathbb{E}(Z_2) = 2.75 \cdot 120 \cdot 2 = 660, \quad Var(Z_2) = 120 \cdot 8.25 \cdot 2 = 1980.$$

2.7. **Spatial Poisson Processes.** Poisson processes (and the variations of these processes discussed in the preceding sections) are used to model the number of events that occur as time progresses. In order to motivate the notion of a spatial Poisson process, we begin by noting that if $\{N_t\} \sim \text{PP}(\lambda)$, then, for $0 \le s < t$, the quantity $N((s,t])$ defined by

$$N((s,t]) \doteq N_t - N_s,$$

describes the number of events that occur in the time interval $(s,t]$. Suppose that we were instead interested in modeling the number of events that take place in a specific region. For example, if we are modeling the the way that trees are distributed throughout a forest, then for each region A in the forest, we would like to define

$$N_A \doteq \text{"the number of trees in the region } A\text{"}.$$

The following definitions will be helpful when introducing the spatial Poisson process.

Definition 2.45. *Fix $d \ge 1$ and recall that*

$$\mathbb{R}^d \doteq \{x = (x_1, x_2, \dots, x_d) : x_1, x_2, \dots, x_d \in \mathbb{R}\}.$$

*We say that a set $A \subseteq \mathbb{R}^d$ is **bounded** if there is some $C > 0$ such that for all $x, y \in A$,*

$$\|x - y\|^2 \doteq \sum_{i=1}^{d} (x_i - y_i)^2 \le C.$$

For a bounded set $A \subseteq \mathbb{R}^d$, we write $|A|$ to denote the size of A. When $d = 2$, size refers to area, when $d = 3$, size refers to volume, and so on.

This brings us to the notion of a spatial Poisson process.

Definition 2.46. *A collection of random variables $\{N_A\}_{A \subseteq \mathbb{R}^d}$ is said to be a **spatial Poisson process** with parameter $\lambda > 0$ if the following hold:*

1. *For each bounded set $A \subseteq \mathbb{R}^d$, $N_A \sim \text{Poisson}(\lambda|A|)$.*
2. *For disjoint sets $A, B \subseteq \mathbb{R}^d$, N_A and N_B are independent random variables.*

*Similarly, for a bounded subset $A \subseteq \mathbb{R}^d$, we say that a collection of random variables $\{\tilde{N}_B\}_{B \subseteq A}$ is a **spatial Poisson process in** A if the following hold:*

1. *For each bounded set $B \subseteq A$, with $p_B \doteq \frac{|B|}{|A|}$, we have that $\tilde{N}_B \sim \text{Poisson}(\lambda p_B)$.*
2. *For disjoint sets $B, C \subseteq A$, $\bar{N}_B$ and $\bar{N}_C$ are independent random variables.*

An important feature of spatial Poisson processes is that, conditional on the number of points occurring in a region A, the positions of the points are uniformly distributed in A. We omit the proof, but the result is stated below.

Proposition 2.47. *Let $\{N_A\}_{A \subseteq \mathbb{R}^d}$ be a spatial Poisson process with parameter $\lambda > 0$. Then for each $A \subseteq \mathbb{R}^d$, and each $B \subseteq A$,*

$$\mathbb{P}(N_B = k | N_A = n) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $p = \frac{|B|}{|A|}$.

Below we consider an example of a spatial Poisson process.

Example 2.48. *Consider a circuit board that is 1 inch wide and 2 inches tall. As time progresses, defects appear in the circuit board; we can model the arrival of these defects as a Poisson process with a rate of $\lambda = 1$ per month. Additionally, it is reasonable to assume that the position of each defect is uniformly distributed across the surface of the circuit board.*

Denote the surface of the circuit board by

$$S = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 2\}.$$

For each $t \geq 0$ and $B \subseteq S$, let $N_B(t)$ denote the number of defects in the region B by time $t \geq 0$. What is the probability mass function, of $N_B(t)$?

We have, for $k \in \mathbb{N}_0$, from the law of total probability, with $p_B \doteq \frac{|B|}{|S|} = \frac{|B|}{2}$,

$$\begin{aligned}
\mathbb{P}(N_B(t) = k) &= \sum_{n=0}^{\infty} \mathbb{P}(N_B(t) = k | N_S(t) = n)\mathbb{P}(N_S(t) = n) \\
&= \sum_{n=k}^{\infty} \mathbb{P}(N_B(t) = k | N_S(t) = n)\mathbb{P}(N_S(t) = n) \\
&= \sum_{n=k}^{\infty} \binom{n}{k} p_B^k (1 - p_B)^{n-k} e^{-t} \frac{t^n}{n!} \\
&= \frac{(p_B t)^k}{k!} e^{-t} \sum_{n=k}^{\infty} \frac{\big((1 - p_B)t\big)^{n-k}}{(n-k)!} \\
&= \frac{(p_B t)^k}{k!} e^{-t} \sum_{n=0}^{\infty} \frac{\big((1 - p_B)t\big)^{n}}{n!} \\
&= \frac{(p_B t)^k}{k!} e^{-t} e^{(1-p_B)t} \\
&= e^{-p_B t} \frac{(p_B t)^k}{k!},
\end{aligned}$$

which says that $N_B(t) \sim$ Poisson($p_B t$). Additionally, it is clear that, since the positions of the defects occur uniformly at random, for each $t \geq 0$ and $B, C \subseteq S$, $N_B(t)$ and $N_C(t)$ are independent. Therefore, for each fixed $t \geq 0$, $\{N_B(t)\}_{B \subseteq S}$ is a spatial Poisson process in S.

3. Continuous-Time Markov Chains

In this section we introduce the notion of a continuous-time Markov chain (CTMCs). Unlike discrete-time Markov chains (DTMCs), these continuous-time processes not only allows us to track the trajectory of a process moving around some state space, but also to measure the length of time that the process spends in each state. To see why we might be interested in studying CTMCs, consider the following model.

A line at a store starts out with no one in it. Customers independently arrive one by one, and the time that it takes for each subsequent customers to arrive follows an exponential distribution with a rate of λ per hour. It takes longer to help some customers than others; in particular, the time that it takes for each customer to be helped (and therefore to leave the line) follows an exponential distribution with a rate of μ per hour. If want to model the number of customers that are in line at each time instant, then it is not immediately clear whether a DTMC would be appropriate for doing so, since the arrival and service times of the customers are continuous random variables.

Now we formally introduce the notion of a CTMC.

Definition 3.1. *A continuous-time stochastic process $\{X_t\}_{t\geq 0}$ with discrete state space $\mathscr{S}$ is a **continuous-time Markov chain (CTMC)** if, for all $s, t \geq 0$,*

$$\mathbb{P}(X_{t+s} = y | X_s = x, X_u = x_u \textit{ for all } u \in [0,s)) = \mathbb{P}(X_{t+s} = y | X_s = x).$$

*Here, for each $u \in [0,s)$, x_u is a (possibly distinct) element of $\mathscr{S}$. A CTMC is said to be **time-homogeneous** if, for all $s, t \geq 0$,*

$$\mathbb{P}(X_{t+s} = y | X_s = x) = \mathbb{P}(X_t = y | X_0 = x).$$

In this course all of the CTMCs we encounter will be time-homogeneous, so we will generally drop that descriptor and simply refer to them as CTMCs.

*The **transition function** of a CTMC is the function $P : [0,\infty) \times \mathscr{S} \times \mathscr{S}$ defined by*

$$P_{x,y}(t) = \mathbb{P}(X_t = y | X_0 = x).$$

That is $P_{x,y}(t)$ describes the probability that if the chain is currently in state x, that after $t \geq 0$ time units, it will be in state y. Note that for each $t \geq 0$, $P(t)$ can be interpreted as an $|\mathscr{S}| \times |\mathscr{S}|$ matrix.

There was a question towards the end of class on April 12 about whether stationary increments and time-homogeneity are the same. Below we describe two situations where they are not.

Example 3.2. *The first is when the process $\{X_t\}$ takes values in a space that is not a vector space; for example if we let $\{X_t\}$ denote whether it is sunny or cloudy at time $t \geq 0$. Then, something like $X_t - X_s$ isn't necessarily defined, since $\mathscr{S}$ = {sunny, rainy} is not a vector space.*

Another setting is as follows; let $\{N_t\}$ be a Poisson process with rate 1, and, for each $t \geq 0$, let $X_t \doteq N_t^2$. Then, for $j, k \in \{0, 1, 4, 9, \ldots\}$, we have

$$\mathbb{P}(X_{t+s} = k | X_s = j) = \mathbb{P}(N_{t+s}^2 = k | N_s^2 = j) = \mathbb{P}(N_{t+s} = \sqrt{k} | N_s = \sqrt{j}) = \mathbb{P}(N_t = \sqrt{k} | N_0 = \sqrt{j}) = \mathbb{P}(X_t = k | X_0 = j)$$

so $\{X_t\}$ is time-homogeneous. On the other hand,

$$\mathbb{E}(X_{t+s} - X_s) = \mathbb{E}(N_{t+s}^2 - N_s^2) = \mathbb{E}(N_{t+s}^2) - \mathbb{E}(N_s^2) = (t+s) + (t+s)^2 - (s + s^2) = t + t^2 + 2st.$$

For example,

$$\mathbb{E}(X_3 - X_2) = 1 + 1 + 4 = 6,$$

while

$$\mathbb{E}(X_4 - X_3) = 1 + 1 + 6 = 8,$$

so $X_3 - X_2$ does not have the same distribution as $X_4 - X_3$ (if the distribution was the same, then their means would be as well). Note that here we are not dealing with any conditional probability or conditional expectation. This says that $\{X_t\}$ does not have stationary increments.

While an example like $\{X_t\}$ above might seem contradictory at first, we can make sense of this difference between time-homogeneity and stationary increments in this example as follows. Here, the probability distribution of the chain's future trajectory is determined by its current state, so if it is currently in state i^2, then we don't care how long it took to get to state i^2. This means that it is time-homogeneous.

On the other hand, the chain does not have stationary increments; each time the chain jumps, it jump size increases. First, it increases by 1, then by $4-1=3$, then by $9-4=5$, and so on. After a long time has passed, on average, it will tend to be in a higher state, and so can be expected to take larger jumps. This means that the chain does not have stationary increments.

Below we summarize some important distinctions between various definitions.

Proposition 3.3. *The following hold both for discrete-time and continuous-time stochastic processes.*

(1) *If $\{X_t\}$ is a stochastic process that has the stationary increments property and the independent increments property, then it is time homogeneous and has the Markov property.*
(2) *If $\{X_t\}$ is a stochastic process that has the independent increments property, then it also has the Markov property.*
(3) *If $\{X_t\}$ has the Markov property, then it does not necessarily have the independent increments property.*
(4) *If $\{X_t\}$ has the Markov property and the stationary increments property, then it is not necessarily time-homogeneous.*
(5) *If $\{X_t\}$ has the Markov property and is time-homogeneous, it does not necessarily have the stationary increments property.*

Proof. The proof is omitted. □

It is not immediately clear how one can calculate the transition function of a CTMC such as the one described in the beginning of this section. However, the following result, known as the Chapman-Kolmogorov equations, captures one of the most important properties of transition functions.

Proposition 3.4. *Let $\{X_t\}$ be a CTMC with transition function P. Then for each $s, t \geq 0$,*

$$P(s+t) = P(s)P(t).$$

That is, for each $x, y \in \mathscr{S}$ and $s, t \geq 0$,

$$P_{x,y}(s+t) = [P(s)P(t)]_{x,y} = \sum_{z \in \mathscr{S}} P_{x,z}(s)P_{z,y}(t).$$

Proof. Fix $x, y \in \mathscr{S}$ and $s, t \geq 0$. Then

$$\begin{aligned} P_{x,y}(s+t) &= \mathbb{P}(X_{s+t} = y | X_0 = x) \\ &= \sum_{z \in \mathscr{S}} \mathbb{P}(X_{s+t} = y | X_0 = x, X_t = z)\mathbb{P}(X_t = z | X_0 = x) \\ &= \sum_{z \in \mathscr{S}} \mathbb{P}(X_{s+t} = y | X_t = z)\mathbb{P}(X_t = z | X_0 = x) \\ &= \sum_{z \in \mathscr{S}} \mathbb{P}(X_s = y | X_0 = z)\mathbb{P}(X_t = z | X_0 = x) \\ &= \sum_{z \in \mathscr{S}} P_{x,z}(s) P_{z,y}(t) \end{aligned}$$

. □

We have already seen an important example of a CTMC. The following result shows that Poisson processes are in fact CTMS and calculates their transition function.

Proposition 3.5. *Let* $\{N_t\} \sim PP(\lambda)$*. Then* $\{N_t\}$ *is a CMTC on* $\mathscr{S} = \{0, 1, 2, \ldots\}$ *with transition function*

$$P_{x,y}(t) = e^{-\lambda t} \frac{(\lambda t)^{y-x}}{(y-x)!}, \quad x, y \in \mathscr{S}, t \geq 0.$$

Proof. The independent and stationary increments property of the Poisson process and the fact that $N_t \sim \text{Poisson}(\lambda t)$ ensure that

$$\begin{aligned} \mathbb{P}(N_{t+s} = y | N_s = x, N_u = x_u, \text{ for } u \in [0, s)) &= \mathbb{P}(N_{t+s} - N_s = y - x | N_s = x, N_u = x_u, \text{ for } u \in [0, s)) \\ &= \mathbb{P}(N_{t+s} - N_s = y - x) \\ &= \mathbb{P}(N_t = y - x) \\ &= e^{-\lambda t} \frac{(\lambda t)^{y-x}}{(y-x)!}, \end{aligned}$$

which shows that

$$P_{x,y}(t) = e^{-\lambda t} \frac{(\lambda t)^{y-x}}{(y-x)!}, \quad x, y \in \mathscr{S}, t \geq 0.$$

□

3.1. **Holding Times of CTMCs.** In this section we develop the necessary ideas to discuss more general classes of DTMC.

Definition 3.6. *Consider a CTMC* $\{X_t\}$ *on* $\mathscr{S}$*. The* ***holding time*** T_x *of state* x *is the amount of time that the chain spends in state* x *before it jumps to another state. Formally,*

$$T_x = \inf\{t > 0 : X_t \neq x \textit{ given that } X_0 = x\}.$$

Recall that since we are studying homogeneous CTMCs, that the distribution of T_x does not depend on the number of time that the chain has visited state x, since each time it arrives at state x, it "forgets" its past history.

Proposition 3.7. *Let* T_x *denote the holding time of state* x*. Then there is some* $q_x > 0$ *such that* $T_x \sim Exp(q_x)$*.*

Proof. Recall from Proposition 2.4 that a continuous random variable is exponentially distributed if and only if it is memoryless, so it suffices to show that T_x is memoryless. Fix $s, t \geq 0$, and observe that

$$\begin{aligned}\mathbb{P}(T_x > s+t|X_0 = x) &= \mathbb{P}(T_x > s+t \text{ and } T_x > s|X_0 = x)\\ &= \mathbb{P}(T_x > s+t|X_0 = x, T_x > s)\mathbb{P}(T_x > s|X_0 = x)\\ &= \mathbb{P}(T_x > s+t|X_u = x \text{ for all } u \in [0,s])\mathbb{P}(T_x > s|X_0 = x)\\ &\overset{1}{=} \mathbb{P}(T_x > s+t|X_s = x)\mathbb{P}(T_x > s|X_0 = x)\\ &\overset{2}{=} \mathbb{P}(T_x > t|X_0 = x)\mathbb{P}(T_x > s|X_0 = x),\end{aligned}$$

where $\overset{1}{=}$ follows from Markov's property and $\overset{2}{=}$ follows from time-homogeneity. The result follows on noting that

$$\mathbb{P}(T_x > s+t|X_0 = x, T_x > s) = \frac{\mathbb{P}(T_x > s+t)}{\mathbb{P}(T_x > s|X_0 = x)} = \mathbb{P}(T_x > t|X_0 = x).$$

□

Proposition 3.7 can be interpreted as follows: each time the Markov chain enters a state $x \in \mathscr{S}$, it stays there for a random amount of time. While the probability distribution of this amount of time is always exponential, the rate parameter will vary depending on which state the chain is in.

3.2. **Transition Rates.** In Section 3.1 we saw that if $\{X_t\}$ is a CTMC on $\mathscr{S}$, then each time $\{X_t\}$ reaches some $x \in \mathscr{S}$, it spends a random (exponentially distributed) amount of time there, before moving to another state. Additionally, we have seen that $\{X_t\}$ can be characterized through its transition function P, where

$$P_{x,y}(t) = \mathbb{P}(X_t = y|X_0 = x).$$

However, as we will see, it can be difficult to work with the transition function directly. In many cases, it will be more convenient to instead study the transition *rates*, which will allow us to simulate and easily construct Markov chains with particular properties.

To motivate this notion, suppose that $\{X_t\}$ starts in state $x \in \mathscr{S}$. Then, we saw that there is some $q_x \geq 0$ such that $T_x \sim \text{Exp}(q_x)$, which is to say that $\{X_t\}$ leaves x and moves to *some* other state after an exponentially long time. However, this characterization doesn't tell us which state $\{X_t\}$ jumps to.

Definition 3.8. *Let $\{X_t\}$ be a CTMC on $\mathscr{S}$. Then, for $x, y \in \mathscr{S}$, define*

$$P_{x,y} \doteq \mathbb{P}(X_{T_x} = y|X_0 = x).$$

Additionally we know that for each $x \in \mathscr{S}$, there is some $q_x \geq 0$ such that $T_x \sim Exp(q_x)$. For $x, y \in \mathscr{S}$ such that $x \neq y$, define

$$q_{x,y} \doteq q_x P_{x,y},$$

and let

$$q_{x,x} \doteq -\sum_{y \neq x} q_{x,y}.$$

We refer to the $|\mathscr{S}| \times |\mathscr{S}|$ matrix Q with entries

$$Q_{x,y} = q_{x,y}, \quad x, y \in \mathscr{S}.$$

*as a **generator matrix** or **infinitesimal generator**. The quantity $q_{x,y}$ is the **transition rate** from state x to state y, and the quantity*

$$q_x \doteq |q_{x,x}| = \sum_{y \neq x} q_{x,y},$$

*is the **holding time parameter** of state x (we will see why in Proposition 3.9).*

Definition 3.8 describes the generator matrix in terms of the underlying transition function of the CTMC. Rather than specifying the transition function of the CTMCs in which we are interested, we will generally specify its generator matrix instead. But it is (theoretically) possible to go between the two. It may not yet be obvious, but the generator matrix of a CTMC plays an analogous role to the transition matrix of a DTMC. The unusual formula for the diagonal entries $\{q_{x,x}\}_{x\in\mathscr{S}}$ will simplify many formulas and equations throughout our study of CTMCs. We begin by noting an important result.

Proposition 3.9. *Let $\{X_t\}$ be a CTMC on $\mathscr{S}$ with generator matrix Q. If we let T_x denote the holding time of state $x \in \mathscr{S}$, then, given that $X_0 = x$, if we let q_x be the holding time parameter of state x, defined as*

$$q_x \doteq |q_{x,x}| = \sum_{y\neq x} q_{x,y},$$

then

$$T_x \sim Exp(q_x).$$

Proposition 3.9 explains why we refer to q_x as the holding time parameter of state x; it measures the rate at which the state leaves state x.

It will sometimes be helpful to reduce our study of CTMCs to particular DTMCs associated with them.

Definition 3.10. *Let $\{X_t\}$ be a CTMC with generator matrix Q, and let $T_0 = 0$ and*

$$T_1 \doteq \inf\{t \geq 0 : X_t \neq X_0\},$$

denote the time of the chains first transition. Similarly, let

$$T_2 \doteq \inf\{t > T_1 : X_t \neq X_{T_1}\},$$

and in general, define

$$T_n \doteq \inf\{t > T_{n-1} : X_t \neq X_{T_{n-1}}\}.$$

Then the process $Y_n \doteq X_{T_n}$ is a DTMC known as the ***embedded chain****.*

The following proposition explains how to calculate the probability that a CTMC jumps from one state to another in terms of the generator matrix.

Proposition 3.11. *Let $\{X_t\}$ be a CTMC with generator matrix Q. The transition matrix of the embedded chain associated with $\{X_t\}$ is given by*

$$P_{x,y} = \begin{cases} \frac{q_{x,y}}{\sum_{z\neq x} q_{x,z}}, & x \neq y \\ 0 & x = y. \end{cases}$$

We will now see how to use the generator matrix to simulate a CTMC with transition function P.

Proposition 3.12. *Consider a transition function P, and recall the definitions of $\{P_{x,y}\}_{x,y\in\mathscr{S}}$ and $\{q_{x,y}\}_{x,y\in\mathscr{S}}$ from Definition 3.8. For each $x \in \mathscr{S}$, let*

$$\mathscr{A}_x \doteq \{y \in \mathscr{S} : P_{x,y} > 0\} = \{y \in \mathscr{S} : q_{x,y} > 0\}$$

denote the collection of states that are accessible from x. Let $X_0 = x$, and for each $y \in \mathscr{A}_x$, let Y_y^1 be an independent Exp$(q_{x,y})$ random variable. Let

$$T_1 \doteq \min\{Y_y^1 : y \in \mathscr{A}_{X_0}\},$$

and let

$$J_1 \doteq \arg\min_{y \in \mathscr{A}_{X_0}} Y_y^1.$$

For all $t < T_1$, let

$$X_t = x,$$

Now for each $y \in \mathscr{A}_{J_1}$, let Y_y^2 be an independent Exp$(q_{J_1,y})$ random variable. Let

$$T_2 \doteq \min\{Y_y^2 : y \in \mathscr{A}_{J_1}\},$$

and let

$$J_2 \doteq \arg\min_{y \in \mathscr{A}_{J_1}} Y_y^2.$$

For all $t \in [T_1, T_1 + T_2)$, let

$$X_t = J_1,$$

and let $X_{T_1} = J_2$. If we continue this process repeatedly, then the process $\{X_t\}$ is a CTMC with transition function P. Additionally, if we let $T_0 \doteq 0$, then the discrete-time process $\{X_{T_n}\}_{n \in \mathbb{N}_0}$ is the associated embedded chain.

In Figure 3.1 we use Proposition 3.12 to simulate a CTMC on $\mathscr{S} = \{0, 1, 2\}$ with generator matrix

$$Q = \begin{pmatrix} -1 & 0.5 & 0.5 \\ 0.5 & -1 & 0.5 \\ 0.5 & 1 & -1.5 \end{pmatrix} \tag{13}$$

over the time interval $[0, 10]$.

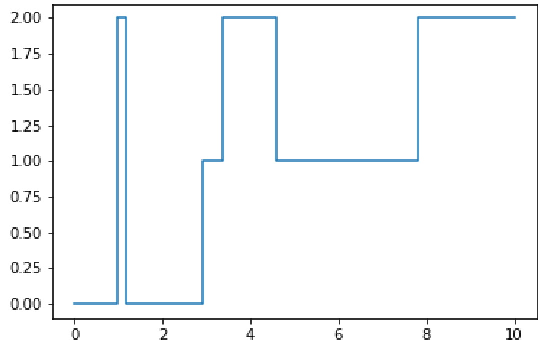


FIGURE 3.1. A realization of a CTMC on $\mathscr{S} = \{0, 1, 2\}$ with the generator matrix Q given in (13) over the time interval $[0, 10]$.

The Python code is given below.

```
import numpy as np
import matplotlib.pyplot as plt
import math
```

```
Q = np.matrix([[-1,0.5,0.5],[0.5,-1,0.5],[0.5,1,-0.5]])

def simCTMC(Q,T,x0):
    states = range(0,len(Q))
    XTrajectory = [x0]
    xt = x0
    times = [0]
    timeCount = 0
    while(timeCount <= T):
        jumpToRates = np.delete(Q[xt,:], xt)
        jumpToStates = np.delete(states,xt)
        minClock = math.inf
        for state in jumpToStates:
            clock = np.random.exponential(1/Q[xt,state])
            if(clock < minClock):
                minClock = clock
                minState = state
        timeCount = timeCount + minClock
        if(timeCount <= T):
            xt = minState
            XTrajectory.append(xt)
            times.append(timeCount)
        if(timeCount > T):
            XTrajectory.append(xt)
            times.append(T)
    plt.plot(times, XTrajectory, drawstyle=’steps-post’)

simCTMC(Q,10,0)
```

In general, even for simple examples, it can be difficult to calculate the transition function of a CTMC directly, so we often describe CTMCs in terms of their generator matrices instead. As we saw in Proposition 3.12, if we know how to calculate probabilities of transitions of the CTMC, then we can use the generator matrix to explicitly construct a CTMC with the desired transition function. In other words, every transition function has a generator matrix associated with it. We will soon see that the opposite is also true; every generator matrix gives rise to a transition function.

In general we will characterize CTMCs through their generator matrices, as this will be more tractable and more intuitive than working characterizing them through their transition functions.

3.3. **Some Examples of CTMCs.** Our discussion of CTMCs so far has been fairly abstract. It will be helpful to start by looking at a few examples. In each of these examples observe that it is much easier to specify the generator matrix of the CTMC than it is to specify the transition function. In the next section we will see how one can use the generator matrix of a CTMC to obtain its transition function.

Example 3.13. *The wifi at UCSB is either up or down at each time instant. If we let X_t denote the state of the wifi at time $t \geq 0$, then the state space of $\{X_t\}$ is $\{U, D\}$. When the wifi is up, the time that it stays up can be modeled as an Exp(μ) random variable. Similarly, when the wifi is down, the time that it stays down before it is repaired can be modeled as an Exp(λ) random variable. Additionally,* *the time that the wifi stays up or down is independent of its past history of being up or down.*

Here $\{X_t\}$ *is a CTMC with generator matrix*

$$Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}.$$

The idea behind the transition rates is that if $\lambda < \mu$ *(i.e., the wifi tends to stay down for longer periods than it stays up), then, in the long-term, the wifi will be down the majority of the time. We will return to this idea later when we discuss limiting and stationary distributions of CTMCs.*

The transition matrix of the embedded chain is given by

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

This makes sense, as the only transition from state U *is to state* D*, so the embedded chain* must *jump from state* U *to state* D*. Similarly, the only transition from state* D *is to state* U*, so the embedded chain must jump from* D *to* U*.*

The next example is slightly more complicated.

Example 3.14. *A machine has two components that must both be functioning in order for it to work properly. Each of the two components has a lifetime that follows an Exp*(λ) *distribution, and the two components' lifetimes are independent of each other. Each time a component fails, a repairperson immediately begins to fix that component. The time for the repairs to finish follows an Exp*(μ) *distribution. Note that the repairperson can only fix one machine at a time. Can we model this problem using a CTMC?*

Let X_t *denote the number of components that are functioning at time* $t \geq 0$*. The state space of* $\{X_t\}$ *is* $\{0, 1, 2\}$.

If both of the components are not functioning, then we know that the repair time follows an Exp(μ) *distribution, and that this is the only possible transition.*

If only one of the components is functioning, this means that the other component is being repaired (since repairs start immediately upon failure). Here there are two possible transitions: either the broken component is fixed, or the functioning component fails.

Finally, if both components are functioning, then the CTMC will move to state 1 *as soon as either one of the components fail. If we let* T_1 *and* T_2 *denote the time that it takes for the first and second component to fail, respectively, then we know that the CTMC moves to state* 1 *after* $\min\{T_1, T_2\}$ *time has passed. Recall from Corollary* 2.9 *that* $\min\{T_1, T_2\} \sim Exp(2\lambda)$.

Our observations tell us that $\{X_t\}$ *is a CTMC with generator matrix*

$$Q = \begin{pmatrix} -\mu & \mu & 0 \\ \lambda & -(\mu + \lambda) & \mu \\ 0 & 2\lambda & -2\lambda \end{pmatrix}$$

The transition matrix of the embedded chain is given by

$$P = \begin{pmatrix} 0 & 1 & 0 \\ \frac{\lambda}{\lambda+\mu} & 0 & \frac{\mu}{\lambda+\mu} \\ 0 & 1 & 0 \end{pmatrix}$$

Now we illustrate how Proposition 3.11 *can be applied; suppose that at time* $t = 10$, *exactly one of the components if functioning. What is the probability that this component is repaired before the other component malfunctions?*

We want to calculate the probability that, given that at time $t = 10$, *the chain is in state* 1, *that it next jumps to state* 2. *If we let, for* $s \geq 0$,

$$T_1(s) \doteq \inf\{t \geq 0 : X_{s+t} \neq 1\},$$

denote the first time after time s that the chain leaves state 1, *then we have, from Markov's property and Proposition* 3.11, *that*

$$\mathbb{P}(X_{T_1(10)} = 2 | X_{10} = 1) = \mathbb{P}(X_{T_1(0)} = 2 | X_0 = 1) = P_{1,2} = \frac{\lambda}{\lambda + \mu}.$$

The following example is known as a **birth and death process**.

Example 3.15. *Let* X_t *denote the size of some population at time* $t \geq 0$. *If* $X_t = x$ *for some* $x \in \mathbb{N}_0$, *then two things can happen: either a new individual is born (which increases the size of the population by 1) or an individual dies (which decreases the size of the population by 1). If the size of the population is* x, *the rate at which an individual is born is* λ_x, *and the rate at which an individual dies is* μ_x. *If there are no individuals left in the population, another individual is born at a rate of* λ_0.

What is the generator matrix of $\{X_t\}$*? What is the transition matrix of the embedded chain?*

The generator matrix is given by

$$Q_{x,y} = \begin{cases} \lambda_x & y = x+1, x \geq 0 \\ \mu_x & y = x-1, x \geq 1 \\ -(\lambda_x + \mu_x) & x = y \geq 1 \\ -\lambda_0 & x = y = 0. \end{cases}$$

or

$$Q_{x,y} = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\mu_1 + \lambda_1) & \lambda_1 & & \\ & \mu_2 & -(\mu_2 + \lambda_2) & \lambda_2 & \\ & & \mu_3 & -(\mu_3 + \lambda_3) & \lambda_3 \\ & & & & \ddots \end{pmatrix}$$

The transition matrix of the embedded chain is given by

$$P_{x,y} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ \frac{\mu_1}{\mu_1+\lambda_1} & 0 & \frac{\lambda_1}{\mu_1+\lambda_1} & 0 & 0 & 0 & \cdots \\ 0 & \frac{\mu_2}{\mu_2+\lambda_2} & 0 & \frac{\lambda_2}{\mu_2+\lambda_2} & 0 & 0 & \cdots \\ 0 & 0 & \frac{\mu_3}{\mu_3+\lambda_3} & 0 & \frac{\lambda_3}{\mu_3+\lambda_3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

This says that, for example, if at time t there are 25 individuals in the population, then the probability that an individual dies before another is born is given by

$$\mathbb{P}(X_{T_{25}(t)} = 24 | X_t = 25) = \frac{\mu_{25}}{\mu_{25}+\lambda_{25}},$$

where, as in Example 3.14, we denote by

$$T_{25}(t) \doteq \inf\{s \geq 0 : X_{t+s} \neq 25\},$$

the first time after time t that the population changes size.

3.4. **Forward and Backward Equations.** In the previous section we saw the generator matrices are well-suited to describe continuous time Markov chains. We begin by recalling the notion of a differentiable matrix-valued function.

Definition 3.16. *Consider a transition function P on state space $\mathscr{S}$. We say that P is differentiable at t if, for all $x, y \in \mathscr{S}$, the function $t \mapsto P_{x,y}(t)$ is differentiable with respect to t. For each $t \geq 0$, we write $P'(t)$ to denote the $|\mathscr{S}| \times |\mathscr{S}|$ matrix with entries $P'_{x,y}(t)$, where*

$$P'_{x,y}(t) \doteq \frac{d}{dt} P_{x,y}(t).$$

The next result explains the connection between the generator matrix and the transition function of a CTMC.

Theorem 3.17. *Let P be the transition probability matrix of a CTMC with state space $\mathscr{S} = \{0, 1, 2, \ldots\}$, and let Q be the corresponding generator matrix. Then P is differentiable, and Q and P satisfy*

$$P'(t) = QP(t), \quad t \geq 0, \tag{14}$$

and

$$P'(t) = P(t)Q, \quad t \geq 0, \tag{15}$$

with initial condition

$$P(0) = I,$$

where I denote the $|\mathscr{S}| \times |\mathscr{S}|$ identity matrix. The equations in (14) are known as the ***backward equations****, and the equations in (15) are the* ***forward equations****.*

Note that we can rewrite (14) component-wise as

$$P'_{x,y}(t) = \sum_{z \in \mathscr{S}} Q_{x,z} P_{z,y}(t), \quad x, y \in \mathscr{S}, t \geq 0$$

and we can rewrite (15) *component-wise as*

$$P'_{x,y}(t) = \sum_{z\in\mathscr{S}} P_{x,z}(t)Q_{z,y}, \quad x,y\in\mathscr{S}, t\geq 0.$$

Additionally, if the state space $\mathscr{S}$ is finite, then the solution to the backward and forward conditions is unique.

Theorem 3.17 tells us that if we know the generator matrix of a CTMC, that we can solve these differential equations to calculate its transition function.

Example 3.18. *Recall the setting from Example* 3.13*, where we modeled the state of the UCSB wifi as a CTMC with state space $\mathscr{S} = \{u,d\}$ and generator matrix*

$$Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}.$$

If we let X_t denote the state of the wifi at time $t\geq 0$, we can use Theorem 3.17 *to calculate the transition function of $\{X_t\}$. First note that the forward equations are given by*

$$\begin{aligned} P'_{u,u}(t) &= Q_{u,u}P_{u,u}(t) + Q_{u,d}P_{d,u}(t) = \mu(P_{d,u}(t) - P_{u,u}(t)) \\ P'_{u,d}(t) &= Q_{u,u}P_{u,d}(t) + Q_{u,d}P_{d,d}(t) = \mu(P_{d,d}(t) - P_{u,d}(t)) \\ P'_{d,u}(t) &= Q_{d,d}P_{d,u}(t) + Q_{d,u}P_{u,u}(t) = \lambda(P_{u,u}(t) - P_{d,u}(t)) \\ P'_{d,d}(t) &= Q_{d,d}P_{d,d}(t) + Q_{d,u}P_{d,u}(t) = \lambda(P_{d,u}(t) - P_{d,d}(t)), \end{aligned} \tag{16}$$

with initial condition

$$P_{u,u}(0) = P_{d,d}(0) = 1, \quad P_{u,d}(0) = P_{d,u}(0) = 0.$$

However, we also know that

$$P_{u,u}(t) + P_{u,d}(t) = P_{d,d}(t) + P_{d,u}(t) = 1,$$

so it suffices to solve

$$\begin{aligned} P'_{u,u}(t) &= \mu(P_{d,u}(t) - P_{u,u}(t)) \\ P'_{d,u}(t) &= \lambda(P_{u,u}(t) - P_{d,u}(t)) \end{aligned} \tag{17}$$

Note that (17) *tells us that*

$$\lambda P'_{u,u}(t) + \mu P'_{d,u}(t) = \lambda\mu(P_{d,u}(t) - P_{u,u}(t)) + \mu\lambda(P_{u,u}(t) - P_{d,u}(t)) = 0,$$

which tells us that the function $t\mapsto \lambda P_{u,u}(t) + \mu P_{d,u}(t)$ is constant. In particular, there is some constant $c\in\mathbb{R}$ such that

$$\lambda P_{u,u}(t) + \mu P_{d,u}(t) = c, \quad t\geq 0. \tag{18}$$

Using our initial condition $P_{u,u}(0) = 1$ and $P_{d,u}(0) = 0$, we see that

$$c = \lambda\cdot 1 + \mu\cdot 0 = \lambda. \tag{19}$$

Combining (18) *and* (19)*, we see that*

$$\mu P_{d,u}(t) = \lambda(1 - P_{u,u}(t)), \tag{20}$$

which, when combined with (17)*, yields*

$$P'_{u,u}(t) = \lambda(1 - P_{u,u}(t)) - \mu P_{u,u}(t) = \lambda - (\lambda+\mu)P_{u,u}(t).$$

Once more using the initial condition $P_{u,u}(0) = 1$, we see that the solution is given by

$$P_{u,u}(t) = \frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}. \tag{21}$$

Now, combine (20) *and* (21) *to see that*

$$\begin{aligned} P_{d,u}(t) &= \frac{\lambda}{\mu}\left(1-\frac{\lambda}{\lambda+\mu}-\frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}\right) \\ &= \frac{\lambda}{\mu}\left(\frac{\mu}{\lambda+\mu}-\frac{\mu}{\lambda+\mu}e^{-(\lambda+\mu)t}\right) \\ &= \frac{\lambda}{\lambda+\mu}-\frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t} \end{aligned}$$

The approach using the backward equations is somewhat simpler and is left as an exercise.

3.5. **Computing Transition Functions with the Generator Matrix.** Rather than directly solving the forward or backward equations, we will establish a general method for computing the solutions of such systems of differential equations. To motivate this approach, note that for a real-valued function $p : \mathbb{R}_+ \to \mathbb{R}_+$, the unique solution to the ODE

$$\begin{aligned} p'(t) &= qp(t) \\ p(0) &= 1, \end{aligned}$$

is given by $p(t) = e^{tq}$. The ODE above looks very similar to the backward equations, which are given by

$$\begin{aligned} P'(t) &= QP(t) \\ P(0) &= I. \end{aligned} \tag{22}$$

As we see below, if we define an appropriate notion of the *matrix exponential*, then we can express the solution to (22) in such terms. To motivate the defnition of the matrix exponential, recall that for each $x \in \mathbb{R}$,

$$e^x \doteq \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Definition 3.19. *Let A be a d × d matrix. The* ***matrix exponential*** *of A, which is the d × d matrix denoted as* e^A *or* $\exp(A)$*, is defined as*

$$\exp(A) \doteq \sum_{k=0}^{\infty} \frac{A^k}{k!},$$

where $A^0 \doteq I$.

The next result provides an affirmative answer to our conjecture regarding the solution to (22).

Proposition 3.20. *Let* $\{X_t\}$ *be a CTMC on a finite state space* $\mathscr{S}$ *with generator matrix Q be the corresponding generator matrix. Then the transition function for* $\{X_t\}$ *is given by*

$$P(t) = \exp(Qt).$$

Proof. Recall from Theorem 3.17 that since $\mathscr{S}$ is finite, the transition function of $\{X_t\}$ is the unique solution to (14) and (15). Thus, it is enough to show that $P(t) \doteq \exp(Qt)$ satisfies those equations. First, note that

$$e^{Qt} = I + \sum_{k=1}^{\infty} \frac{(Qt)^k}{k!}, \quad t \geq 0,$$

and that the series above is uniformly convergent, so

$$\frac{d}{dt}e^{Qt} = \frac{d}{dt}\left(I + \sum_{k=1}^{\infty}\frac{(Qt)^k}{k!}\right) = \sum_{k=1}^{\infty}\frac{d}{dt}\left(\frac{Q^k}{k!}t^k\right) = \sum_{k=1}^{\infty}\frac{Q^k}{k!}kt^{k-1} = Q\sum_{k=1}^{\infty}\frac{Q^{k-1}}{(k-1)!}t^{k-1} = Qe^{Qt},$$

which shows (14). If we instead factored $Q^k = Q^{k-1}Q$, then we would obtain

$$\frac{d}{dt}e^{Qt} = e^{Qt}Q,$$

so (15) holds as well. Therefore, $P(t) = e^{Qt}$ is the unique solution to the forward and backward equations with initial condition $P(0) = I$, and is the transition function of $\{X_t\}$. □

While this tells us how to calculate the transition function of a CTMC with a particular generator matrix, in practice it can be numerically difficult to calculate matrix exponentials. However, for certain types of matrices the calculation can be simplified greatly. Let us recall the notion of diagonalizable matrices.

3.5.1. *Diagonalizable Matrices and Related Concepts.*

Definition 3.21. *A $d \times d$ matrix D is said to be* ***diagonal*** *if its off-diagonal entries are 0. If D is diagonal and*

$$D_{i,j} = \begin{cases} \lambda_i & i = j \\ 0, & i \neq j, \end{cases}$$

then we write $D = diag(\lambda_1, \dots, \lambda_d)$. A $d \times d$ matrix A is said to be ***diagonalizable*** *if there is an invertible matrix U and a diagonal matrix D such that*

$$A = UDU^{-1}.$$

Observe that if $D = diag(\lambda_1, \dots, \lambda_d)$, then its matrix exponential is given by

$$\exp(D) = diag(e^{\lambda_1}, \dots, e^{\lambda_d}).$$

We now recall some important definitions from linear algebra.

Definition 3.22. *We say a collection of vectors $\{\boldsymbol{v}_i\}_{i=1}^n$ in $\mathbb{R}^d$ is* ***linearly independent*** *if the only constants such that*

$$\sum_{i=1}^{n}\alpha_i \boldsymbol{v}_i = \mathbf{0} \in \mathbb{R}^d,$$

are $\alpha_1 = \cdots = \alpha_n = 0$. Recall also that a non-zero vector $\boldsymbol{v} \in \mathbb{R}^d$ is an ***eigenvector*** *for an $d \times d$ matrix M if there is some λ such that*

$$M\boldsymbol{v} = \lambda\boldsymbol{v}.$$

The constant λ is an ***eigenvalue*** *of M, and the pair $(\boldsymbol{v}, \lambda)$ is an* ***eigenpair*** *of M.*

The following proposition explains how to diagonalize a matrix.

Proposition 3.23. *Let M be a $d \times d$ matrix with d linearly independent eigenvectors, denoted by $\boldsymbol{v}_1, \dots, \boldsymbol{v}_d$. Denote the corresponding eigenvalues by $\lambda_1, \dots, \lambda_d$. Then M can be diagonalized as*

$$M = UDU^{-1},$$

where $D = diag(\lambda_1, \dots, \lambda_d)$ and $U = \begin{pmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_d \end{pmatrix}$.

Proof. The result follows on showing that $MU = UD$. Observe that

$$MU = \begin{pmatrix} M\boldsymbol{v}_1 & \dots & M\boldsymbol{v}_d \end{pmatrix} = \begin{pmatrix} \lambda_1\boldsymbol{v}_1 & \dots & \lambda_d\boldsymbol{v}_d \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_d \end{pmatrix} \operatorname{diag}(\lambda_1, \dots, \lambda_d) = UD,$$

so M can be diagonalized as claimed. □

We begin by demonstrating how to diagonalize a 2 × 2 matrix.

Example 3.24. *Let*

$$M \doteq \begin{pmatrix} 3 & 2 \\ 5 & 1 \end{pmatrix}.$$

We first calculate the eigenvalues of M by solving for all $\lambda \in \mathbb{R}$ such that

$$det(M - \lambda I) = det\left(\begin{pmatrix} 3-\lambda & 2 \\ 5 & 1-\lambda \end{pmatrix}\right) = (3-\lambda)(1-\lambda) - 10 = 0.$$

Using the quadratic formula, we see that

$$\lambda_1 = 2 - \sqrt{11}, \quad \lambda_2 = 2 + \sqrt{11}.$$

We now solve for the eigenvectors corresponding to these eigenvalues; first we note that if $\boldsymbol{v}_i$ is an eigenvector corresponding to the eigenvalue λ_i, then

$$(M - \lambda_i I)\boldsymbol{v}_i = M\boldsymbol{v}_i - \lambda_i I\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i - \lambda_i\boldsymbol{v}_i = \mathbf{0},$$

so it suffices to solve the above equation for $i = 1, 2$. Denoting the j^{th} entry of $\boldsymbol{v}_i$ by v_i^j, this means that we need to solve

$$\begin{pmatrix} 3-\lambda_1 & 2 \\ 5 & 1-\lambda_1 \end{pmatrix}\begin{pmatrix} v_1^1 \\ v_1^2 \end{pmatrix} = \begin{pmatrix} 1+\sqrt{11} & 2 \\ 5 & -1+\sqrt{11} \end{pmatrix}\begin{pmatrix} v_1^1 \\ v_1^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and we can see that a nonzero solution is given by

$$\begin{pmatrix} v_1^1 \\ v_1^2 \end{pmatrix} = \begin{pmatrix} \frac{1-\sqrt{11}}{5} \\ 1 \end{pmatrix}.$$

Similarly, a nonzero solution to

$$\begin{pmatrix} 3-\lambda_2 & 2 \\ 5 & 1-\lambda_2 \end{pmatrix}\begin{pmatrix} v_2^1 \\ v_2^2 \end{pmatrix} = \begin{pmatrix} 1-\sqrt{11} & 2 \\ 5 & -1-\sqrt{11} \end{pmatrix}\begin{pmatrix} v_2^1 \\ v_2^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

is given by

$$\begin{pmatrix} v_2^1 \\ v_2^2 \end{pmatrix} = \begin{pmatrix} \frac{1+\sqrt{11}}{5} \\ 1 \end{pmatrix}.$$

Then, as in Proposition 3.23*, if we let*

$$U \doteq \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{pmatrix} = \begin{pmatrix} \frac{1-\sqrt{11}}{5} & \frac{1+\sqrt{11}}{5} \\ 1 & 1 \end{pmatrix}$$

and

$$D \doteq diag(\lambda_1, \lambda_2) = \begin{pmatrix} 2-\sqrt{11} & 0 \\ 0 & 2+\sqrt{11} \end{pmatrix},$$

we have

$$U^{-1} = -\frac{5}{2\sqrt{11}}\begin{pmatrix} 1 & -\frac{1+\sqrt{11}}{5} \\ -1 & \frac{1-\sqrt{11}}{5} \end{pmatrix},$$

and we can check that $M = UDU^{-1}$.

We now outline how to diagonalize a 3 × 3 matrix.

Example 3.25. *Let*

$$Q \doteq \begin{pmatrix} -2 & 2 & 0 \\ 2 & -3 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

We begin by calculating the eigenvalues of Q, namely by solving for all $\lambda \in \mathbb{R}$ such that

$$det(Q - \lambda I) = det\left(\begin{pmatrix} -2-\lambda & 2 & 0 \\ 2 & -3-\lambda & 1 \\ 1 & 0 & -1-\lambda \end{pmatrix}\right) = 0.$$

Note that the determinant of $Q - \lambda I$ is given by

$$\begin{aligned} det(Q - \lambda I) &= (-2-\lambda)\,[(-3-\lambda)(-1-\lambda) - 1\cdot 0] - 2\,[(2(-1-\lambda)) - 1\cdot 1] + 0\,[2\cdot 0 - (-3-\lambda)\cdot 1] \\ &= (-2-\lambda)(3+\lambda)(1+\lambda) + 4(1+\lambda) + 2 \\ &= -\lambda^3 - 6\lambda^2 - 7\lambda \\ &= -\lambda(\lambda^2 + 6\lambda + 7) \\ &= \lambda(-\lambda + \sqrt{2} - 3)(\lambda + \sqrt{2} + 3), \end{aligned}$$

where the final line is obtained by applying the quadratic formula to $\lambda^2 + 6\lambda + 7$. It follows that the eigenvalues of Q are given by

$$\lambda_1 \doteq 0, \quad \lambda_2 = \sqrt{2} - 3, \quad \lambda_3 = -\sqrt{2} - 3.$$

We now solve for the eigenvectors corresponding to these eigenvalues; first we note that if $\boldsymbol{v}_i$ is an eigenvector corresponding to the eigenvalue λ_i, then

$$(Q - \lambda_i I)\boldsymbol{v}_i = Q\boldsymbol{v}_i - \lambda_i I \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i - \lambda_i \boldsymbol{v}_i = \mathbf{0},$$

so it suffices to solve the above equation for $i = 1, 2, 3$. Denoting the j^{th} entry of $\boldsymbol{v}_i$ by v_i^j, this means that we need to solve

$$(Q - \lambda_i I)\boldsymbol{v}_i = \begin{pmatrix} -2-\lambda_i & 2 & 0 \\ 2 & -3-\lambda_i & 1 \\ 1 & 0 & -1-\lambda_i \end{pmatrix} \begin{pmatrix} v_i^1 \\ v_i^2 \\ v_i^3 \end{pmatrix}.$$

This can be simplified by rewriting $Q - \lambda_i I$ in reduced row echelon form, then solving the resulting system of equations. After solving for $\boldsymbol{v}_1, \boldsymbol{v}_2$, and $\boldsymbol{v}_3$, the diagonalization of Q is given by $Q = UDU^{-1}$, where

$$U = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 \end{pmatrix}, \quad D = diag(\lambda_1, \lambda_2, \lambda_3).$$

The following result will simplify some calculations involving matrix exponentials; it says that if a matrix can be diagonalized, then we can easily calculate its matrix exponential.

Proposition 3.26. *Let A be a diagonalizable matrix of the form*

$$A = UDU^{-1},$$

Then

$$\exp(A) = U \exp(D) U^{-1}.$$

Proof. Observe that for each $k \in \mathbb{N}$,

$$(UDU^{-1})^k = UD^kU^{-1},$$

so

$$\exp(A) = \sum_{k=0}^{\infty} \frac{(UDU^{-1})^k}{k!} = \sum_{k=0}^{\infty} \frac{UD^kU^{-1}}{k!} = U\left(\sum_{k=0}^{\infty} \frac{D^k}{k!}\right)U^{-1} = U\exp(D)U^{-1}.$$

□

3.5.2. *Computing Transition Functions.* We now re-derive the transition function from Example 3.18 using this result.

Example 3.27. *Consider the CTMC with state space $\mathscr{S} = \{u, d\}$ and generator matrix*

$$Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}.$$

As shown in Example 3.24, *we can write $Q = UDU^{-1}$, where*

$$U = \begin{pmatrix} 1 & \frac{\mu}{\mu+\lambda} \\ 1 & -\frac{\lambda}{\mu+\lambda} \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 \\ 0 & -(\mu+\lambda) \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \\ 1 & -1 \end{pmatrix}.$$

Observe that

$$P(t) = U\exp(Dt)U^{-1} = \begin{pmatrix} 1 & \frac{\mu}{\mu+\lambda} \\ 1 & -\frac{\lambda}{\mu+\lambda} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & e^{-(\mu+\lambda)t} \end{pmatrix}\begin{pmatrix} \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \\ 1 & -1 \end{pmatrix}$$

In the example below we use diagonalization to find the transition function of the CTMC from Example 3.14

Example 3.28. *A machine has two components that must both be functioning in order for it to work properly. Each of the two components has a lifetime that follows an Exp*(7) *distribution, and the two components' lifetimes are independent of each other. Each time a component fails, a repairperson immediately begins to fix that component. The time for the repairs to finish follows an Exp*(5) *distribution. Note that the repairperson can only fix one machine at a time. Let X_t denote the number of machines functioning at time t. Recall that the generator matrix is given by*

$$Q = \begin{pmatrix} -5 & 5 & 0 \\ 7 & -12 & 5 \\ 0 & 14 & -14 \end{pmatrix}$$

Using numerical techniques, one can verify that Q can be diagonalized as $Q = UDU^{-1}$, where

$$U = \begin{pmatrix} 1 & -35-15\sqrt{21} & 5\left(3\sqrt{21}-7\right) \\ 1 & -7\left(3-3\sqrt{21}\right) & -7\left(3+3\sqrt{21}\right) \\ 1 & 196 & 196 \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} \frac{98}{193} & \frac{70}{193} & \frac{25}{193} \\ -\frac{63+31\sqrt{21}}{48636} & \frac{11\sqrt{21}-15}{16212} & \frac{54-\sqrt{21}}{24318} \\ \frac{31\sqrt{21}-63}{48636} & -\frac{15+11\sqrt{21}}{16212} & \frac{\sqrt{21}+54}{24318} \end{pmatrix},$$

and

$$D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{-31+3\sqrt{21}}{2} & 0 \\ 0 & 0 & \frac{-31-3\sqrt{21}}{2} \end{pmatrix},$$

From Proposition 3.20 *and Proposition* 3.26*, we see that the transition function of the CTMC is given by, at time* $t = 0.1$*,*

$$P(0.1) = U\exp(0.1D)U^{-1} = \begin{pmatrix} 0.6982273 & 0.2525381 & 0.04923469 \\ 0.3535533 & 0.4825311 & 00.16391562 \\ 0.1930000 & 0.4589637 & 0.34803629 \end{pmatrix}.$$

This tells us that if both components are functioning at time $t = 0$*, then there is a 34.8% chance that they are both functioning at time* $t = 0.1$ *as well.*

3.6. **Poisson Subordination.** In Proposition 3.12 we saw that if $\{X_t\}$ is a CTMC with generator matrix Q, if we can compute the transition probabilities

$$P_{x,y} \doteq \mathbb{P}(X_{T_x} = y | X_0 = x),$$

where $T_x \doteq \inf\{t \geq 0 : X_t \neq x\}$, is the holding time of state x, then we can simulate the chain by generating independent exponential random variables.

We begin by showing how we can construct a CTMC using a DTMC and a Poisson process. First we show that if we add together a geometrically distributed number of iid exponentially distributed random variables, the sum is also exponentially distributed.

Lemma 3.29. *Let* $\{E_m\} \overset{iid}{\sim}$ *Exp*(λ) *be independent of* $X \sim$ *Geom*(p)*, and define*

$$S_m \doteq \sum_{k=1}^{m} E_k.$$

Then $S_X \sim$ *Exp*(λp)*.*

Proof. Recall that the mgf of E_k is

$$\mathbb{E}[e^{E_k t}] = \frac{\lambda}{\lambda - t}, \quad t < \lambda,$$

so the conditional mgf of S_X given that $X = m$ can be computed as

$$\mathbb{E}\left[e^{tS_X} | X = m\right] = \mathbb{E}\left[e^{t\sum_{k=1}^{m} E_k} \middle| X = m\right] = \mathbb{E}\left[e^{t\sum_{k=1}^{m} E_k}\right] = \prod_{k=1}^{m} \mathbb{E}[e^{tE_k}] = \left(\frac{\lambda}{\lambda - t}\right)^m.$$

In the previous calculation, the second equality is due to the independence of $\{E_k\}$ and X and the third is due to the independence of the $\{E_k\}$. Using the law of total expectation, we see that, for $t < p\lambda$,

$$\mathbb{E}[e^{tS_X}] = \sum_{m=1}^{\infty} \mathbb{E}[e^{tS_m} | X = m]\mathbb{P}(X = m) = \sum_{m=1}^{\infty} \left(\frac{\lambda}{\lambda - t}\right)^m (1-p)^{m-1} p = \frac{p\lambda}{\lambda - t} \sum_{m=0}^{\infty} \left(\frac{(1-p)\lambda}{\lambda - t}\right)^m = \frac{p\lambda}{p\lambda - t},$$

where the final equality follows from using the geometric series formula. Since mgfs uniquely determine probability distributions, it follows that $S_X \sim \text{Exp}(p\lambda)$. □

The next theorem illustrates how we can use a DTMC and a Poisson process to construct a CTMC with a particular generator matrix.

Theorem 3.30. *Let* $\{Y_n\}$ *be a DTMC on a finite state space* $\mathscr{S} = \{1, \dots, d\}$ *with transition matrix* P*. Let* $\{N_t\}$ *be a Poisson process with rate* λ *that is independent of* $\{Y_n\}$*. Let*

$$X_t \doteq Y_{N_t}, \quad t \geq 0.$$

Then $\{X_t\}$ is a CTMC on $\mathscr{S}$ with generator matrix

$$Q \doteq \lambda(P - I),$$

where I denotes the $d \times d$ identity matrix. We say that $\{X_t\}$ is ***subordinated to a Poisson process****.*

Proof. Let τ_x denote the (integer-valued) holding time of the DTMC in state x, so that

$$\tau_x \doteq \inf\{n \geq 0 : Y_n \neq x\}.$$

Note that $\tau_x = m$ if and only if $\{Y_n\}$ stays in state x for $m-1$ time instants, then moves to another state. The probability that $\{Y_n\}$ jumps from x to x in a single time instant is $P_{x,x}$, and the probability that it jumps to another state from x is $1 - P_{x,x}$, so, using Markov's property, we can see that

$$\mathbb{P}(\tau_x = m | Y_0 = x) = P_{x,x}^{m-1}(1 - P_{x,x}).$$

This says that, given that $X_0 = x$,

$$\tau_x \sim \text{Geom}(1 - P_{x,x}). \tag{23}$$

Now, let

$$T_x \doteq \inf\{t \geq 0 : X_t \neq x\},$$

and note that we can write

$$T_x = S_{\tau_x}, \quad S_m \doteq \sum_{k=1}^{m} E_k \tag{24}$$

where

$$\begin{aligned} E_1 &\doteq \inf\{t \geq 0 : N_t = 1\}, \\ E_m &\doteq \inf\left\{t \geq 0 : N_{t+S_{m-1}} = N_{S_{m-1}} + 1\right\}, \quad m \geq 2. \end{aligned}$$

Recall from Definition 2.16 that $E_m \overset{iid}{\sim} \text{Exp}(\lambda)$, which, when combined with (23), (24), and Lemma 3.29, tells us that, given that $Y_0 = x$,

$$T_x \sim \text{Exp}(\lambda(1 - P_{x,x})). \tag{25}$$

Observe also that, for all $y \neq x$,

$$\mathbb{P}(Y_{\tau_x} = y | Y_0 = x) = \frac{P_{x,y}}{1 - P_{x,x}} \tag{26}$$

and $P(Y_{\tau_x} = x | Y_0 = x) = 0$. Define

$$q_{x,y} \doteq \begin{cases} \lambda P_{x,y} & x \neq y \\ -\lambda(1 - P_{x,x}) & x = y, \end{cases} \tag{27}$$

and let $q_x \doteq -q_{x,x} = \lambda(1 - P_{x,x})$. Together (25) and (26) show that, given that $X_0 = x$, the time that $\{X_t\}$ stays in state x follows an $\text{Exp}(q_x)$ distribution, and then, at that point, the probability that $\{X_t\}$ jumps to state y is given by $q_{x,y}$. This tells us that the generator matrix of $\{X_t\}$ is given by

$$Q_{x,y} = \begin{cases} q_{x,y} & x \neq y \\ q_x & x = y. \end{cases}$$

Recalling from (27) the definition of $q_{x,y}$, we see that $Q = \lambda(P - I)$, as claimed. □

Theorem 3.30 showed that if we have a DTMC with transition matrix P, then by simulating a Poisson process with rate λ, we can construct a CTMC with generator matrix $Q = \lambda(P - I)$. The next result shows how we can leverage this to compute the transition function of a CTMC in terms of the transition matrix of a related DTMC.

Theorem 3.31. *Let $\{X_t\}$ be a CTMC on $\mathscr{S} = \{1, \ldots, d\}$ with generator matrix Q. Let*

$$\lambda \doteq \max_{1 \le x \le d} |q_{x,x}|,$$

and define

$$P \doteq I + \frac{Q}{\lambda}.$$

Then the transition function $P(t)$ of $\{X_t\}$ is given by

$$P(t) = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n, \quad t \ge 0.$$

Proof. We begin by checking that P is stochastic. Note that, for each $x \neq y \in \mathscr{S}$, $\lambda \ge q_{x,y} \ge 0$, and $0 \ge q_{x,x} \ge -\lambda$, and

$$P_{x,y} = \begin{cases} \frac{q_{x,y}}{\lambda} & x \neq y \\ 1 + \frac{q_{x,x}}{\lambda} & x = y, \end{cases}$$

so $0 \le P_{x,y} \le 1$. Additionally, for each $x \in \mathscr{S}$,

$$q_{x,x} = -\sum_{y \neq x} q_{x,y},$$

so

$$\sum_{y \in \mathscr{S}} P_{x,y} = P_{x,x} + \sum_{y \neq x} P_{x,y} = 1 + \frac{q_{x,x}}{\lambda} + \sum_{y \neq x} \frac{q_{x,y}}{\lambda} = 1 - \sum_{y \neq x} q_{x,x}\lambda + \sum_{y \neq x} q_{x,x}\lambda = 1,$$

so P is a stochastic matrix. Now, let $\{Y_n\}$ be a DTMC with transition matrix P, and let $\{N_t\} \sim \text{PP}(\lambda)$ be independent of $\{Y_n\}$. Theorem 3.30 tells us that, with $X_t \doteq Y_{N_t}\}$, $\{X_t\}$ is a CTMC with generator matrix Q, as

$$\lambda(P - I) = \lambda\left(I + \frac{Q}{\lambda} - I\right) = Q.$$

For $x, y \in \mathscr{S}$ and $t \ge 0$,

$$\begin{aligned} P_{x,y}(t) &= \mathbb{P}(X_t = y | X_0 = x) \\ &= \mathbb{P}(Y_{N_t} = y | Y_0 = x) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(Y_{N_t} = y | N_t = n, Y_0 = x)\mathbb{P}(N_t = n | Y_0 = x) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(Y_n = y | N_t = n, Y_0 = x)\mathbb{P}(N_t = n) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(Y_n = y | Y_0 = x)\mathbb{P}(N_t = n) \\ &= \sum_{n=0}^{\infty} (P^n)_{x,y} e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \end{aligned}$$

where the final equality is due to Proposition E.7. □

Theorem 3.31 illustrates how we one can calculate the transition function of a given CTMC on a finite state space. However, in practice it may be difficult to obtain a closed expression for the infinite series

$$\sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n,$$

so one typically truncates the series after a finite number of terms to estimate the probability. In doing so, one obtains the estimate

$$\hat{P}[M](t) \doteq \sum_{n=0}^{M} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n.$$

By choosing a larger value of M, we can improve our approximation. The next result shows us how to choose M large enough so that our approximation is as accurate as we would like. Note that this truncation is only guaranteed to be this accurate for the specific time instant t we have chosen.

Proposition 3.32. *Let $\{X_t\}$ be a CTMC on $\mathscr{S} = \{1, \dots, d\}$ with generator matrix Q. Let*

$$\lambda \doteq \max_{1 \le x \le d} |q_{x,x}|,$$

and define

$$P \doteq I + \frac{Q}{\lambda}.$$

For each $M \in \mathbb{N}$, let

$$\hat{P}[M](t) = \sum_{n=0}^{M} e^{-\lambda t} \frac{(\lambda t)^n}{n!} P^n, \quad t \ge 0.$$

Fix $\epsilon > 0$, and let $N^t \sim PP(\lambda t)$. If we choose $M \in \mathbb{N}$ large enough so that

$$\mathbb{P}(N^t > M) \le \epsilon,$$

then, for all $s \in [0, t]$,

$$|P_{x,y}(s) - \hat{P}[M]_{x,y}(s)| \le \epsilon, \quad x, y \in \mathscr{S}.$$

Proof. Observe that for each $x, y \in \mathscr{S}$, and $s \in [0, t]$,

$$\begin{aligned} |P_{x,y}(s) - \hat{P}[M]_{x,y}(s)| &= \left| \sum_{k=0}^{\infty} (P^k)_{x,y} e^{-\lambda s} \frac{(\lambda s)^k}{k!} - \sum_{k=0}^{M} (P^k)_{x,y} e^{-\lambda s} \frac{(\lambda s)^k}{k!} \right| \\ &= \sum_{k=M+1}^{\infty} (P_{x,y})^k e^{-\lambda s} \frac{(\lambda s)^k}{k!} \\ &\le \sum_{k=M+1}^{\infty} e^{-\lambda s} \frac{(\lambda s)^k}{k!} \\ &= \mathbb{P}(N^s > M) \\ &\le \mathbb{P}(N^t > M), \end{aligned}$$

where the final inequality used the fact that $s \mapsto \mathbb{P}(N^s > M)$ is increasing in s. □

The next example illustrates how we can apply this result.

Example 3.33. *Let $\{X_t\}$ be a CTMC on $\mathscr{S} = \{1, 2, 3\}$ with generator matrix*

$$Q = \begin{pmatrix} -1 & 0.5 & 0.5 \\ 0 & -3 & 3 \\ 2 & 2 & -4 \end{pmatrix}$$

Let P denote the transition function of $\{X_t\}$ and find an estimate $\hat{P}(2)$ for $P(2)$ such that

$$|P_{x,y}(2) - \hat{P}_{x,y}(2)| < 0.001, \quad x, y \in \mathscr{S}.$$

Let $\lambda \doteq 4$ and define

$$P \doteq I + \frac{Q}{\lambda} = \begin{pmatrix} 0.75 & 0.125 & 0.125 \\ 0 & 0.25 & 0.75 \\ 0.5 & 0.5 & 0 \end{pmatrix}.$$

If we let $M = 18$, then, if $N \sim Poisson(\lambda \cdot t = 8)$, we have

$$\mathbb{P}(N > M) \approx 0.00065,$$

so our estimate is given by

$$\hat{P}(2) \doteq \hat{P}[18](2) \doteq \sum_{n=0}^{18} e^{-8} \frac{8^n}{n!} P^n = \begin{pmatrix} 0.50836069 & 0.24502253 & 0.24502253 \\ 0.48546768 & 0.2564721 & 0.25646596 \\ 0.49462243 & 0.25188859 & 0.25189473 \end{pmatrix},$$

which was calculated using Python as follows:

```
import numpy as np
import scipy.linalg
from scipy.stats import poisson
from numpy.linalg import matrix_power

def EstimatePHat(Q, epsilon, t):
    lam = max(abs(np.diag(Q)))
    d = len(Q)
    I = I = np.identity(d)
    P = I + np.multiply(Q,1/lam)
    M = int(poisson.ppf(1 - epsilon, lam*t))
    PHat = np.zeros([d,d])
    for n in range(0,M+1):
        PHat += np.multiply(matrix_power(P, n),poisson.pmf(n,lam*t))
    return(PHat)

Q = np.array([[-1,.5,.5],[0,-3,3],[2,2,-4]])
I = np.identity(3)
lam = max(abs(np.diag(Q)))
P = I + np.multiply(Q,1/lam)

EstimatePHat(Q,0.001,2)
```

3.7. **Long Term Behavior of CTMCs.**

Thus far we have spent most of our time with CTMCs discussing different methods to simulate them and estimate their transition functions. We now move towards studying their long-term behavior. Many of the results and definitions will be similar to their analogues for DTMCs, so it may be helpful to review Appendix E.

3.7.1. *Examples of Different Long-Term Behaviors.*

In this section we will see two examples of the different limiting behaviors that can arise for CTMCs.

Example 3.34. *Consider a CTMC with generator matrix*

$$Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix}.$$

In Example 3.27 *we saw that the transition function P can be written as*

$$P(t) = \begin{pmatrix} 1 & \frac{\mu}{\mu+\lambda} \\ 1 & -\frac{\lambda}{\mu+\lambda} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-(\mu+\lambda)t} \end{pmatrix} \begin{pmatrix} \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \\ 1 & -1 \end{pmatrix},$$

so if we note that $\lim_{t\to\infty} e^{-(\mu+\lambda)t} = 0$, *we see that*

$$\lim_{t\to\infty} P(t) = \begin{pmatrix} 1 & \frac{\mu}{\mu+\lambda} \\ 1 & -\frac{\lambda}{\mu+\lambda} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \\ \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \end{pmatrix}$$

In this case we see that as $t \to \infty$, $P(t)$ *converges to a stochastic matrix with identical rows.*

The following example illustrates different limiting behavior for the associated transition function.

Example 3.35. *Consider a CTMC* $\{X_t\}$ *on* $\mathscr{S} = \{1,2,3\}$ *with generator matrix*

$$Q = \begin{pmatrix} 0 & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda \\ 0 & 0 & 0 \end{pmatrix}.$$

We can diagonalize Q as

$$Q = UDU^{-1},$$

where

$$U = \begin{pmatrix} -\frac{\lambda}{\mu} & \frac{\lambda+\mu}{\mu} & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -(\lambda+\mu) \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ \frac{\mu}{\lambda+\mu} & 0 & \frac{\lambda}{\lambda+\mu} \\ -\frac{\lambda}{\lambda+\mu} & 1 & -\frac{\lambda}{\lambda+\mu} \end{pmatrix}$$

so we can apply Proposition 3.20 *and Proposition* 3.26 *to see that*

$$P(t) = U\exp(tD)U^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{\mu}{\lambda+\mu}(1-e^{-(\lambda+\mu)t}) & e^{-(\lambda+\mu)t} & \frac{\lambda}{\lambda+\mu}(1-e^{-(\lambda+\mu)t}) \\ 0 & 0 & 1 \end{pmatrix},$$

from which we can see that

$$\lim_{t\to\infty} P(t) = \begin{pmatrix} 1 & 0 & 0 \\ \frac{\mu}{\lambda+\mu} & 0 & \frac{\lambda}{\lambda+\mu} \\ 0 & 0 & 1 \end{pmatrix}.$$

In this case we see that as $t \to \infty$, $P(t)$ *converges to a stochastic matrix whose rows are not identical.*

It is also possible to construct examples where the limit of the transition function is not even a stochastic matrix. However, such an example is not important for our current purpose, so is omitted for now.

If we recall the notion of irreducibility from the theory of DTMCs, then our intuition suggests that the CTMC in Example 3.34, the chain can jump back and forth between the two states. However, in Example 3.35, once the chain reaches either state 1 or state 3, it cannot ever transition back to state 2. This suggests that the chain is not irreducible.

3.7.2. *Limiting Distribution of CTMCs.* The definition of the limiting distribution of a CTMC is very similar to that of the limiting distribution of a DTMC.

Definition 3.36. *Let $\{X_t\}$ be a CTMC on $\mathscr{S}$ with transition function P. A probability distribution μ on $\mathscr{S}$ is the **limiting distribution** of $\{X_n\}$ if, for all $x, y \in \mathscr{S}$,*

$$\lim_{t\to\infty} P_{x,y}(t) = \mu_y.$$

Observe that the CTMC in Example 3.34 has a limiting distribution, while the CTMC in Example 3.35 does not.

3.7.3. *Stationary Distributions of CTMCs.* As in the previous section, we can easily extend the notion of stationary distributions to CTMCs as well.

Definition 3.37. *ALet $\{X_t\}$ be a CTMC on $\mathscr{S}$ with transition function P. A probability distribution π on $\mathscr{S}$ is the **stationary distribution** of $\{X_n\}$ if*

$$\pi = \pi P(t), \quad t \geq 0.$$

Equivalently, π is a stationary distribution for $\{X_t\}$ if for all $x \in \mathscr{S}$ and $t \geq 0$,

$$\pi_x = \sum_{y\in\mathscr{S}} \pi_y P_{y,x}(t).$$

3.7.4. *Existence of a Unique Stationary Distribution for CTMCs.* We now introduce several definitions for CTMCs that are again analogous to the definitions for DTMCs.

Definition 3.38. *Let $\{X_t\}$ be a CTMC on $\mathscr{S}$ with transition function P. We say that a collection of states $\mathscr{C} \subseteq \mathscr{S}$ is a **communication class** if for all $x, y \in \mathscr{S}$, there are $s, t \geq 0$ such that*

$$P_{x,y}(s) > 0, \quad P_{y,x}(s) > 0.$$

*A communication class $\mathscr{C}$ is **closed** if for all $y \in \mathscr{S}$ such that there is some $x \in \mathscr{C}$ and $t \geq 0$ such that $P_{x,y}(t) > 0$, we have $y \in \mathscr{C}$. We say that $\{X_t\}$ is **irreducible** if for all $x, y \in \mathscr{S}$, there is some $t > 0$ such that $P_{x,y}(t) > 0$. Equivalently, $\{X_t\}$ is irreducible if and only if $\mathscr{S}$ consists of a single closed communication class.*

Note that we have not defined periodicity for CTMCs. Intuitively, all CTMCs of interest are aperiodic, since $\{X_t\}$ always remains in each state for some nonzero length of time. The proof of the following result is omitted, but it will still be useful to refer to. Intuitively, the proposition says that if there is some probability that a CTMC goes from x to y in t time units, then it would be possible for the chain to go from x to y in *any* amount of time.

Proposition 3.39. *Let P be a transition function on $\mathscr{S}$. For $x, y \in \mathscr{S}$, if $P_{x,y}(t) > 0$ for some $t > 0$, then $P_{x,y}(s) > 0$ for all $s > 0$.*

The following result provides conditions under which a CTMC has a unique stationary distribution (and under which this stationary distribution is the limiting distribution).

Theorem 3.40. *Let $\{X_t\}$ be a CTMC on a finite state space $\mathscr{S}$ with transition function P. If $\{X_t\}$ is irreducible, then there is a unique stationary distribution π, which is also a limiting distribution. In* *particular, for each $x, y \in \mathscr{S}$,*

$$\lim_{t\to\infty} P_{x,y}(t) = \pi_y,$$

or, equivalently,

$$\lim_{t\to\infty} P(t) = \Pi$$

where Π is an $|\mathscr{S}| \times |\mathscr{S}|$ matrix with entries given by

$$\Pi_{x,y} = \pi_y, \quad x, y \in \mathscr{S}.$$

Additionally, as with DTMCs, the unique stationary distribution describes, in the long term, the proportion of time that the chain spends in each state.

Returning to Example 3.34, we can see that the unique stationary distribution of a CTMC with generator matrix

$$Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix},$$

is given by

$$\pi = \begin{pmatrix} \frac{\lambda}{\lambda+\mu} & \frac{\mu}{\lambda+\mu} \end{pmatrix}$$

On the other hand, the CTMC from Example 3.35 with generator matrix

$$Q = \begin{pmatrix} 0 & 0 & 0 \\ \mu & -(\lambda+\mu) & \lambda \\ 0 & 0 & 0 \end{pmatrix}.$$

has no limiting distribution, and, for each $\alpha \in [0,1]$,

$$\pi(\alpha) \doteq \begin{pmatrix} \alpha & 0 & 1-\alpha \end{pmatrix},$$

is a stationary distribution. Recall from early on in our study of CTMCs that we said that the generator matrix plays a role similar to that of the transition matrix of a DTMC. Thus we might expect that stationary distributions of a CTMC can be characterized in terms of the generator matrix as well. As the next result shows, this is indeed the case.

Theorem 3.41. *Consider a CTMC on $\mathscr{S}$ with generator matrix Q. A probability distribution π on $\mathscr{S}$ is a stationary distribution for the chain if and only if*

$$\pi Q = \mathbf{0},$$

where $\mathbf{0} \in \mathbb{R}^{|\mathscr{S}|}$ is a vector of zeroes. Coordinate-wise this says that π is a stationary distribution if and only if for each $x \in \mathscr{S}$,

$$\sum_{y\in\mathscr{S}} \pi_y Q_{y,x} = 0.$$

Proof. Suppose that π is a stationary distribution for the CTMC. If we denote the chain's transition function by P, then

$$\pi = \pi P(t), \quad t \ge 0.$$

Using Theorem 3.17, we obtain, from differentiating the expression above with respect to t and using the fact that $P(0) = I$, that

$$\mathbf{0} = \pi Q P(0) = \pi Q I = \pi Q.$$

Now suppose that $\pi Q = \mathbf{0}$. Once more using Theorem 3.17, we see that

$$\mathbf{0} = \mathbf{0} P(t) = \pi Q P(t) = \pi P'(t),$$

which says that $\pi P(t)$ is constant in t. Additionally, $P(0) = I$, so

$$\pi P(t) = \pi P(0) = \pi I = \pi,$$

which says that π is a stationary distribution for the CTMC. □

The result above tells us that when a stationary distribution exists, it can be calculated by solving a system of linear equations. We demonstrate this with an example below.

Example 3.42. *Recall the setting from Example* 3.14*, where* $\{X_t\}$ *was a CTMC on* $\mathscr{S} = \{0,1,2\}$ *describing the number of machine components that were functioning at time* $t \geq 0$*. We saw that the generator matrix was given by*

$$Q = \begin{pmatrix} -\mu & \mu & 0 \\ \lambda & -(\mu+\lambda) & \mu \\ 0 & 2\lambda & -2\lambda \end{pmatrix}.$$

Let's determine, in the long run, for what proportion of time both components are functional.

From Theorem 3.40*, we know that since* $\{X_t\}$ *is irreducible, it has a unique stationary distribution (that is also the limiting distribution). Using Theorem* 3.41*, we see that the stationary distribution satisfies*

$$\pi Q = \mathbf{0}$$
$$\pi_0 + \pi_1 + \pi_2 = 1.$$

We have that

$$\begin{aligned} -\mu\pi_0 + \lambda\pi_1 &= 0 \\ \mu\pi_0 - (\mu+\lambda)\pi_1 + 2\lambda\pi_2 &= 0 \\ \mu\pi_1 - 2\lambda\pi_2 &= 0. \end{aligned}$$

This says that $\pi_1 = \frac{\mu}{\lambda}\pi_0$*, and that*

$$\pi_2 = \frac{\mu}{2\lambda}\pi_1 = \frac{\mu}{2\lambda}\frac{\mu}{\lambda}\pi_0 = \frac{\mu^2}{2\lambda^2}\pi_0.$$

And

$$\pi_0 + \frac{\mu}{\lambda}\pi_0 + \frac{\mu^2}{2\lambda^2}\pi_0 = 1,$$

so

$$\pi_0 = \frac{1}{1 + \frac{\mu}{\lambda} + \frac{\mu^2}{2\lambda^2}} = \frac{2\lambda^2}{2\lambda^2 + 2\mu\lambda + \mu^2},$$

which we can use to calculate

$$\pi_1 = \frac{2\lambda\mu}{2\lambda^2 + 2\mu\lambda + \mu^2}, \quad \pi_2 = \frac{\mu^2}{2\lambda^2 + 2\mu\lambda + \mu^2}.$$

This tells us that, in the long term, the proportion of time in which both components are functional is

$$\frac{\mu^2}{2\lambda^2 + 2\mu\lambda + \mu^2}.$$

Since π *is the limiting distribution, we can also compute, using the fact that if* $X_0 \sim \pi$*, then* $X_t \sim \pi$ *for all* $t \geq 0$*,*

$$\mathbb{E}[X_t | X_0 \sim \pi] = \sum_{x \in \mathscr{S}} x\pi_x = 0 \cdot \pi_0 + 1 \cdot \pi_1 + 2 \cdot \pi_2.$$

Additionally, since π *is a limiting distribution, we know that, for all* $x, y \in \mathscr{S}$*,*

$$\lim_{t\to\infty} \mathbb{P}[X_t = x | X_0 = y] = \pi_x,$$

so it follows that, for all $y \in \mathscr{S}$,

$$\lim_{t\to\infty} \mathbb{E}[X_t | X_0 = y] = \sum_{x\in\mathscr{S}} x\pi_x = 0\cdot\pi_0 + 1\cdot\pi_1 + 2\cdot\pi_2.$$

We consider another example below.

Example 3.43. *Recall the machine with two components from Example* 3.14*; here X_t denotes the number of functioning components at time t, and has generator matrix (here we are setting $\lambda = 2, \mu = 1$)*

$$Q = \begin{pmatrix} -1 & 1 & 0 \\ 2 & -3 & 1 \\ 0 & 4 & -4 \end{pmatrix},$$

where the rates are in hours. Suppose that the cost to fix a machine is 40 dollars per hour. In the long term, over each 24 hour period, how much should the company expect to pay the repairperson?

We can solve the equation $\pi Q = \mathbf{0}$ to see that the stationary distribution is given by

$$\pi = \begin{pmatrix} 8/13 & 1/13 & 4/13 \end{pmatrix},$$

and, since the chain is irreducible, π is a limiting distribution. In the long term, machines are being fixed whenever $X_t = 0$ or $X_t = 1$, so the proportion of time in which machines are being fixed is $5/13$*. Thus, in the long term, the company should expect to pay* $(5/13)\cdot 24\cdot 40 = 369.23$ *dollars per day to the repairperson.*

3.7.5. *Balance Equations.* As a consequence of Theorem, 3.41 we can derive the balance equations, which provide another characterization of the stationary distribution(s) of a CTMC. Observe that for a generator matrix Q on a state space $\mathscr{S}$, since, for each $x \in \mathscr{S}$, $Q_{x,x} = -\sum_{y\neq x} Q_{x,y}$, so we can write

$$0 = \sum_{y\in\mathscr{S}} \pi_y Q_{y,x} = \sum_{y\neq x} \pi_y Q_{y,x} + \pi_x Q_{x,x} = \sum_{y\neq x} \pi_y Q_{y,x} - \pi_x \sum_{y\neq x} Q_{x,y},$$

which says that

$$\pi_x \sum_{y\neq x} Q_{x,y} = \sum_{y\neq x} \pi_y Q_{y,x}. \tag{28}$$

Note that $\sum_{y\neq x} Q_{x,y}$ describes the rate at which the chain moves out of state x given that it starts in state x, so

$$\pi_x \sum_{y\neq x} Q_{x,y}$$

describes, in the long-term, the rate at which the chain leaves state x. Similarly, $Q_{y,x}$ describes the rate at which the chain moves to state x from state y, so

$$\sum_{y\neq x} \pi_y Q_{y,x},$$

describes, in the long-term, the rate at which the chain enters state x. Therefore, the global balance equations in (28) say that, in the long term, the rate at which the chain enters each state is equal to the rate at which it leaves the state.

The following example illustrates this idea.

Example 3.44. *Consider a CTMC $\{X_t\}$ on $\mathscr{S} = \{1,2,3\}$ with generator matrix*

$$Q = \begin{pmatrix} -5 & 3 & 2 \\ 2 & -3 & 1 \\ 0 & 2 & -2 \end{pmatrix}.$$

Since $\{X_t\}$ is irreducible, its unique stationary distribution π satisfies

$$\pi Q = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 \end{pmatrix} \begin{pmatrix} -5 & 3 & 2 \\ 2 & -3 & 1 \\ 0 & 2 & -2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} = \mathbf{0},$$

which yields

$$\pi = \begin{pmatrix} \frac{4}{23} & \frac{10}{23} & \frac{9}{23} \end{pmatrix}.$$

The global balance equations in (28) *say that, in the long term, every 23 time units, the chain will enter and then exit state 1 four times, will enter and exit state 2 ten times, and will enter and then exit state 3 nine times.*

We consider another example below.

Example 3.45. *When someone holds the one ring for too long a time, they become overly attached to it. While on their journey to Mordor, Frodo, Sam, Merry, and Pippin develop a strategy to ensure that no one holds the ring too long continuously. Their strategy works as follows; when Frodo feels himself becoming overly attached to the ring, he flips a fair coin. If the coin lands on heads, Frodo passes the ring on to Merry. If it lands on tails, Frodo passes it on to Pippin. If the ring is passed on to Pippin, he holds onto it as long as he can bear before passing it on to Sam. Similarly, if Frodo passes it to Merry, he holds onto then passes it to Sam. Sam holds onto the ring until he becomes overly attached, at which point he passes it back to Frodo, and the cycle repeats.*

The time that it takes for Frodo to become overly attached to the ring can be modeled an Exp(0.2) *random variable, so that, on average, Frodo can hold the ring for 5 days before passing it on. The time that it takes for Merry and Pippin to pass the ring on can be modeled by Exp*(0.5) *and Exp*(0.7) *random variables, respectively. Sam, who is strong of heart, can hold the ring for the longest; the time unitl he passes it on can be modeled as an Exp*(0.05) *random variable.*

Their journey to Mordor is very long. During a typical 30 day stretch of their journey, how many times do we expect the ring to be passed to Sam?

We can model the current bearer of the ring as a CTMC $\{X_t\}$ on $\mathscr{S} = \{F, M, P, S\}$ with generator matrix

$$Q = \begin{pmatrix} -0.2 & 0.1 & 0.1 & 0 \\ 0 & -0.5 & 0 & 0.5 \\ 0 & 0 & -0.7 & 0.7 \\ 0.05 & 0 & 0 & -0.05 \end{pmatrix}.$$

Since the chain is irreducible, we can calculate its stationary distribution as the solution π to

$$\pi Q = \mathbf{0},$$

which is given by

$$\pi = \begin{pmatrix} \frac{35}{187} & \frac{7}{187} & \frac{5}{187} & \frac{140}{187} \end{pmatrix}$$

From the balance equations, we see that, in the long term, over a typical 30 day stretch of their journey, the ring will be passed to (and away from) Sam about $30 \cdot \frac{140}{187} = 22.46$ *times.*

3.8. **Stationary Distribution of the Embedded Chain.** Consider an irreducible CTMC $\{X_t\}$ on a finite state space $\mathscr{S}$ and recall the associated embedded chain $\{Y_n\}$ from Definition 3.10). If we denote the generator matrix of $\{X_t\}$ by Q and the transition matrix of $\{Y_n\}$ by P, then we know that $\{X_t\}$ has a stationary distribution π satisfying

$$\pi Q = \mathbf{0}.$$

Recalling Theorem 3.40, we know that π is also the limiting distribution of $\{X_t\}$, so it describes, in the long term, the probability that $\{X_t\}$ is in each state.

Now, let us considerAdditionally, $\{Y_n\}$ is irreducible as well (though it may not be aperiodic; see Example 3.46 below). Recall from Corollary E.57 that since $\{Y_n\}$ is an irreducible DTMC on a finite state space, it also has a unique (positive) stationary distribution, which we will denote by $\tilde{\pi}$; recall that this stationary distribution satisfies

$$\tilde{\pi} P = \tilde{\pi}.$$

However, $\{Y_n\}$ is not necessarily aperiodic, so it might be the case that $\tilde{\pi}$ is **not** a limiting distribution of $\{Y_n\}$, so we would like to understand how we should interpret $\tilde{\pi}$.

We begin by considering a simple example below.

Example 3.46. *Consider a CTMC* $\{X_t\}$ *on* $\mathscr{S} = \{1,2\}$ *with generator matrix*

$$Q = \begin{pmatrix} -\mu & \mu \\ \lambda & -\lambda \end{pmatrix},$$

and recall, from Example 3.34*, that the limiting distribution (which is the unique stationary distribution) of* $\{X_t\}$ *is given by*

$$\pi = \begin{pmatrix} \frac{\lambda}{\mu+\lambda} & \frac{\mu}{\mu+\lambda} \end{pmatrix}$$

The associated embedded chain $\{Y_n\}$ *has transition matrix*

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

so its unique stationary distribution is given by

$$\tilde{\pi} = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}.$$

Here we should interpret $\tilde{\pi}_x$ *as describing, in the long term, the proportion of transitions that the chain* $\{X_t\}$ *makes into state* $x \in \mathscr{S}$*. Every time the chain is in state* 1*, it eventual,ly jumps into state* 2*. Similarly, every time the chain is in state* 2*, it eventually jumps into state* 1*. Together, this means that, in the long term, half of the chains jumps will be into state* 1*, and the other half will be into state* 2*.*

On the other hand, π *describes the proportion of time that* $\{X_t\}$ *spends in each state; the information about how long* $\{X_t\}$ *spends in each state is lost when we consider the embedded chain, since it only keeps track of the jumps, and not how long the CTMC remains in each state before jumping.*

The following result describes the relationship between the stationary distribution of a CTMC and the stationary distribution of the associated embedded chain.

Proposition 3.47. *Let $\{X_t\}$ be an irreducible CTMC on a finite state space $\mathscr{S}$ with generator matrix Q. Let $\{Y_n\}$ be the associated embedded DTMC with transition matrix P. Denote the unique stationary distributions of $\{X_t\}$ and $\{Y_n\}$ by π and $\tilde{\pi}$, respectively. Then, for each $x \in \mathscr{S}$,*

$$\tilde{\pi}_x = \frac{\pi_x q_x}{\sum\limits_{y \in \mathscr{S}} \pi_y q_y},$$

where, as before,

$$q_y \doteq \sum_{z \neq y} q_{y,z}.$$

Similarly, for each $x \in \mathscr{S}$,

$$\pi_x = \frac{\frac{\tilde{\pi}_x}{q_x}}{\sum\limits_{y \in \mathscr{S}} \left(\frac{\tilde{\pi}_y}{q_y} \right)}$$

Proof. For the first part of the result, it suffices to show that if π is a stationary distribution for $\{X_t\}$ then, with

$$\tilde{\pi}_x = \frac{\pi_x q_x}{\sum\limits_{y \in \mathscr{S}} \pi_y q_y}, \quad x \in \mathscr{S},$$

we have that $\tilde{\pi} P = \tilde{\pi}$. Recall that

$$-\pi_x q_x + \sum_{y \neq x} \pi_y q_{y,x} = \pi_x q_{x,x} + \sum_{y \neq x} \pi_y q_{y,x} = \sum_{y \in \mathscr{S}} \pi_y q_{y,x} = (\pi Q)_x = 0,$$

so

$$\pi_x q_x = \sum_{y \neq x} \pi_y q_{y,x}.$$

Let $\tilde{\psi}_x \doteq \pi_x q_x$, and recall that $P_{y,x} \doteq \frac{q_{y,x}}{q_y}$, so the above calculation can be rewritten as saying

$$\tilde{\psi}_x = \sum_{y \neq x} \tilde{\pi}_y q_{y,x} = \sum_{y \neq x} \frac{\tilde{\psi}_y}{q_y} q_{y,x} = \sum_{y \neq x} \tilde{\psi}_y P_{y,x}.$$

Then, since $P_{x,x} = 0$,

$$(\tilde{\psi} P)_x = \sum_{y \in \mathscr{S}} \tilde{\psi}_y P_{y,x} = \sum_{y \neq x} \tilde{\psi}_y P_{y,x} + \psi_x P_{x,x} = \tilde{\psi}_x + 0 = \tilde{\psi}_x,$$

meaning that $\tilde{\psi} P = \tilde{\psi}$. The result follows on noting that

$$\tilde{\pi} = \frac{\tilde{\psi}}{\sum\limits_{y \in \mathscr{S}} \tilde{\psi}_y}.$$

The other part of the proof is similar and is omitted.

□

3.9. **Time Reversible Chains.** In this section we discuss the notion of time reversibility. Intuitively, a chain is time reversible if it behaves the same as time progresses normally as it does when time is reversed. It is helpful to begin with an example of a Markov chain that is not time reversible.

Example 3.48. *Let $\{X_t\}$ be a CTMC on $\mathscr{S} = \{0,1,2\}$ with generator matrix*

$$Q = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

If $X_0 = 0$, then the embedded chain $\tilde{X}_n$ will make the transitions

$$0 \mapsto 1 \mapsto 2 \mapsto 0 \mapsto 1 \mapsto 2 \mapsto \cdots.$$

On the other hand, if we fix some $N \in \mathbb{N}$ such that $\tilde{X}_N = 0$, then the reversed chain $\{\tilde{X}_n^{r,N}\}_{n=0}^N$ defined by

$$\tilde{X}_n^{r,N} = \tilde{X}_{N-n}, \quad n = 0, \ldots, N,$$

makes the transitions

$$0 \mapsto 2 \mapsto 1 \mapsto 0 \mapsto 2 \mapsto \cdots.$$

This means that if we observed $\{\tilde{X}_n\}$ and $\{\tilde{X}_n^{r,N}\}$ from time $n = 0$ to time $n = N$, we could differentiate between the two. Consequently, $\{\tilde{X}_n\}$ is not reversible, and neither is $\{X_t\}$.

We now formally introduce the notion of time reversibility.

Definition 3.49. *Consider a CTMC on $\mathscr{S}$ with generator matrix Q and unique stationary distribution π. The chain is said to be* ***time reversible*** *if*

$$\pi_x q_{x,y} = \pi_y q_{y,x}, \quad x, y \in \mathscr{S}.$$

These equations are known as the ***local (or detailed) balance equations****.*

We can use the local balance equations to calculate stationary distributions.

Proposition 3.50. *Consider an irreducible CTMC $\{X_t\}$ on $\mathscr{S}$ with generator matrix Q. If λ is a probability distribution on $\mathscr{S}$ such that*

$$\lambda_x q_{x,y} = \lambda_y q_{y,x}, \quad x, y \in \mathscr{S},$$

then λ is the unique stationary distribution of the chain.

Proof. According to Theorem 3.41, it suffices to show that

$$\lambda Q = \mathbf{0}.$$

Note that, for each $x \in \mathscr{S}$,

$$(\lambda Q)_x = \sum_{y \in \mathscr{S}} \lambda_y q_{y,x} = \sum_{y \in \mathscr{S}} \lambda_x q_{x,y} = \lambda_x \sum_{y \in \mathscr{S}} q_{x,y} = 0,$$

since each row of the the generator matrix sums to 0. Therefore, λ is the unique stationary distribution. □

Now, we consider a less trivial example of a chain that is not reversible.

Example 3.51. *Consider a CTMC on* $\mathscr{S} = \{1,2,3\}$ *with generator matrix*

$$Q = \begin{pmatrix} -1 & \lambda & 1-\lambda \\ 1-\lambda & -1 & \lambda \\ \lambda & 1-\lambda & -1 \end{pmatrix},$$

where $\lambda \in [0,1]$*. Since the chain is irreducible and* $\mathscr{S}$ *is finite, there is a unique stationary distribution satisfying* $\pi Q = \mathbf{0}$*, meaning that*

$$\begin{cases} -\pi_1 + (1-\lambda)\pi_2 + \lambda\pi_3 & = 0 \\ \lambda\pi_1 - \pi_2 + (1-\lambda)\pi_3 & = 0 \\ (1-\lambda)\pi_1 + \lambda\pi_2 - \pi_3 & = 0. \end{cases}$$

We can see that the solution is given by $\pi = \begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}$*, that is* π *is the uniform distribution on* $\mathscr{S}$*. But, is the chain reversible?*

The detailed balance equations are then given by

$$\begin{cases} (1/3)\lambda = (1/3)(1-\lambda) \\ (1/3)(1-\lambda) = (1/3)\lambda \\ (1/3)\lambda = (1/3)(1-\lambda) \end{cases},$$

which is satisfied if and only if $\lambda = 1/2$*. Thus, the CTMC is reversible if and only if* $\lambda = 1/2$*. This makes sense; for example, if* $\lambda = 0.99$*, then the chain will tend, on average, to move from 0 to 1 to 2 to 0, and so on. Thus, if you were to observe the chain backwards in time, you would see it tending to follow a trajectory of 0 to 2 to 1 to 2, and so on. This means that we could distinguish between its behavior forwards and backwards in time.*

Below we apply Proposition 3.50 to derive the stationary distribution of the birth and death process.

Theorem 3.52. *Let* $\{X_t\}$ *be a CTMC on* $\mathscr{S} = \{0,1,2,\dots\}$ *with generator matrix* $Q = [q_{i,j}]_{i,j\in\mathscr{S}}$*, where*

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & & \\ & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & \\ & & & & \ddots \end{pmatrix}.$$

Then $\{X_t\}$ *has a stationary distribution if and only if*

$$\sum_{k=0}^{\infty} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} < \infty,$$

in which case the statinary distribution is given by

$$\pi_0 = \left(\sum_{k=0}^{\infty} \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i} \right)^{-1},$$

and

$$\pi_k = \pi_0 \prod_{i=1}^{k} \frac{\lambda_{i-1}}{\mu_i}, \quad k \geq 1.$$

Proof. It suffices to show that the detailed balance equations hold with this choice of π. Here the detailed balance equations are given by

$$\pi_k\lambda_k = \pi_{k+1}\mu_{k+1}, \quad k \in \mathscr{S},$$

and we have that

$$\pi_k\lambda_k = \left(\pi_0\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i}\right)\lambda_k = \left(\pi_0\prod_{i=1}^{k+1}\frac{\lambda_{i-1}}{\mu_i}\right)\mu_{k+1} = \pi_{k+1}\mu_{k+1}.$$

Additionally, we can see that π is a probability distribution, as $\pi_k \geq 0$ for each $k \in \mathscr{S}$, and

$$\sum_{k\in\mathscr{S}}\pi_k = \sum_{k=0}^{\infty}\left(\pi_0\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i}\right) = \pi_0\sum_{k=0}^{\infty}\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i} = 1.$$

□

Recall that a discrete-time random walk on $\mathscr{S} = \{0, 1, \dots\}$ with transition matrix

$$P = \begin{pmatrix} 1-p & p & & & \\ 1-p & & p & & \\ & 1-p & & p & \\ & & & & \ddots \end{pmatrix}$$

has a stationary distribution if and only if $p < 0.5$. In this case, the stationary distribution is given by

$$\pi_0 = \sum_{i=0}^{\infty}\left(\frac{p}{1-p}\right)^i$$
$$\pi_k = \left(\frac{p}{1-p}\right)^k\pi_0, \quad k \geq 1.$$

A birth and death process whose birth and death rates are state-independent is the continuous-time analogue of such a discrete-time random walk, so we should expect that a similar condition will need to hold for there to be a stationary distribution. We verify this in the next example.

Example 3.53. *Let $\{X_t\}$ be a CTMC on $\mathscr{S} = \{0, 1, 2, \dots\}$ with generator matrix $Q = [q_{i,j}]_{i,j\in\mathscr{S}}$, where*

$$Q = \begin{pmatrix} -\lambda & \lambda & & & \\ \mu & -(\mu+\lambda) & \lambda & & \\ & \mu & -(\mu+\lambda) & \lambda & \\ & & & & \ddots \end{pmatrix}$$

Using Theorem 3.52, we know that $\{X_t\}$ has a stationary distribution if and only if

$$\sum_{k=0}^{\infty}\prod_{i=1}^{k}\frac{\lambda}{\mu} = \sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k < \infty.$$

The series above converges if and only if $\mu > \lambda$, as expected. Once more applying Theorem 3.52, we see that the stationary distribution π is given by

$$\pi_0 = \left(\sum_{k=0}^{\infty}\prod_{i=1}^{k}\frac{\lambda}{\mu}\right)^{-1} = \left(\sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\right)^{-1} = \left(\frac{\mu}{\mu-\lambda}\right)^{-1} = \frac{\mu-\lambda}{\mu},$$

and

$$\pi_k = \pi_0\prod_{i=1}^{k}\frac{\lambda}{\mu} = \frac{\mu-\lambda}{\mu}\left(\frac{\lambda}{\mu}\right)^k, \quad k \geq 1.$$

Additionally, if $\mu > \lambda$, then π is also the limiting distribution of the chain, and therefore describes the long-term probability of the system being in each state.

The previous examples show that, in some settings, when a CTMC is reversible, the local balance equations provide a straightforward way to calculate the stationary distribution. The next example illustrates how we can use time reversibility to study the long-term behavior of a simple queue.

Example 3.54. *Suppose that customers arrive at a bank according to a Poisson process with a rate of 10 customers per hour. There is a single teller at the bank whose service times are independent and distributed according to an Exponential distribution with a mean of* 4 *minutes. On average, in the long term, how many people are in the queue?*

Let X_t *denote the number of customers in the queue at time* t*. Then* $\{X_t\}$ *is a birth and death process with state-independent birth and death rates of* $\lambda = 1/6$ *and* $\mu = 1/4$*, respectively. From Example* 3.53*, it follows that the stationary distribution of the system is*

$$\pi_k = \frac{\mu-\lambda}{\mu}\left(\frac{\lambda}{\mu}\right)^k = \frac{\mu-\lambda}{\mu}\left(1-\frac{\mu-\lambda}{\mu}\right)^k, \quad k \geq 0,$$

and that π *is also the limiting distribution. Thus if there are* k_0 *customers at time* 0*, then, in the long term, the average number of customers is given by*

$$\lim_{t\to\infty} \mathbb{E}[X_t|X_0 = k_0] = \sum_{k=0}^{\infty} k\pi_k = \frac{\mu}{\mu-\lambda} - 1 = \frac{1/4}{1/4-1/6} - 1 = 3-1 = 2.$$

In the calculation above, we have used the fact that π *is a geometric distribution on* $\{0,1,2,\ldots\}$ *with parameter* $p \doteq \frac{\mu-\lambda}{\mu}$*. Recall, that if* X *follows a geometric distribution on* $\{0,1,2\ldots\}$ *with parameter* p*, then* $\mathbb{E}(X) = p^{-1} - 1$.

4. Brownian Motion

In this section we introduce our first example of a continuous-time stochastic process with continuous state space. This process, known as a Brownian motion or Weiner process is a fundamental object in probability theory, and is both deeply mathematically interesting and incredibly useful; some of the fields in which it is applied include finance, biology, economics, operations research, physics, and statistics.

Just as the normal distribution is "universal" in that it arises in very general settings as the limiting distribution of the (properly rescaled) fluctuations of sums around the mean of a distribution, Brownian motion arises as the limiting distribution of (properly rescaled) fluctuations of random walks around their "expected," or average, trajectories.

We begin by introducing the definition of a standard Brownian motion. It may be helpful to recall the definition of stationary and independent increments from Definition 2.21. We also briefly review some of the important properties of normally-distributed random variables.

4.1. **Normal Random Variables.** Recall that we write $X \sim \mathcal{N}(\mu, \sigma^2)$ when X is a random variable with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

As in Example C.14 in Appendix C, one can check that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the mgf of X is given by

$$m_X(t) \doteq \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right),$$

from which we can see that

$$\frac{d}{dt} m_X(t) = (\mu + \sigma^2 t) \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right),$$

which yields

$$\mathbb{E}(X) = \frac{d}{dt} m_X(t)\Big|_{t=0} = \mu.$$

Furthermore,

$$\begin{aligned} \frac{d^2}{dt^2} m_X(t) &= \frac{d}{dt}\left((\mu + \sigma^2 t) \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)\right) \\ &= \sigma^2 \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) + (\mu + \sigma^2 t)^2 \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \end{aligned}$$

so

$$\mathbb{E}(X^2) = \frac{d^2}{dt^2} m_X(t)\Big|_{t=0} = \sigma^2 + \mu^2.$$

It follows that

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2.$$

Additionally, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $Z \doteq \frac{X-\mu}{\sigma}$ follows a $\mathcal{N}(0,1)$ distribution; this can also be checked using the basic properties of moment generating functions. As we mentioned in the introduction to this section, Brownian motion can be thought of as a stochastic process satisfying a generalization of the central limit theorem. We recall the central limit theorem from PSTAT120A below.

Theorem 4.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be iid random variables with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Consider the sequence $\{S_n\}_{n\in\mathbb{N}}$ defined as*

$$S_n \doteq \sum_{i=1}^{n} X_i.$$

Then, as $n \to \infty$,

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0,1)$ and $\xrightarrow{d}$ denotes convergence in distribution (see Definition C.17*). We can informally interpret this as saying that for large values of n,*

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} \approx \mathcal{N}(0,1).$$

We now briefly discuss multivariate normal random variables.

4.2. **Multivariate Normal Random Variables.** In this section we briefly introduce the notion of multivariate random variables and note several important properties of multivariate normal random variables. Just as (univariate) normal random variables are described through their mean and variance, so too are multivariate normal random variables; they only difference is that here the mean is a vector, and the variance is a matrix.

Definition 4.2. *For a multivariate random variable $\boldsymbol{X} = \begin{pmatrix} X_1 & \dots & X_d \end{pmatrix} \in \mathbb{R}^d$, we define the d-dimensional **mean vector** $\mathbb{E}(\boldsymbol{X})$ by*

$$\mathbb{E}(\boldsymbol{X}) = \begin{pmatrix} \mathbb{E}(X_1) & \dots & \mathbb{E}(X_d) \end{pmatrix}^T,$$

*provided that each of the above expected values is well-defined. Similarly, we define the $d \times d$ **covariance matrix**, $Cov(\boldsymbol{X})$, of $\boldsymbol{X}$ by*

$$(Cov(\boldsymbol{X}))_{i,j} = Cov(X_i, X_j), \quad 1 \le i, j \le d.$$

We begin by introducing the definition of a multivariate normal random variable.

Definition 4.3. *A vector $\boldsymbol{X} = (X_1, \dots, X_d)^T \in \mathbb{R}^d$ is said follow a **multivariate normal distribution** if for every vector $\boldsymbol{a} = (a_1, \dots, a_d)^T \in \mathbb{R}^d$, there are some $\mu \in \mathbb{R}$ and $\sigma \ge 0$ (both, possibly, depending on a) such that*

$$\boldsymbol{a}^T \boldsymbol{X} = \begin{pmatrix} a_1 & \dots & a_d \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} = \sum_{i=1}^{d} a_i X_i \sim \mathcal{N}(\mu, \sigma^2).$$

*We say that random variables $X_1, \dots, X_n$ are **jointly normal** if the vector $\boldsymbol{X} \doteq (X_1, \dots, X_n)$ follows a multivariate normal distribution.*

Below we describe multivariate normal distributions through their mean and covariance structure.

Definition 4.4. *Let $\boldsymbol{X} = (X_1,\dots,X_d)^T$ be a multivariate normal random variable and define the vector $\boldsymbol{\mu} \doteq \mathbb{E}(\boldsymbol{X}) \in \mathbb{R}^d$, and the matrix $\Sigma \doteq Cov(\boldsymbol{X}) \in \mathbb{R}^{d\times d}$, so that*

$$\boldsymbol{\mu} = (\mathbb{E}(X_1),\dots,\mathbb{E}(X_d))^T,$$

and

$$\Sigma_{i,j} \doteq Cov(X_i, X_j).$$

We refer to $\boldsymbol{\mu}$ as the ***mean vector*** *of $\boldsymbol{X}$ and Σ as the* ***covariance matrix*** *of $\boldsymbol{X}$, and we write $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$.*

The definition of the multivariate normal distribution is somewhat unintuitive. Below we see why it is necessary, and why it is not sufficient for X_i to individually follow a normal distribution. We also see that uncorrelated normal random variables are not necessarily independent of one another (however, as we see in Theorem 4.7, if they are *jointly normal*, then they are independent if and only if they are uncorrelated).

Example 4.5. *Let $X \sim \mathcal{N}(0,1)$ and let W be a random variable that is independent of X with distribution*

$$\mathbb{P}(W = 1) = \mathbb{P}(W = -1) = \frac{1}{2}.$$

Define $Z = WX$. Then (X, Z) are not jointly normal. To see this, observe that

$$X + Z = X + WX = (1+W)X,$$

and recall that the mgf of X is

$$m(t) = \mathbb{E}[e^{tX}] = e^{\frac{t^2}{2}},$$

so we can use the law of total expectation to see that the mgf of $X + Z$ is given by

$$\mathbb{E}[e^{t(X+Z)}] = \mathbb{E}\left[\mathbb{E}[e^{t(1+W)X}]|W]\right] = \mathbb{E}\left[e^{\frac{(t(1+W))^2}{2}}\right] = e^{\frac{(t(1+1))^2}{2}} \cdot \frac{1}{2} + e^{\frac{(t(1+(-1)))^2}{2}} \cdot \frac{1}{2} = \frac{1}{2}\left(1 + e^{\frac{4t^2}{2}}\right).$$

Since the mgf of a $\mathcal{N}(\mu, \sigma^2)$ random variable is given by

$$m_{\mu,\sigma^2}(t) \doteq e^{\mu t + \frac{(t\sigma)^2}{2}},$$

it is clear that $X + Z$ does not follow a normal distribution (i.e., there is no choice of μ and σ^2 such that the mgf of $X + Z$ is the same as the mgf of the $\mathcal{N}(\mu,\sigma^2)$ distribution). Thus, X and Z are ***not*** *jointly normal, even though they are both (marginally) normal.*

Note also that, since X and W are independent,

$$\mathbb{E}(XZ) = \mathbb{E}(X^2W) = \mathbb{E}(X^2)\mathbb{E}(W) = 1 \cdot 0 = 0,$$

which means that X and Z are uncorrelated, as

$$Cov(X, Z) = \mathbb{E}(XZ) - \mathbb{E}(X)\mathbb{E}(Z) = 0.$$

However, X and Z are not independent, as, for each $z > 0$,

$$\mathbb{P}(Z \geq z|X \geq z) = \mathbb{P}(WX \geq z|X \geq z) = \mathbb{P}(W = 1) = \frac{1}{2} \neq \mathbb{P}(Z \geq z).$$

Above we saw that (X, Z) are not jointly normal, since $X + Z$ does not follow a normal distribution. Then, since $X + Z$ does not follow a normal distribution, what distribution does it follow?

If we denote the conditional distribution of X given Y by $\mathscr{L}(X|Y)$, thenthe conditional distribution of $X+Z$ given that $W=1$ is

$$\mathscr{L}(X+Z|W=1)=\mathscr{L}(X+X|W=1)=\mathscr{L}(2X|W=1)=\mathscr{L}(2X)=\mathscr{N}(0,4).$$

So, if $W=1$, which happens with probability $1/2$, $X+Z$ follows a $\mathscr{N}(0,4)$ distribution. On the other hand, if $W=-1$, then the conditional distribution of $X+Z$ given that $W=-1$ is

$$\mathscr{L}(X+Z|W=1)=\mathscr{L}(X-X|W=1)=\mathscr{L}(0|W=1)=\mathscr{L}(0).$$

This says that if $W=-1$, then $X+Z=0$, which means that with probability $1/2$, $X+Z$ behaves like a $\mathscr{N}(0,4)$ random variable, and with probability $1/2$, it equals 0. We can interpret $X+Z$ as a mixture *distribution of the form*

$$\mathscr{L}(X+Z)=\frac{1}{2}\delta_0+\frac{1}{2}\mathscr{N}(0,4),$$

where $Y\sim\delta_0$ means that $\mathbb{P}(Y=0)=1$.

The following Python code illustrates Example 4.5.

```
import numpy as np
import matplotlib.pyplot as plt

n = 100
x = np.random.normal(size = n)
w = np.random.choice([-1,1], p = [1/2,1/2], size = n)
z = np.multiply(x,w)
xz = np.stack((x,z), axis = 1)

plt.scatter(xz[:,0], xz[:,1])
```

The results are plotted in Figure 4.1 below.

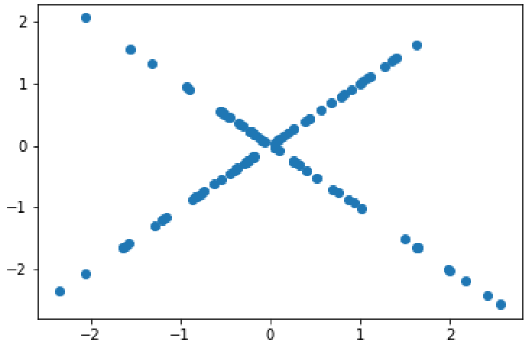


FIGURE 4.1. The random variable from Example 4.5 are not jointly normally distributed.

The next result shows how to construct bivariate normal random variables using iid standard normal random variables. The proof is omitted, but easily follows from the fact that moment generating functions uniquely characterize the distributions of random variables.

Theorem 4.6. *Let $Z_1, Z_2 \overset{iid}{\sim} \mathcal{N}(0,1)$. Then, the bivariate random variable (X,Y) defined by*

$$X \doteq \sigma_X Z_1 + \mu_X, \quad Y \doteq \sigma_Y(\rho Z_1 + \sqrt{1-\rho^2} Z_2) + \mu_Y,$$

follows a $\mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$ distribution, where

$$\boldsymbol{\mu} = (\mu_X, \mu_Y), \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

The following result can be useful when studying stochastic processes whose increments follow normal distributions.

Theorem 4.7. *Suppose that (X,Y) are jointly normal. Then X and Y are independent if and only if Cov$(X,Y) = 0$.*

Proof. If X and Y are independent, then, since $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) = 0$, we have $\text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$.

Denote the mean vector of (X,Y) by $\boldsymbol{\mu} = (\mu_X, \mu_Y)$ and the covariance matrix of (X,Y) by

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho \\ \rho & \sigma_Y^2. \end{pmatrix}$$

and suppose that $\rho = \text{Cov}(X,Y) = 0$. From 4.6, we have that

$$(X,Y) \overset{d}{=} (\tilde{X}, \tilde{Y}),$$

where

$$\tilde{X} \doteq \sigma_X Z_1 + \mu_X, \quad \tilde{Y} \doteq \sigma_Y Z_2 + \mu_Y,$$

and $Z_1, Z_2 \overset{iid}{\sim} \mathcal{N}(0,1)$. The result follows.

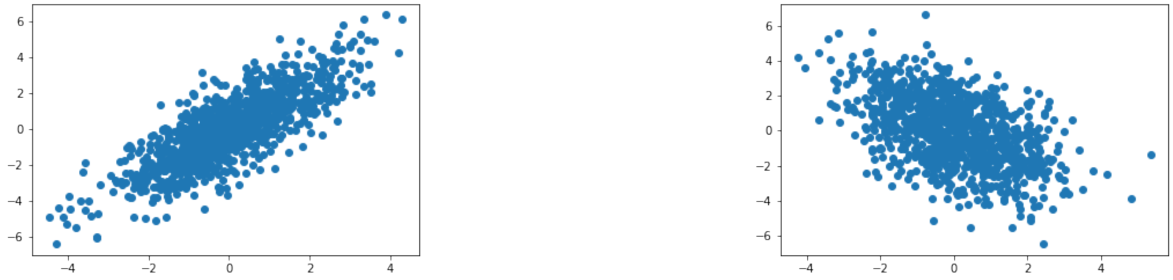
□

The Python code below illustrates how we can use Theorem 4.7 to simulate bivariate normal random variables.

```
import numpy as np
import matplotlib.pyplot as plt

def simBVN(mux, muy, rho, varx, vary, n):
    z1 = np.random.normal(size = n)
    z2 = np.random.normal(size = n)
    x = (varx**(1/2))*z1 + np.full(n,mux)
    y = (vary**(1/2))*(rho*z1 + ((1 - rho**2)**(1/2))*z2) + np.full(n,muy)
    return x, y

rho = 0.8
mux = 0
muy = 0
varx = 2
vary = 4
n = 1000
xy = simBVN(mux, muy, rho, varx, vary, n)
```

```
plt.scatter(xy[0], xy[1])

rho = -0.5
mux = 0
muy = 0
varx = 2
vary = 4
n = 1000
xy = simBVN(mux, muy, rho, varx, vary, n)
plt.scatter(xy[0], xy[1])
```

The results are plotted below in Figure 4.2a and Figure 4.2b below.



(A) One thousand observations from a bivariate normal distribution where $\rho > 0$.

(B) One thousand observations from a bivariate normal distribution where $\rho < 0$.

FIGURE 4.2. Samples from two different bivariate normal distributions.

The next proposition says that linear transformations of multivariate normal random variables are also multivariate normal random variables.

Proposition 4.8. *Let $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$, and let $A \in \mathbb{R}^{n\times d}$ be a deterministic matrix. Then, $A\boldsymbol{X} \sim \mathcal{N}_n(A\boldsymbol{\mu}, A\Sigma A^T)$.*

Proof. The proof is a homework exercise. □

This concludes our short review of the normal distribution and the multivariate normal distribution.

4.3. **Introduction to Brownian Motion.** The definition of a Brownian motion, which is given below, is deceptively simple.

Definition 4.9. *An $\mathbb{R}$-valued stochastic process $\{W_t\} = \{W_t\}_{t\geq 0}$ is said to be a* ***standard Brownian motion (SBM)*** *or* ***Wiener process*** *if:*

(1) The increments of $\{W_t\}$ are stationary and independent.
(2) For each $t \geq 0$, $W_t \sim \mathcal{N}(0,t)$.
(3) $\mathbb{P}(W_t$ is continuous at all $t \geq 0) = 1$.

Note the difference between Definition 4.9, where we introduce Brownian motion, and Definition 2.16 in Section 2.2, where we introduced the Poisson process. When defining Poisson processes, we did so in a **constructive** manner; in particular, we showed how, using exponentially-distributed random variables, one can construct a Poisson process. Thus, no work was needed for us to show that Poisson processes exist, as we provided a concrete method to construct them. On the other hand, the definition of a Brownian motion in Definition 4.9 is **descriptive**; it simply says that a stochastic process with properties (1), (2), and (3) is a Brownian motion. Surprisingly, it is very difficult to prove the existence of or construct a Brownian motion, as such proofs generally rely on techniques from measure theory and functional analysis. Consequently, a complete proof of the existence of Brownian motion is beyond the scope of this class. However, it will still be important (for simulations and applications!) for us to understand the main ideas of some of these proofs.

Before we continue, it may be helpful to take a moment to clarify what we mean when we say that a stochastic process with particular properties does not exist.

Remark 4.10. *For example, suppose that we were interested in constructing a stochastic process* $\{X_t\}$ *with the following properties:*

(i) The collection $\{X_t\}$ *consists of iid random variables with Var*$(X_t) > 0$ *(this just excludes the case when* $X_t = c$ *for some constant* $c \in \mathbb{R}$ *for all* $t \geq 0$*, i.e., when* X_t *is some fixed, deterministic constant).*

(ii) $\mathbb{P}(X_t$ *is continuous at all* $t \geq 0) = 1$.

Why does such a process not exist? Below we outline the argument showing that for each $s \geq 0$, $\mathbb{P}(X_s$ *is continuous at* $s) = 0$*, from which we see that*

$$\mathbb{P}(X_t \textit{ is continuous at all } t \geq 0) \leq \mathbb{P}(X_s \textit{ is continuous at } s) = 0.$$

To see this, fix $s \geq 0$ *and let* $\{s_n\}$ *be a sequence of time instants converging to* s*. Then, using the fact that Var*$(X_{s_n}) > 0$ *for all* n*, we can show that with* $\mu \doteq \mathbb{E}(X_s) = \mathbb{E}(X_{s_n})$*, there is some* $\epsilon > 0$ *such that*

$$\mathbb{P}(X_{s_n} > \mu + \epsilon) > 0, \textit{ and } \mathbb{P}(X_{s_n} < \mu - \epsilon) > 0.$$

From the second Borel-Cantelli lemma, since the X_{s_n} *are iid, it follows that*

$$\mathbb{P}(X_{s_n} > \mu + \epsilon \textit{ for infinitely many } n \in \mathbb{N}) = \mathbb{P}(X_{s_n} < \mu - \epsilon \textit{ for infinitely many } n \in \mathbb{N}) = 1. \tag{29}$$

It follows that

$$\mathbb{P}(X_{s_n} \to X_s \textit{ as } n \to \infty) = 0,$$

since (29) *ensures that, with probability 1, the sequence* $\{X_{s_n}\}$ *does not converge.*

Now we can begin talking about one of the approaches that one can take to show that the Brownian motion does exist, namely, that there are stochastic processes satisfying the conditions in Definition 4.9. We begin by considering iid random variables $\{\xi_n\}_{n\in\mathbb{N}}$ satisfying

$$\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = \frac{1}{2}.$$

Then the discrete-time stochastic process $\{S_n\}_{n\in\mathbb{N}_0}$ defined by

$$\begin{aligned} S_0 &= 0 \\ S_n &= \sum_{i=1}^{n} \xi_i, \quad n \in \mathbb{N}. \end{aligned} \tag{30}$$

is a simple symmetric random walk. Note that over an interval $[0, k]$, where $k \in \mathbb{N}$, the process $\{S_n\}_{n=0}^{k}$ makes a total of k jumps, and each jump has a size of 1. Using $\{S_n\}_{n\in\mathbb{N}_0}$, we can construct a collection of continuous-time stochastic process that "rescale" time and space by considering what happens if the process were to take many small jumps in a fixed time interval. We begin by defining the continuous-time process $\{Y_t\}$ obtained by linear interpolation of $\{S_n\}$:

$$Y_t = S_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)\xi_{\lfloor t \rfloor + 1}, \quad t \geq 0. \tag{31}$$

where $\lfloor x \rfloor$ denotes the integer part of a real number x; in particular, for $x \in \mathbb{R}$, we have

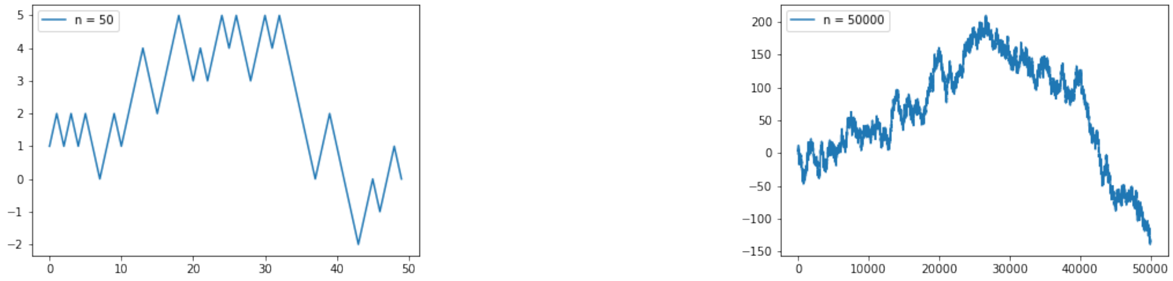
$$\lfloor x \rfloor \doteq \sup\{n \in \mathbb{N} : n \leq x\}.$$

Observe that at time $k \in \mathbb{N}_0$,

$$Y_k = S_k + (k - \lfloor k \rfloor)\xi_{\lfloor k \rfloor + 1} = S_k, \tag{32}$$

and that from time k to time $k+1$, Y_t simply follows a straight line from S_k to S_{k+1}. Thus $\{Y_t\}$ is the process obtained by linear interpolation of $\{S_n\}$.

In Figure 4.3, we plot two realizations of the process $\{Y_t\}$; the first realization is on the time interval $[0, 50]$ (so that there are a total of 50 jumps) and the second realization is on the time interval $[0, 50,000]$ (so that there are a total of 50,000 jumps).



(A) A realization of the process $\{Y_t\}$ on the time interval $[0, 50]$.

(B) A realization of the process $\{Y_t\}$ on the time interval $[0, 50000]$.

FIGURE 4.3. Two realizations of the process $\{Y_t\}$ on different time intervals. Note the difference in the values on the y-axis.

The code for the simulation and plots is provided below.

```
import numpy as np
import matplotlib.pyplot as plt

def plotYt(n):
    rw = np.cumsum(np.random.choice(a = [-1,1], size = n, replace = True, p = [0.5,0.5]))
    rw = np.array(rw)
    rw = np.insert(rw, 0, 0, axis=0)
    plt.plot(rw)

plotYt(50)
plt.legend(loc = "upper left")
plotYt(50000)
plt.legend(loc = "upper left")
```

As mentioned above, we are interested in what happens when the process undergoes a large number of small jumps in each fixed time interval. Accordingly, it will be helpful to, for each $n \in \mathbb{N}$, define the continuous-time process $\{X_t^{(n)}\}_{t\geq 0}$ by

$$X_t^{(n)} \doteq \frac{1}{\sqrt{n}} Y_{nt}, \quad t \geq 0. \tag{33}$$

In Figure 4.4 we plot a single realization of the processes $\{X_t^{(50)}\}_{t\in[0,1]}$, $\{X_t^{(500)}\}_{t\in[0,1]}$, and $\{X_t^{(50000)}\}_{t\in[0,1]}$ on the interval $[0,1]$. Note that y-axis has been scaled down for each of the processes, so that they all "live," more or less, within the same range of values. Similarly, the x-axis has been rescaled; the process $\{X_t^{(n)}\}_{t\in[0,1]}$ is generated (for $n = 50, 500, 50000$) from a random walk that undergoes n jumps in the interval $[0,1]$.

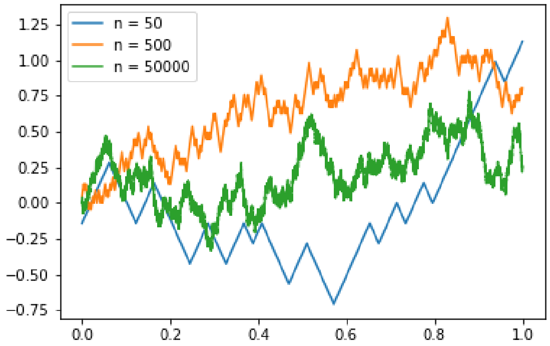


FIGURE 4.4. The continuous-time processes obtained by linearly interpolating a random walk that takes n jumps of size $\frac{1}{\sqrt{n}}$ in the time interval $[0,1]$.

The Python code used to generate the plots is below (note that since the processes are random, each time you run the code you may get a very different looking plot).

```
import numpy as np
import matplotlib.pyplot as plt

def plotXnt(n):
    rw = np.cumsum(np.random.choice(a = [-1,1], size = n, replace = True, p = [0.5,0.5]))
    rw = np.array(rw)
    rw = np.insert(rw, 0, 0, axis=0)
    rwScaled = [x / n**.5 for x in rw]
    time = [i / (n) for i in range(0,n+1)]
    plt.plot(time, rwScaled, label = "n = " "{}".format(n))

k = [50,500,50000]

for n in k:
    plotXnt(n)
    plt.legend(loc = "upper left")
```

The factor of $\frac{1}{\sqrt{n}}$ "shrinks" each jump by a factor of $\sqrt{n}$, and the fact that we are looking at the nt^{th} time instant of $\{Y_t\}$ says that time (and therefore the number of jumps) is being sped up by a factor of n. For a concrete example, consider the process $\{X_t^{(5)}\}$; each jump of the underlying discrete-time process is of size $\frac{1}{\sqrt{5}}$, and, at time $t = 1$, there have been a total of $nt = 5 \cdot 1$ jumps. Similarly, if we consider $\{X_t^{(10)}\}$, then each jump is of size $\frac{1}{\sqrt{10}}$, and at time $t = 1$, there have been a total of $nt = 10 \cdot 1$ jumps. Thus, for large n, $\{X_t^{(n)}\}$ is the continuous-time process obtained by linear interpolation of a random walk that takes many small jumps.

Additionally, we know from PSTAT 160A that random walks have stationary and independent increments, so it follows that each process $\{X_t^{(n)}\}$ has stationary and independent increments as well. For a more detailed discussion of this point, see Section D.1. Additionally, from (32) and (33) we see that if $t = \frac{k}{n}$ for some $k \in \mathbb{N}$, then

$$X_t^{(n)} = \frac{1}{\sqrt{n}} Y_{n\frac{k}{n}} = \frac{1}{\sqrt{n}} Y_k = \frac{S_k}{\sqrt{n}},$$

which says that

$$\mathbb{E}(X_t^{(n)}) = \mathbb{E}\left(\frac{S_k}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \mathbb{E}(\xi_i) = 0,$$

and

$$\mathrm{Var}(X_t^{(n)}) = \mathrm{Var}\left(\frac{S_k}{\sqrt{n}}\right) = \frac{1}{n} \sum_{i=1}^{k} \mathrm{Var}(\xi_i) = \frac{k}{n}.$$

Additionally, if k is large and n is large and $t = \frac{k}{n}$, then $X_t^{(n)}$ is the sum of a large number of iid random variables (namely, $\{\xi_i, 1 \le i \le k\}$), and according to the central limit theorem and our observations above, is approximately a $\mathcal{N}\left(0, t = \frac{k}{n}\right)$ random variable.

From the above discussion we have seen that for each $n \in \mathbb{N}$, $\{X_t^{(n)}\}_{t \ge 0}$ has stationary and independent increments, and, if n is large, that $X_t^{(n)}$ is approximately a $\mathcal{N}(0, t)$ random variable. Additionally, each $\{X_t^{(n)}\}$ is a continuous function. This suggests that, as $n \to \infty$, the processes $\{X_t^{(n)}\}_{t \ge 0}$ should, in some appropriate sense, converge to process satisfying the conditions in Definition 4.9. That is, as $n \to \infty$, $\{X_t^{(n)}\}_{t \ge 0}$ should converge to a Brownian motion $\{W_t\}_{t \ge 0}$. The details of this argument are very technical and are typically encountered in advanced graduate courses in probability; for example, it is not even obvious what it "should" mean for a sequence of stochastic processes to converge to another stochastic process. [1]

Remark 4.11. *The key takeaways of the discussion above are as follows:*

1. *A process satisfying the conditions in Definition* 4.9 *exists. Namely, Brownian motion exists.*
2. *We can (approximately) generate a Brownian motion by sampling iid random variables. In particular, if n is large and $\{X_t^{(n)}\}$ is defined as above, then $\{X_t^{(n)}\}$ is "approximately" a Brownian motion.*

In Figure 4.5 we apply Donsker's theorem to plot three (approximate) realizations of a Brownian motion.

At the beginning of this section we mentioned that Brownian motion is analogous to the normal distribution, as it shows up "universally" as the limit of a large class of random walks. This is summarized in the result below, known as Donsker's Theorem or the functional central limit theorem.

[1] For those interested, this is discussed in detail in Chapters 1 and 2 of this book.

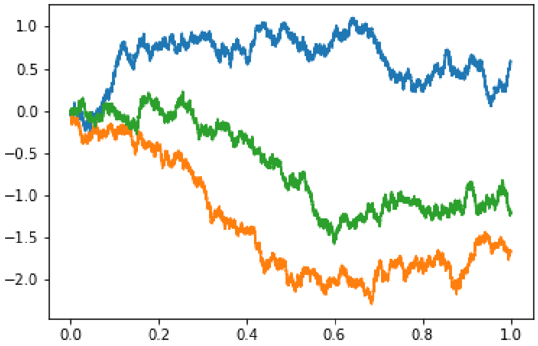


FIGURE 4.5. Three (approximate) realizations of a Brownian motion on the time interval $[0,1]$. In particular, we generate three realizations of $\{X_t^{(n)}\}$ on the time interval $[0,1]$, where $n = 100000$.

Theorem 4.12. *Let $\{\xi_i\}_{i=1}^{\infty}$ be an iid sequence of random variables with $\mathbb{E}(\xi_i) = \mu$ and $Var(\xi_i) = \sigma^2 < \infty$. For each $i \in \mathbb{N}$, let $\eta_i \doteq \xi_i - \mu$, so that $\mathbb{E}(\eta_i) = 0$, and for each $n \in \mathbb{N}$ define*

$$S_n \doteq \sum_{i=1}^{n} \eta_i,$$

and define the continuous time processes $\{X_t^{(n)}\}_{t\geq 0}$ by

$$X_t^{(n)} \doteq \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor} + (nt - \lfloor nt \rfloor)\frac{1}{\sigma\sqrt{n}} \eta_{\lfloor nt \rfloor}.$$

Then, as $n \to \infty$, the processes $\{X_t^{(n)}\}_{t\geq 0}$ converge (in an appropriate sense) to a Brownian motion $\{W_t\}_{t\geq 0}$. Consequently, Brownian motion exists.

Now that we have seen that Brownian motion does in fact exist, in the next section we will start investigating some of the consequences of Definition 4.9.

4.4. **Fundamental Properties of Brownian Motion.**

We begin by proving a few properties of Brownian motion.

4.4.1. *Brownian Motion as a Markov Process.*

In this section we discuss the Markov property in the context of Brownian motion.

Proposition 4.13. *Let $\{W_t\}_{t\geq 0}$ be an SBM. The following hold:*

(1) $\mathbb{P}(W_0 = 0) = 1$, meaning that the process starts at 0.
(2) $\{W_t\}$ has the Markov property.

Proof. (1) By definition, if $W_0 \sim \mathcal{N}(0,0)$, which means that $\mathbb{P}(W_0 = 0) = 1$.

(2) Observe that, for $s, t \geq 0$, and $y, x \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(W_{t+s} \leq x | W_s = y, W_u = x_u \text{ for } u \in [0,s]) &= \mathbb{P}(W_{t+s} - W_s \leq x - y | W_s = y, W_u = x_u \text{ for } u \in [0,s]) \\ &\overset{1}{=} \mathbb{P}(W_{s+t} - W_s \leq x - y) \\ &\overset{2}{=} \mathbb{P}(W_t - W_0 \leq x - y) \\ &= \Phi_{0,t}(x-y)\end{aligned}$$

where $\overset{1}{=}$ is due to the independent increments property, $\overset{2}{=}$ is due to the stationary increments property, and

$$\Phi_{\mu,\sigma^2}(z) \doteq \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx, \quad z \in \mathbb{R},$$

denotes the cdf of the $\mathcal{N}(\mu, \sigma^2)$ distribution. Additionally, using the independent increments property and then the stationary increments property of Brownian motion, we see that

$$\begin{aligned}\mathbb{P}(W_{t+s} \leq y | W_s = x) &= \mathbb{P}(W_{t+s} - W_s \leq y - x | W_s = x) \\ &= \mathbb{P}(W_{t+s} - W_s \leq y - x) \\ &= \mathbb{P}(W_t \leq y - x) \\ &= \Phi_{0,t}(y-x) \\ &= \int_{-\infty}^{y-x} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{z^2}{2t}\right) dt \\ &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(z-x)^2}{2t}\right) dt\end{aligned}$$

Thus, for all $s, t \geq 0$ and $x, y \in \mathbb{R}$,

$$\mathbb{P}(W_{t+s} \leq y | W_s = x, W_u = x_u \text{ for } u \in [0,s]) = \mathbb{P}(W_{t+s} \leq y | W_s = x),$$

which shows that Markov's property holds.

□

Note that even though Brownian motion is a continuous-time process with the Markov property, it is quite different from the CTMCs we studied in Section 3. As time progresses, a Brownian motion moves through its state space, $\mathbb{R}$, continuously, meaning that it never jumps between two states. On the other hand, CTMCs move through their state space exclusively by jumping from state to state (at random time instants). Accordingly, we will abstain from referring to Brownian motion as a CTMC and will instead refer to it as a **(continuous-time) Markov process**. Note, from the proof of Proposition 4.13, that the CDF of the transition probabilities of an SBM are given by

$$\mathbb{P}(W_{t+s} \leq y | W_s = x) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(z-x)^2}{2t}\right) dz$$

This says that the conditional cdf of W_{t+s} given that $W_s = y$ is given by

$$F_t(x|y) \doteq \int_{0}^{x} \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(z-y)^2}{2t}\right) dz$$

Differentiating, we see that the conditional pdf of W_{t+s} given that $W_s = x$ is given by

$$K_t(x,y) \doteq \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(x-y)^2}{2t}\right) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right)$$

We refer to the functions $\{K_t, t \geq 0\}$ as the **transition kernel** of $\{W_t\}$. In particular, we have that, for any event $A \subseteq \mathbb{R}$,

$$\mathbb{P}(W_{s+t} \in A | W_s = x) = \int_A K_t(x,y) dy.$$

For notational convenience, we often write

$$K_t(x,A) \doteq \int_A K_t(x,y)\,dy, \quad A \subseteq \mathbb{R}.$$

The next result shows that the Chapman-Kolmogorov equations hold for Brownian motion.

Proposition 4.14. *Let K denote the transition kernel of an SBM. Then, for $s,t \geq 0$,*

$$K_{t+s}(x,y) = \int_{-\infty}^{\infty} K_s(x,z)K_t(z,y)\,dz.$$

Proof. Observe that, for $y \in \mathbb{R}$, since $K_s(x,z)$ is the conditional density of W_s given that $W_0 = x$,

$$\begin{aligned}\int_{-\infty}^{y} K_{s+t}(x,z)\,dz &= K_{s+t}(x,(-\infty,y]) \\ &= \mathbb{P}(X_{s+t} \leq y | X_0 = x) \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X_{s+t} \leq y | X_0 = x, X_s = z)K_s(x,z)\,dz \\ &= \int_{-\infty}^{\infty} \mathbb{P}(X_t \leq y | X_0 = z)K_s(x,z)\,dz \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} K_t(z,w)\,dw\right) K_s(x,z)\,dz \\ &= \int_{-\infty}^{y} \left(\int_{-\infty}^{\infty} K_s(x,z)K_t(z,w)\,dz\right) dw.\end{aligned}$$

Differentiating both sides with respect to y, we see that

$$K_{s+t}(x,z) = \int_{-\infty}^{\infty} K_s(x,z)K_t(z,y)\,dz.$$

□

The following result illustrates how we can modify one Brownian motion to construct several others.

Lemma 4.15. *Let $\{W_t\}$ be an SBM.*

*(1) The **mirrored** process $\{\check{W}_t\}_{t\geq 0}$ defined by*

$$\check{W}_t = -W_t, \quad t \geq 0,$$

is an SBM.

*(2) The **time-inverted** process $\{\bar{W}_t\}_{t\geq 0}$ defined by $\bar{W}_t = 0$ and*

$$\bar{W}_t = tW_{\frac{1}{t}}, \quad t > 0.$$

is an SBM.

*(3) For each $T > 0$, the **time-reversed** process $\{\tilde{W}_t\}_{t\geq 0}$ defined by,*

$$\tilde{W}_t \doteq W_T - W_{T-t}, \quad t \in [0,T],$$

is an SBM on $[0,T]$.

*(4) For each $\alpha > 0$, the **scaled** process $\{\hat{W}_t\}_{t\geq 0}$ defined by*

$$\hat{W}_t \doteq \frac{1}{\sqrt{\alpha}} W_{\alpha t}, \quad t \geq 0,$$

is an SBM.

Proof. We prove only part (1); the other parts are either similar or are beyond the scope of this class.

(1) We need to show that $\{\check{W}_t\}$ satisfies the conditions in Definition 4.9. First, note that if $t_1 < t_2 \leq t_3 < t_4$, then

$$\begin{aligned}\mathbb{P}(\check{W}_{t_4} - \check{W}_{t_3} \leq x, \check{W}_{t_2} - \check{W}_{t_1} \leq y) &= \mathbb{P}(-(W_{t_4} - W_{t_3}) \leq x, -(W_{t_2} - W_{t_1}) \leq y) \\ &= \mathbb{P}(-x \leq W_{t_4} - W_{t_3}, -y \leq W_{t_2} - W_{t_1}) \\ &= \mathbb{P}(-x \leq W_{t_4} - W_{t_3})\mathbb{P}(-y \leq W_{t_2} - W_{t_1}) \\ &= \mathbb{P}(\check{W}_{t_4} - \check{W}_{t_3} \leq x)\mathbb{P}(\check{W}_{t_2} - \check{W}_{t_1} \leq y)\end{aligned}$$

where the second-to-last equality used the independent increments property of $\{W_t\}$. This shows that $\{\check{W}_t\}$ has independent increments. To show that $\{\check{W}_t\}$ has stationary increments, fix $s_1 < s_2$ and $t_1 < t_2$ such that $t_2 - t_1 = s_2 - s_1$ and observe that, for each $x \in \mathbb{R}$,

$$\mathbb{P}(\check{W}_{t_2} - \check{W}_{t_1} \leq x) = \mathbb{P}(-x \leq W_{t_2} - W_{t_1}) = \mathbb{P}(-x \leq W_{s_2} - W_{s_1}) = \mathbb{P}(\check{W}_{s_2} - \check{W}_{s_1} \leq x),$$

where the second-to-last equality is due to the stationary increments property of $\{W_t\}$. This shows that $\{\check{W}_t\}$ has stationary increments. Additionally, we know that for $\alpha \in \mathbb{R}$, if $X \sim \mathcal{N}(0, \sigma^2)$, then $\alpha X \sim \mathcal{N}(0, (\alpha\sigma)^2)$, so it follows that $\check{W}_t \sim \mathcal{N}(0, t)$. Finally, we note that if a function $f : \mathbb{R}_+ \to \mathbb{R}$ is continuous at all $t \geq 0$, then the function $g : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$g(t) \doteq -f(t),$$

is also continuous at all $t \geq 0$, so it follows that $\mathbb{P}(\check{W}_t \text{ is continuous for all } t \geq 0) = 1$.

□

The following result shows how we can calculate the covariance between two time instants of an SBM.

Lemma 4.16. *Let $\{W_t\}$ be an SBM. Then*

$$\mathit{Cov}(W_s, W_t) = \min\{s, t\}.$$

Proof. The proof is similar to the homework problem dealing with the covariance of two time instants of a Poisson process. First, note that, if $t \geq s$, then, using the independent increments property and the fact that $W_u \sim \mathcal{N}(0, u)$, we have

$$\mathbb{E}(W_s W_t) = \mathbb{E}(W_s(W_t - W_s + W_s)) = \mathbb{E}(W_s(W_t - W_S)) + \mathbb{E}(W_s^2) = \mathbb{E}(W_s)\mathbb{E}(W_t - W_s) + \mathbb{E}(W_s^2) = 0 + s = s.$$

Therefore,

$$\text{Cov}(W_s, W_t) = \mathbb{E}(W_s W_t) - \mathbb{E}(W_s)\mathbb{E}(W_t) = s - 0 = s.$$

Similarly, if $s \geq t$, then $\text{Cov}(W_s, W_t) = t$, so it follows that $\text{Cov}(W_s, W_t) = \min\{s, t\}$. □

Sample paths of a Brownian motion exhibit many fascinating and unusual properties. Perhaps most interestingly, even though every Brownian motion sample path is a continuous function, it is also true that every path is not differentiable at any point. That is, at every $t \geq 0$, the sample path at t is so irregular that its derivative does not exist.

Proposition 4.17. *Let $\{W_t\}$ be an SBM. Then*

$$\mathbb{P}(\textit{for all } t \geq 0, \textit{ the function } t \mapsto W_t \textit{ is not differentiable at } t) = 1.$$

Proof. While a detailed proof of this proposition is beyond the scope of this course, we provide a heuristic argument. Note that due to the stationary increments property of Brownian motion, for each $t, h > 0$,

$$\frac{W_{t+h} - W_t}{h} \overset{d}{=} \frac{W_h}{h} \sim \mathcal{N}(0, h^{-1}).$$

This means that for small values of h, the different quotient above is a random variable with very large variance, so on a small window $[t, t+h]$, we expect that the rate of change of the Brownian motion will, on average, tend to become arbitrarily large in magnitude. Thus, we expect that the limit of these different quotients does not exist. □

Proposition 4.17 tells us that every sample path of a Brownian motion is so irregular that it cannot be differentiated at any point. At this point, it is natural to wonder whether this is possible due to the fact that $\{W_t\}$ is a stochastic process. However, there are well-known examples of deterministic functions that are also continuous at every point, but differentiable at no point. One such example, known as the Weierstrass function, is given below.

Proposition 4.18. *Let $b \in \mathbb{N}$ and $a \in (0,1)$ be such that b is odd and $ab > 1 + \frac{3}{2}\pi$. Then, the function $f : \mathbb{R} \to \mathbb{R}$ defined by*

$$f(x) \doteq \sum_{n=0}^{\infty} a^n \cos(b^n \pi x),$$

is continuous at all $x \in \mathbb{R}$, but is differentiable nowhere.

Just like a Brownian motion sample path, the Weierstrass function is an example of a function with a *fractal* structure.

Below we introduce the notion of a Brownian motion started at a point other than 0.

Definition 4.19. *Fix $x \in \mathbb{R}$ and let $\{W_t\}$ be a SBM. We say that the stochastic process $\{X_t\}$ defined by*

$$X_t \doteq x + W_t,$$

is a Brownian motion started at x, since $X_0 = x + W_0 = x$.

Below we illustrate how we can calculate probabilities involving a Brownian motion started at some $x \in \mathbb{R}$.

Example 4.20. *Let $\{X_t\}$ be a Brownian motion started at 1. Then, if we let $\Phi_{0,5}$ denote the cdf of a $\mathcal{N}(0,5)$ random variable, we have that*

$$\mathbb{P}(X_5 \geq 3) = \mathbb{P}(1 + W_5 \geq 3) = \mathbb{P}(W_5 \geq 2) = 1 - \Phi_{0,5}(2) \approx 1 - 0.8144533 = 0.1855467.$$

4.5. **First Hitting Time of Brownian Motion.** In this section we analyze the *first hitting time* of a Brownian motion. We begin by recalling the definition of the first hitting time of a stochastic process.

Definition 4.21. *Let $\{X_t\}$ be a stochastic process. The **first fitting time** of a state $a \in \mathbb{R}$ is the random variable*

$$T_a \doteq \inf\{t \geq 0 : X_t = a\}.$$

The next result characterizes the distribution of the first hitting time of an SBM.

Lemma 4.22. *Let $\{W_t\}$ be an SBM, and let $T_a \doteq \inf\{t \geq 0 : W_t = a\}$, for some $a \in \mathbb{R}$. Then, the CDF of T_a is given by*

$$\mathbb{P}(T_a \leq t) = 2\mathbb{P}(W_t \geq |a|) = 2\left(1 - \Phi\left(\frac{|a|}{\sqrt{t}}\right)\right),$$

where Φ denotes the CDF of a $\mathcal{N}(0,1)$ random variable. Consequently, the pdf of T_a is given by

$$f_{T_a}(t) = \frac{|a|}{\sqrt{2\pi t^3}} \exp\left(-\frac{a^2}{2t}\right), \quad t > 0.$$

Proof. Suppose that $a > 0$. We omit some details, but the basic idea relies on noting that, if $T_a < t$, then any time after T_a, the SBM is equally likely to be above or below a; namely that

$$\mathbb{P}(W_t > a | T_a < t) = \frac{1}{2},$$

which gives,

$$\frac{1}{2} = \mathbb{P}(W_t > a | T_a < t) = \frac{\mathbb{P}(W_t > a, T_a < t)}{\mathbb{P}(T_a < t)} = \frac{\mathbb{P}(W_t > a)}{\mathbb{P}(T_a < t)}.$$

Rearranging, we obtain

$$\mathbb{P}(T_a < t) = 2\mathbb{P}(W_t > a),$$

which completes the proof. To obtain the last part, simply differentiate with respect to t to obtain the pdf of T_a. □

The following corollary says that for each $a \in \mathbb{R}$, on average, the time that it takes an SBM to reach a is infinite.

Corollary 4.23. *Let T_a be as in Lemma 4.22. Then $\mathbb{E}(T_a) = \infty$.*

Proof. First, note that since $T_a \geq 0$, $\mathbb{E}(T_a)$ is well-defined, see Definition A.10. We have,

$$\mathbb{E}(T_a) = \int_0^\infty t f_{T_a}(t)\,dt = \frac{|a|}{\sqrt{2\pi}} \int_0^\infty t^{-1/2} \exp\left(-\frac{a^2}{2t}\right) dt.$$

Note that if

$$T \geq \frac{a^2}{2\log 2},$$

then, for all $t \geq T$,

$$\exp\left(-\frac{a^2}{2t}\right) \geq \frac{1}{2}.$$

Thus,

$$\begin{aligned}\mathbb{E}(T_a) &= \frac{|a|}{\sqrt{2\pi}} \int_0^\infty t^{-1/2} \exp\left(-\frac{a^2}{2t}\right) dt \\ &\geq \frac{|a|}{\sqrt{2\pi}} \int_T^\infty t^{-1/2} \exp\left(-\frac{a^2}{2t}\right) dt \\ &\geq \frac{|a|}{\sqrt{2\pi}} \frac{1}{2} \int_T^\infty t^{-1/2} dt,\end{aligned}$$

so the result follows. □

We note in the next corollary that Lemma 4.22 also tells us about the distribution of the maximum of a Brownian motion.

Corollary 4.24. *Let $\{W_t\}$ be an SBM. Consider the process $\{M_t\}$ defined by*

$$M_t \doteq \max_{0\le s\le t} W_t.$$

Then, for each $a > 0$,

$$\mathbb{P}(M_t \ge a) = 2\mathbb{P}(W_t \ge a).$$

Proof. Note that $\{T_a \le t\}$ if and only if $\{M_t \ge a\}$. □

The following example illustrates how to apply Lemma 4.22.

Example 4.25. *Let $\{X_t\}$ be a Brownian motion starting at 2, so that*

$$X_t = W_t + 2, \quad t \ge 0,$$

where $\{W_t\}$ is an SBM (starting at 0). What is the probability that $\{X_t\}$ hits state 3 by time 4?

Note that

$$\inf\{t\ge 0 : X_t = 3\} = \inf\{t \ge 0 : W_t + 2 = 3\} = \inf\{t \ge 0 : W_t = 1\},$$

so

$$\mathbb{P}\left(\max_{0\le t\le 4} X_t \ge 3\right) = \mathbb{P}\left(\max_{0\le t\le 4} W_t \ge 1\right) = \mathbb{P}(M_4 \ge 1) = 2\mathbb{P}(W_4 \ge 1) = 0.6170751.$$

4.6. **Brownian Motion with Drift and Scaling.** In this section we introduce a slight generalization of SBM; in particular, we consider a process whose expected value changes over time, and whose variance, at time t, may be larger (or smaller) than t.

Definition 4.26. *Let $\{W_t\}$ be an SBM. For $\mu \in \mathbb{R}$ and $\sigma > 0$, consider the process $\{X_t\}$ defined by*

$$X_t \doteq \mu t + \sigma W_t.$$

Then $\{X_t\}$ is a ***Brownian motion with drift and scaling****, which we sometimes denote as BM(μ,σ). We refer μ as the* ***drift parameter*** *and σ as the* ***scaling parameter****.*

The following proposition describes several properties such processes.

Proposition 4.27. *Let $\{X_t\}$ be a BM(μ,σ) process. Then:*

1. *$\{X_t\}$ has stationary and independent increments.*
2. *$\mathbb{P}(X_t$ is continuous at all $t \ge 0) = 1$.*
3. *For each $t \ge 0$, $X_t \sim \mathcal{N}(\mu t, \sigma^2 t)$.*
4. *For each $s, t \ge 0$, Cov$(X_s, X_t) = \sigma^2 \min\{s,t\}$.*

Proof. (1) Note that, for $s \le t$,

$$X_t - X_s = \mu(t-s) + \sigma(W_t - W_s),$$

so the independent increments property follows on recalling that $\{W_t\}$ has independent increments. The stationary increments property follows on noting that, since $\{W_t\}$ has stationary increments, for $s \le t$,

$$X_t - X_s = \mu(t-s) + \sigma(W_t - W_s) \overset{d}{=} \mu(t-s) + \sigma W_{t-s} = X_{t-s}.$$

(2) This follows immediately from the fact that $t \mapsto \mu t$ and $t \mapsto \sigma W_t$ are continuous functions.

(3) This follows from basic properties of the normal distribution; note that

$$\mathbb{E}(X_t) = \mathbb{E}(\mu t + \sigma W_t) = \mu t + \sigma \mathbb{E}(W_t) = \mu t,$$

and

$$\text{Var}(X_t) = \text{Var}(\mu t + \sigma W_t) = \sigma^2 \text{Var}(W_t) = \sigma^2 t.$$

(4) Observe that, if $s \le t$, then, from (1),

$$\begin{aligned}
\mathbb{E}(X_s X_t) &= \mathbb{E}(\mathbb{E}[X_s X_t | X_s]) \\
&= \mathbb{E}(\mathbb{E}[X_s(X_t - X_s + X_s) | X_s]) \\
&= \mathbb{E}(\mathbb{E}[X_s(X_t - X_s) + X_s^2 | X_s]) \\
&= \mathbb{E}(\mathbb{E}[X_s(X_t - X_s) | X_s]) + \mathbb{E}(\mathbb{E}[X_s^2 | X_s]) \\
&= \mathbb{E}(X_s \mathbb{E}[X_t - X_s | X_s]) + \mathbb{E}(X_s^2) \\
&= \mathbb{E}(X_s \mathbb{E}(X_t - X_s)) + \mathbb{E}(X_s^2) \\
&= \mathbb{E}(X_s \mu(t - s)) + \mathbb{E}(X_s^2) \\
&= \mu^2 s(t - s) + (\sigma^2 s + (\mu s)^2) \\
&= \mu^2 st + \sigma^2 s^2.
\end{aligned}$$

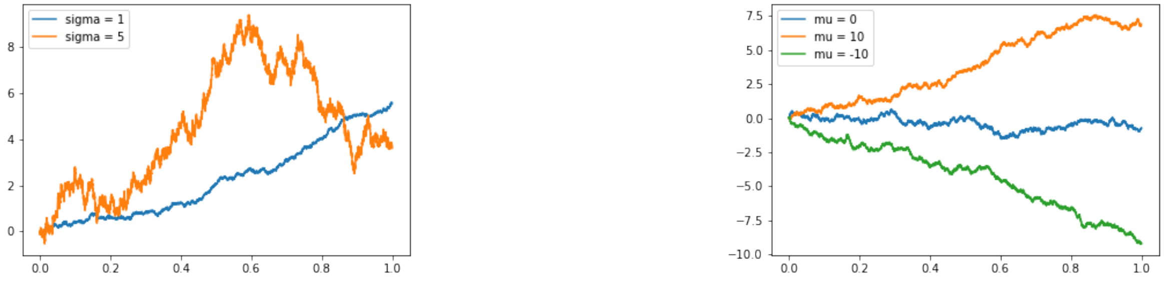
Thus,

$$\text{Cov}(X_s, X_t) = \mathbb{E}(X_s X_t) - \mathbb{E}(X_s)\mathbb{E}(X_t) = \mu^2 st + \sigma^2 s - \mu^2 st = \sigma^2 s.$$

The result follows.

□

Note that if $\mu > 0$, then the process tends to, over time, move upwards, and if $\mu < 0$, then it tends to move downwards. In Figure 4.6a we plot an example of Brownian motion with drift and scaling.



(A) Realization of Brownian motions with drift parameter $\mu = 5$ and scaling parameters $\sigma = 1$ and $\sigma = 5$.

(B) Realization of Brownian motions with scaling parameter $\sigma = 2$ and drift parameters $\mu = 0$, $\mu = -10$, and $\mu = 10$.

FIGURE 4.6. Observe the effect that the choice of scaling parameter and drift parameter has on the process.

The Python code used to simulate the processes is given below.

```
def simBmDriftScaling(n, mu, sigma):
    rw = np.cumsum(np.random.choice(a = [-1,1], size = n, replace = True, p = [0.5,0.5]))
    rw = np.array(rw)
    rw = np.insert(rw, 0, 0, axis=0)
    bm = [x / n**.5 for x in rw]
    Scaledbm = [x*sigma for x in bm]
    time = [i / (n) for i in range(0,n+1)]
    DriftScaledbm = Scaledbm + np.multiply(time, mu)
    plt.plot(time, DriftScaledbm)

sigma = [1,5]
for sig in sigma:
    simBmDriftScaling(10000,5,sig)

mu = [0,10, -10]
for m in mu:
    simBmDriftScaling(10000,m,2)
```

Brownian motions with scaling and drift are, as we see in Section 6, are some of the simplest example of processes satisfying stochastic differential equations (SDE). In particular, a BM(μ,σ) process solves the SDE

$$dX_t = \mu dt + \sigma dW_t, \tag{34}$$

where $\{W_t\}$ is an SBM. Note that it is not yet clear what is meant by the equation in (34), especially since, as we saw in Proposition 4.17, an SBM is nowhere differentiable.

4.7. **Brownian Bridge.** In this section we study another stochastic process related to Brownian motion; this one, known as a Brownian bridge, essentially behaves like a Brownian motion that is forced to return to 0 at some time instant. First, we introduce the notion of a Gaussian process.

Definition 4.28. *A stochastic process $\{X_t\}$ is said to be a **Gaussian process** if, for all $0 \le t_1 < \cdots < t_d$, $(X_{t_1}, \ldots, X_{t_d})$ follows a multivariate normal distribution.*

The following proposition says that an SBM is a Gaussian process.

Proposition 4.29. *Let $\{W_t\}$ be an SBM. Then $\{W_t\}$ is a Gaussian process.*

Proof. Fix time instants $0 = t_0 < t_1 < \cdots < t_d$, and constants $a_1, \ldots, a_d \in \mathbb{R}$. Then, if we define

$$\alpha_i \doteq \sum_{j=i}^{d} a_j,$$

we have

$$\sum_{i=1}^{d} a_i W_{t_i} = \sum_{i=1}^{d} \alpha_i (W_{t_i} - W_{t_{i-1}}),$$

so, using the stationary and independent increments properties of $\{W_t\}$, we see that $\sum_{i=1}^{d} a_i W_{t_i}$ follows a normal distribution[2], which means that $(W_{t_1}, \ldots, W_{t_d})$ is a multivariate normal random variable. The result follows. □

Now we define a Brownian bridge.

[2] What are the mean and variance?

Definition 4.30. *A stochastic process $\{B_t\}_{0\le t\le 1}$ is said to be a Brownian bridge (on $[0,1]$) if the following hold:*

(1) $\mathbb{P}(B_0 = 0) = \mathbb{P}(B_1 = 0) = 1$.
(2) $\{B_t\}$ *is a Gaussian process.*
(3) $\mathbb{E}(B_t) = 0$ *for all* $t \in [0,1]$.
(4) *Cov*$(B_s, B_t) = \min\{s,t\} - st$ *for* $s,t \in [0,1]$.
(5) $\mathbb{P}(B_t$ *is continuous at all* $t \ge 0) = 1$.

The next result shows how to construct a Brownian bridge with a Brownian motion.

Theorem 4.31. *Let $\{W_t\}$ be an SBM. The process $\{B_t\}_{0\le t\le 1}$, defined by*

$$B_t \doteq W_t - tW_1, \quad t \in [0,1],$$

is a Brownian bridge.

Proof. We show that the conditions in Definition 4.30 hold.

(1) Note that $B_0 = W_0 - 0\cdot W_1 = 0$ and $B_1 = W_1 - W_1 = 0$.
(2) This follows from the fact that $\{W_t\}$ is a Gaussian process.
(3) $\mathbb{E}(B_t) = \mathbb{E}(W_t) - t\mathbb{E}(W_1) = 0$.
(4) Note that, with $s \le t$, using the fact that $\mathbb{E}(W_s W_t) = s$,

$$\begin{aligned}\mathrm{Cov}(B_s, B_t) &= \mathbb{E}(B_s B_t) - \mathbb{E}(B_s)\mathbb{E}(B_t)\\ &= \mathbb{E}[(W_s - sW_1)(W_t - tW_1)]\\ &= \mathbb{E}(W_s W_t) - s\mathbb{E}(W_1 W_t) - t\mathbb{E}(W_s W_1) + st\mathbb{E}(W_1 W_1)\\ &= s - st - st + st\\ &= s - st.\end{aligned}$$

The result follows.
(5) This follows from the fact that, with probability 1, the functions $t \mapsto W_t$ and $t \mapsto tW_1$ are continuous.

□

Before we discuss more process related to Brownian motion in Section 6, we will briefly introduce martingales.

5. Martingales

In this section we provide a brief introduction to martingales, which are a class of stochastic processes that are particularly amenable to analysis. Indeed, martingales (respectively, submartingales and supermartingales) can be thought of as stochastic analogues of constant (respectively, non-decreasing and non-increasing) sequences. Recall, from calculus, the following fact.

Proposition 5.1. *Let $\{a_n\}$ be a non-decreasing sequence of real numbers that is bounded above by some M, so that for all $n \in \mathbb{N}$, $a_n \leq a_{n+1} \leq M$, then the sequence $\{a_n\}$ converges.*

Similarly, if $\{b_n\}$ is a non-increasing sequence of real numbers bounded below by some L, so that for all $n \in \mathbb{N}$, $b_n \geq b_{n+1} \geq L$, then the sequence $\{b_n\}$ converges.

We begin by introducing some relevant definitions.

Definition 5.2. *A sequence of random variables $\{X_n\}_{n\in\mathbb{N}_0}$ is called a **(discrete-time) martingale** (with respect to the natural filtration) if the following hold:*

(1) For all $n \in \mathbb{N}_0$, $\mathbb{E}(|X_n|) < \infty$.

(2) For each $n \in \mathbb{N}_0$, $\mathbb{E}[X_{n+1}|X_0, X_1, \dots, X_n] = X_n$.

If instead of (2) *we have that*

$$\mathbb{E}[X_{n+1}|X_0, X_1, \dots, X_n] \geq X_n \quad n \in \mathbb{N}_0,$$

*then the process is a **submartingale**. Similarly, if instead of* (2) *we have that*

$$\mathbb{E}[X_{n+1}|X_0, X_1, \dots, X_n] \leq X_n \quad n \in \mathbb{N}_0,$$

*then the process is a **supermartingale**.*

*Similarly, a continuous-time stochastic process $\{Y_t\}_{t\geq 0}$ is called a **(continuous-time) martingale** (with respect to the natural filtration) if the following hold:*

(1) For all $t \geq 0$, $\mathbb{E}(|Y_t|) < \infty$.

(2) For each $0 \leq s \leq t$, $\mathbb{E}[Y_t|Y_u, u \in [0,s]] = Y_s$.

If instead of (2) *we have that*

$$\mathbb{E}[Y_t|Y_u, u \in [0,s]] \geq Y_s, \quad 0 \leq s \leq t,$$

*then the process is a **submartingale**. Similarly, if instead of* (2) *we have that*

$$\mathbb{E}[Y_t|Y_u, u \in [0,s]] \leq Y_s, \quad 0 \leq s \leq t,$$

*then the process is a **supermartingale**.*

Note that a process is a martingale if and only if it is a supermartinagle and a submartingale.

Intuitively, one can think of a martingale as the wealth that a gambler has during a sequence of fair bets. A submartingale represents the wealth a gambler has during a sequence of bets that are biased towards them (i.e., the gambler's tends to increase, on average), while a supermartingale represents the wealth a gambler has during a sequence of bets that are biased against them (i.e., the gambler's wealth tends to decrease, on average). This is shown in the example below.

Example 5.3. *A gambler is playing a game of chance, where each round of the game is independent of the others. Each time they play, the probability that they win is p, and the probability that they lose is $1-p$. Suppose that they bet 10 dollars on each game. If we let X_n denote the player's wealth after the n-th game, then for what value(s) of p is $\{X_n\}$ a martingale? For what value(s) is it a submartingale or supermartingale?*

Note that $X_0 = 50$, and, in general, if we let E_{n+1} denote the winnings/losses from the $n+1$ game, then

$$\mathbb{P}(E_{n+1} = 10) = p, \quad \mathbb{P}(E_{n+1} = -10) = 1-p,$$

and

$$X_{n+1} = X_n + E_{n+1} = \sum_{i=1}^{n+1} E_{n+1}.$$

Observe that, for each $n \in \mathbb{N}$,

$$\mathbb{E}(|X_{n+1}|) = \mathbb{E}\left(\left|\sum_{i=1}^{n+1} E_{n+1}\right|\right) \leq \sum_{i=1}^{n+1} \mathbb{E}(|E_{n+1}|) = 10(n+1) < \infty.$$

Since E_{n+1} is independent of $X_0, X_1, \dots, X_n$, we have

$$\mathbb{E}[X_{n+1}|X_0,\dots,X_n] = \mathbb{E}[X_n + E_{n+1}|X_0, X_1,\dots,X_n] = X_n + \mathbb{E}[E_{n+1}] = X_n + 20p - 10.$$

If $p = \frac{1}{2}$, we obtain

$$\mathbb{E}[X_{n+1}|X_0,\dots,X_n] = X_n,$$

meaning that $\{X_n\}$ is a martingale. If $p < \frac{1}{2}$ (i.e., if the player is more likely to lose than win), then $20p - 10 < 0$, so

$$\mathbb{E}[X_{n+1}|X_0,\dots,X_n] = X_n + 20p - 10 \leq X_n,$$

so $\{X_n\}$ is a supermartingale. Finally, if $p > \frac{1}{2}$ (i.e., if the player is more likely to win than lose), then $20p - 10 > 0$, so

$$\mathbb{E}[X_{n+1}|X_0,\dots,X_n] = X_n + 20p - 10 \geq X_n,$$

so $\{X_n\}$ is a submartingale.

Now we provide examples of continuous-time sub/super/martingales.

Example 5.4. *Let $\{W_t\}$ be an SBM, and let, for some $\mu \in \mathbb{R}$,*

$$X_t \doteq \mu t + W_t,$$

so that $\{X_t\}$ is a Brownian motion with drift parameter μ. Note that for all $t \geq 0$,

$$\mathbb{E}[|X_t|] = \mathbb{E}[|\mu t + W_t|] \leq |\mu| t + \mathbb{E}[|W_t|] = |\mu| t + \sqrt{\frac{2t}{\pi}} < \infty.$$

Additionally, since $\{X_t\}$ has independent increments (see Proposition 4.27), for $0 \leq s \leq t$,

$$\begin{aligned}\mathbb{E}[X_t|X_u, u \in [0,s]] &= \mathbb{E}[X_t - X_s + X_s|X_u, u \in [0,s]] \\ &= \mathbb{E}[X_t - X_s] + X_s \\ &= \mu(t-s) + X_s.\end{aligned}$$

If $\mu = 0$, then $X_t = W_t$, and we see that an SBM is a martingale. If $\mu > 0$, then

$$\mathbb{E}[X_t|X_u, u \in [0,s]] = \mu(t-s) + X_s \geq X_s,$$

so $\{X_t\}$ is a submartingale. Finally, if $\mu < 0$,

$$\mathbb{E}[X_t | X_u, u \in [0,s]] = \mu(t-s) + X_s \leq X_s,$$

so $\{X_t\}$ is a supermartingale.

The following remark illustrates an important point, namely that martingales are not necessarily Markov processes. Intuitively, the Markov property says that a process' distribution, given the present and past, depends only on the present. On the other hand, the martingale property says that, given the past and present, the **expected value** of the process is simply the present value.

Example 5.5. *Let $\{E_n\}_{n\in\mathbb{N}_0}$ be iid random variables where*

$$\mathbb{P}(E_n = 1) = \mathbb{P}(E_n = -1) = \frac{1}{2}.$$

Let X_0 be a random variable independent of $\{E_n\}$ with distribution

$$\mathbb{P}(X_0 = 0) = \mathbb{P}(X_0 = 1) = \frac{1}{2},$$

and define the process $\{X_n\}_{n\in\mathbb{N}_0}$ by, for $n \in \mathbb{N}_0$,

$$X_{n+1} = X_n + E_{n+1} X_0.$$

Note that, for $n \in \mathbb{N}_0$,

$$\begin{aligned}\mathbb{E}[X_{n+1} | X_0, \ldots, X_n] &= \mathbb{E}[X_n + E_{n+1}X_0 | X_0, \ldots, X_n] \\ &= X_n + X_0 \mathbb{E}[E_{n+1} | X_0, \ldots, X_n] \\ &= X_n + X_0 \mathbb{E}[E_{n+1}] \\ &= X_n.\end{aligned}$$

Additionally, for each $n \in \mathbb{N}_0$,

$$\mathbb{E}(|X_{n+1}|) \leq 2(n+2),$$

so we have shown that $\{X_n\}$ is a martingale. However,

$$X_2 = X_1 + E_2 X_0$$

so

$$\mathbb{P}[X_2 = 1 | X_0 = 1, X_1 = 0] = \mathbb{P}[X_1 + E_2 X_0 = 1 | X_0 = 1, X_1 = 0] = \frac{1}{2},$$

while

$$\mathbb{P}[X_2 = 1 | X_0 = 0, X_1 = 0] = \mathbb{P}[X_1 + E_2 X_0 = 1 | X_0 = 0, X_1 = 0] = 0.$$

This tells us that $\{X_n\}$ does not have the Markov property.

The follow proposition summarizes an important property of sub/super/martingales.

Proposition 5.6. *Let $\{X_n\}$, $\{Y_n\}$, and $\{Z_n\}$ be a martingale, submartingale, and supermartingale respectively. Then, for all $n \in \mathbb{N}_0$,*

$$\mathbb{E}(X_n) = \mathbb{E}(X_0), \quad \mathbb{E}(Y_n) \geq \mathbb{E}(Y_0), \quad \mathbb{E}(Z_n) \leq \mathbb{E}(Z_0).$$

Similarly, if $\{X_t\}$, $\{Y_t\}$, and $\{Z_t\}$ are continuous-time martingale, submartingale, and supermartingales, respectively, then for all $t \geq 0$,

$$\mathbb{E}(X_t) = \mathbb{E}(X_0), \quad \mathbb{E}(Y_t) \geq \mathbb{E}(Y_0), \quad \mathbb{E}(Z_t) \leq \mathbb{E}(Z_0).$$

Proof. We prove the result only for $\{X_n\}$; the other proofs are similar. Using the law of total expectation, note that

$$\mathbb{E}(X_n) = \mathbb{E}[\mathbb{E}[X_n | X_0, \dots, X_{n-1}]] = \mathbb{E}(X_{n-1}).$$

Repeating this process inductively, we see that $\mathbb{E}(X_n) = \mathbb{E}(X_0)$. □

We briefly note that we can define martingales more generally.

Definition 5.7. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ and $\{Y_n\}_{n\in\mathbb{N}_0}$ be stochastic processes. If the following hold:*

(1) For all $n \in \mathbb{N}_0$, $\mathbb{E}(|X_n|) < \infty$.

(2) For each $n \in \mathbb{N}_0$, $\mathbb{E}[X_{n+1}|Y_0, \dots, Y_n] = X_n$.

Then we say that $\{X_n\}$ is a ***martingale with respect the filtration generated by*** *$\{Y_n\}$. The definitions for submartingales and supermartingales, and for continuous-time processes are all analogous.*

The following example illustrates this concept in the continuous-time setting.

Example 5.8. *Let $\{W_t\}$ be an SBM, and define $\{X_t\}$ by*

$$X_t \doteq W_t^2 - t, \quad t \geq 0.$$

Then $\{X_t\}$ is a martingale with respect to the filtration generated by $\{W_t\}$. To see this, note that for each $t \geq 0$,

$$\mathbb{E}(|X_t|) = \mathbb{E}(|W_t^2 - t|) \leq \mathbb{E}(W_t^2) + t = 2t < \infty.$$

Additionally, for $s \leq t$,

$$\begin{aligned}
\mathbb{E}[X_t|W_s] &= \mathbb{E}[W_t^2 - t|W_s] \\
&= \mathbb{E}[(W_t - W_s + W_s)^2 - t|W_s] \\
&= \mathbb{E}[(W_t - W_s)^2|W_s] + 2\mathbb{E}[(W_t - W_s)W_s|W_s] + \mathbb{E}[W_s^2|W_s] - t \\
&= \mathbb{E}((W_t - W_s)^2) + 2W_s\mathbb{E}[W_t - W_s|W_s] + W_s^2 - t \\
&= \mathbb{E}(W_{t-s}^2) + 0 + W_s^2 - t \\
&= (t-s) + W_s^2 - t \\
&= W_s^2 - s \\
&= X_s.
\end{aligned}$$

Thus, $\{X_t\}$ is a martingale with respect to $\{W_t\}$.

The next example shows how to construct a martingale from a Poisson process.

Example 5.9. *Let $\{N_t\}$ be a Poisson process with rate λ. Then, the compensated Poisson process*

$$\tilde{N}_t \doteq N_t - t,$$

is a martingale with respect to $\{N_t\}$. Note that for each $t \geq 0$,

$$\mathbb{E}(|\tilde{N}_t|) \leq \mathbb{E}(N_t) + t = \lambda t + t < \infty,$$

and, for $s \leq t$,

$$\mathbb{E}[\tilde{N}_t|N_s] = \mathbb{E}[N_t - t|N_s] = \mathbb{E}[N_t - N_s + N_s - t|N_s] = \mathbb{E}[N_t - N_s] + N_s - t = N_s - s = \tilde{N}_s,$$

so $\{\tilde{N}_t\}$ is a martingale with respect to $\{N_t\}$.

In the next section we see some of the reasons martingales, submartingales, and supermartingales are so useful.

5.1. **Optional Stopping.** Recall from Proposition 5.6 that if $\{X_t\}$ is a martingale, then, for any fixed time $t \geq 0$, $\mathbb{E}(X_t) = \mathbb{E}(X_0)$. This means that for any deterministic time, the expected value of a martingale is the same as at time 0. Of course, this is not true for *random* times; consider an SBM $\{W_t\}$, and let, for some $a \neq 0$,

$$T_a \doteq \inf\{t \geq 0 : W_t = a\}.$$

Then, $\mathbb{P}(W_{T_a} = a) = 1$, so $\mathbb{E}(W_{T_a}) = a$. However, $\mathbb{E}(W_t) = 0$ for all $t \geq 0$.

We now introduce the definition of a *stopping time* more generally.

Definition 5.10. *Let $\{X_n\}_{n\in\mathbb{N}_0}$ be a discrete-time stochastic process. Consider a $\mathbb{N}_0$-valued random variable τ and define events $\{A_n\}_{n\in\mathbb{N}_0}$ by $A_n \doteq \{\tau \leq n\}$.*

We say that τ is a ***stopping time*** *for $\{X_n\}$ if, for each $n \in \mathbb{N}_0$, we can determine from observing $\{X_0, \dots, X_n\}$ whether the event A_n occurs.*

Similarly, let $\{Y_t\}_{t\geq 0}$ be a continuous-time stochastic process. Consider a non-negative random variable T and, define events $\{B_t\}_{t\geq 0}$ by $B_t \doteq \{T \leq t\}$.

We say that T is a ***stopping time*** *for $\{Y_t\}$ if, for each $t \geq 0$, we can determine from observing $\{Y_u,\ u \in [0, t]\}$ whether the event B_t occurs.*

Note that T_a defined above is a stopping time for $\{W_t\}$, since if we observe the trajectory of process over the time interval $[0, t]$, we can determine whether there is some $u \in [0, t]$ such that $W_u = a$; this would mean that $T_a \leq u \leq t$.

The following theorem is known as the Optional Stopping Theorem. We state the result for continuous-time martingales, but the analogous result holds for discrete-time martingales as well. The result requires some technical conditions on the process $\{X_t\}$ that are satisfied by all examples of interest that we have considered. Consequently, we omit these technical conditions - see e.g., this book for complete details. More general Optional Stopping Theorems are discussed there as well.

Theorem 5.11. *Let $\{Y_t\}$ be a stochastic process, and let T be a stopping time for some stochastic process $\{X_t\}$. Suppose that at least one of the following conditions holds:*

1. *There is some $M > 0$ such that $\mathbb{P}(T \leq M) = 1$.*
2. *$\mathbb{P}(T < \infty) = 1$ and there is some $M > 0$ such that*

$$\sup_{t\geq 0} \mathbb{E}\left(|Y_t| \mathbf{1}_{\{T>t\}}\right) \leq M.$$

Then, if $\{Y_t\}$ is a martingale with respect to $\{X_t\}$, $\mathbb{E}(Y_T) = \mathbb{E}(Y_0)$. If $\{Y_t\}$ is a submartingale with respect to $\{X_t\}$, then $\mathbb{E}(Y_T) \geq \mathbb{E}(Y_0)$, and if $\{Y_t\}$ is a supermartingale with respect to $\{X_t\}$, then $\mathbb{E}(Y_T) \leq \mathbb{E}(Y_0)$.

Note that condition (2) in Theorem 5.11 says that the stopping time is almost-surely finite, and that, on the event that the stopping time T has yet passed, the process has bounded expected value.

The following example illustrates how we can apply Theorem 5.11.

Example 5.12. *Let $\{W_t\}$ be an SBM. For $a, b > 0$, let $T_{a,-b} \doteq \inf\{t \geq 0 : W_t = a \text{ or } W_t = -b\}$. What is the probability that $\{W_t\}$ hits level a before it hits level $-b$?*

We begin by noting that this is equivalent to calculating $\mathbb{P}(W_{T_{a,-b}} = a)$. We will apply the optional stopping theorem; we show that condition (2) of that result holds.

First note that if $T_{a,-b} > t$, then $|W_t| < \max\{a, b\}$, which means that

$$\mathbb{E}(|W_t|\mathbf{1}_{\{T_{a,-b}>t\}}) \leq \mathbb{E}(\max\{a,b\}\mathbf{1}_{\{T_{a,-b}>t\}}) \leq \max\{a,b\}.$$

Additionally, from Lemma 4.22, we know that

$$\mathbb{P}(T_{a,-b} < \infty) \geq \mathbb{P}(T_a < \infty) = \lim_{t\to\infty} \mathbb{P}(T_a \leq t) = 1.$$

Thus, conditional (2) holds. Note also that if we let $p \doteq \mathbb{P}(W_{T_{a,-b}} = a)$, then $\mathbb{P}(W_{T_{a,-b}} = -b) = 1 - p$. Finally, $\{W_t\}$ is a martingale, so our last observation and the optional stopping theorem say that

$$pa + (1-p)(-b) = \mathbb{E}(W_{T_{a,-b}}) = \mathbb{E}(W_0) = 0,$$

so

$$p = \frac{b}{a+b}.$$

6. Stochastic Calculus

In this section we discuss some of the fundamental ideas in stochastic calculus. We will, for example, learn how to interpret equations of the form

$$dX_t = f(X_t)dt + g(X_t)dW_t, \tag{35}$$

where $\{W_t\}$ is an SBM and f, g are functions. As we will see, the equation in (35) is shorthand for an equation of the form

$$X_t = \int_0^t f(X_s)ds + \int_0^t g(X_s)dW_s.$$

From calculus class, we know that as long as the function $s \mapsto f(X_s)$ is sufficiently well-behaved (e.g., is continuous), we should be able to make sense of the term

$$\int_0^t f(X_s)ds,$$

as a (Riemann) integral. However, it is not yet clear how the term

$$\int_0^t g(X_s)dW_s,$$

should be defined, or what it represents. Thus, before we formally introduce such integrals, we begin with a motivating example to build our intuition.

6.1. A Motivating Example.

Suppose that we are interested in modeling the growth of a population over time; the size of the population at time t is denoted by $x(t)$. Assume also that the rate of change is proportional to the population size, namely that the larger the population is, the faster it grows. This leads to an ordinary differential equation (ODE) of the form

$$\begin{aligned} \frac{d}{dt}x(t) &= \alpha x(t) \\ x(0) &= x_0, \end{aligned} \tag{36}$$

where the parameter α determines the rate of growth, and x_0 is the initial size of the population. The equation in (36) can be formally rewritten as

$$\begin{aligned} dx(t) &= \alpha x(t)dt \\ x(0) &= x_0. \end{aligned} \tag{37}$$

A solution to (37) is then a function $x(t)$ satisfying

$$\begin{aligned} x(t) &= \int_0^t dx(s) = \int_0^t \alpha x(s)ds \\ x(0) &= x_0. \end{aligned} \tag{38}$$

We can verify that the unique solution to (38) is given by

$$x(t) = x_0 e^{\alpha t}, \quad t \geq 0,$$

meaning that (36) is an exponential growth model.

Let suppose instead that we are interested in modeling a population whose size has some additional randomness; then we might have an equation of the form

$$X_t = \int_0^t \alpha X_s ds + W_t,$$

where $\{W_t\}$ is a Brownian motion. Rewriting this in its differential form, we obtain

$$\begin{aligned} \frac{d}{dt}X_t &= \alpha X_t + \frac{d}{dt}W_t \\ X_0 &= x_0. \end{aligned} \tag{39}$$

or, equivalently,

$$\begin{aligned} dX_t &= \alpha X_t dt + dW_t \\ X_0 &= x_0. \end{aligned} \tag{40}$$

Of course, from Proposition 4.17, we know that Brownian motion sample paths are nowhere differentiable, and therefore that the quantity $\frac{d}{dt}W_t$ is not well-defined. Nonetheless, we would like to develop a reasonable interpretation for the equation in (40). Formally, integrating (40), we obtain

$$X_t = X_0 + \int_0^t \alpha X_s ds + \int_0^t dW_s.$$

This suggests that, at the very least, we should have

$$W_t = \int_0^t dW_s.$$

However, we might be interested in processes that have more general dynamics. For example, if we want to allow the noise term to depend somehow on the current size of the population, then we might obtain an equation of the form,

$$dX_t = \alpha X_t dt + g(X_t)dW_t,$$

where g is some function. After formally integrating both sides we obtain

$$X_t = \int_0^t \alpha X_s ds + \int_0^t g(X_s)dW_s,$$

so we will need to understand integrals of the form

$$\int_0^t g(X_s)dW_s,$$

more generally.

6.2. **Constructing the Itô Integral.** In this section we provide an introduction to stochastic integrals of the form discussed in Section 6.1. In particular, we will begin by studying integrals of the form

$$I_T \doteq \int_0^T X_s dW_s,$$

where $\{W_t\}$ is an SBM and $\{X_t\}$ is a stochastic process satisfying certain (very general) conditions. The full construction of such integrals is outside the scope of this class, but we provide an outline and discuss some important properties below.

First, we recall the notion of L^2 convergence of random variables.

Definition 6.1. *Let $\{X_n\}, X$ be random variables such that $\mathbb{E}(X^2) < \infty$, and, for each $n \in \mathbb{N}$, $\mathbb{E}(X_n^2) < \infty$. We say that X_n converges to X in L^2 if, as $n \to \infty$,*

$$\mathbb{E}[(X_n - X)^2] \to 0.$$

We sometimes say that as $n \to \infty$, $X_n \overset{L^2}{\to} X$.

6.2.1. *Integrating Simple Processes.* We begin by defining stochastic integrals for particularly simple processes. A **simple process** is a stochastic process $\{\phi_t\}$ of the form

$$\phi_t \doteq \sum_{j=0}^{n} E_j \mathbf{1}_{[t_j, t_{j+1})}(t),$$

where $0 < t_1 < \cdots < t_n = t$, and $\{E_j\}$ is a sequence of random variables such that the outcome of E_j can be determined by observing $\{W_u,\ u \in [0, t_j]\}$. Note that a simple process ϕ_t takes on only a finite number of (random) values in the time-interval $[0, t]$; it is piecewise constant on the intervals $[t_j, t_{j+1})$.

For a simple process we define

$$I_t \doteq \int_0^t \phi_s dW_s \doteq \sum_{j=0}^{n} \phi(t_j)(W_{t_{j+1}} - W_{t_j}) = \sum_{j=0}^{n} E_j (W_{t_{j+1}} - W_{t_j}). \tag{41}$$

Note that this resembles a Riemann sum, where the length of the partition is random, and where the left endpoint is taken.

An important observation in the construction of more general Itô integrals is Itô's isometry, which says that for a simple process $\{\phi_t\}$, for $0 \leq s \leq t$,

$$\mathbb{E}\left[\left(\int_s^t \phi_u dW_u\right)^2\right] = \mathbb{E}\left[\int_s^t \phi_u^2 du\right].$$

6.2.2. *Integrating More General Processes.* Now, we are interested in extending the definition from (41) to processes that are not simple. The general idea is to consider a stochastic process $\{X_t\}$ satisfying the following very general conditions:

(1) The outcome of X_t can be determined by observing $\{W_u,\ u \in [0, t]\}$. This is true, for example, if $X_t = f(W_t)$, where f is some function.

(2) For each $0 < s < t$,

$$\mathbb{E}\left[\int_s^t X_u^2 du\right] < \infty.$$

For a process $\{X_t\}$ satisfying (1) and (2), we can find a sequence $\{\phi_t^n\}$ of simple processes such that

$$\mathbb{E}\left[\int_s^t (X_u - \phi^n(u))^2 du\right] \to 0. \tag{42}$$

Then, using techniques from real analysis, we define the integral

$$I_s^t \doteq \int_s^t X_u dW_u,$$

as being the unique random variable satisfying, as $n \to \infty$,

$$\int_s^t \phi_u^n dW_u \xrightarrow{L^2} I_s^t.$$

Recall from Definition 6.1 that this means that, as $n \to \infty$,

$$\mathbb{E}\left[\left(\int_s^t \phi_u^n dW_u - I_s^t\right)^2\right] \to 0. \tag{43}$$

The key takeaways from the construction are that:

(1) We can easily define the integral of a simple process with respect to Brownian motion.
(2) We can approximate more general processes with a sequence of simple processes. Then, the integrals of these simple processes converge to the integral of the general process.

6.3. **Properties of the Itô Integral.** In this section we summarize without proof some of the most important properties of the Itô integral.

Theorem 6.2. *Let $\{W_t\}$ be an SBM and let $\{X_t\}$ and $\{Y_t\}$ be processes for which the stochastic integrals*

$$I_t^X \doteq \int_0^t X_s dW_s, \quad I_t^Y \doteq \int_0^t Y_s dW_s, \quad 0 \le t \le T,$$

are defined. Then, the following hold:

(1) For $\alpha, \beta \in \mathbb{R}$,

$$\int_0^t (\alpha X_s + \beta Y_s) dW_s = \alpha \int_0^t X_s dW_s + \beta \int_0^t Y_s dW_s.$$

(2) For $0 \le s \le t$,

$$\int_0^t X_u dW_u = \int_0^s X_u dW_u + \int_s^t X_u dW_u.$$

(3) For each $t \ge 0$,

$$\mathbb{E}\left(\int_0^t X_s dW_s\right) = 0.$$

(4) The outcome of $\int_s^t X_u dW_u$ can be determined by observing $\{W_u,\ u \in [0,t]\}$.

(5) The process $\{I_t^X\}$ is a martingale with respect to the filtration generated by $\{W_t\}$; namely,

$$\mathbb{E}[|I_t^X|] < \infty,$$

and, for $0 \le s \le t$,

$$\mathbb{E}[I_t^X | W_u,\ u \in [0,s]] = I_s^X.$$

(6) The process $\{I_t^X\}$ is continuous at each t with probability 1.

6.4. **Explicitly Calculating an Itô Integral.** In this section we use the construction of the Itô integral in Section 6.2 to explicitly calculate the integral

$$\int_0^t W_s dW_s.$$

To begin, we note that

$$\int_0^t s ds = \frac{t^2}{2},$$

so we might guess that

$$\int_0^t W_s dW_s \overset{?}{=} \frac{W_t^2}{2}.$$

As we see in the following proposition, however, this is not correct.

Proposition 6.3. *Let* $\{W_t\}$ *be an SBM. Then*

$$\int_0^t W_s dW_s = \frac{1}{2}\left(W_t^2 - t\right).$$

Proof. Consider partitions $0 = t_0^n < \cdots < t_n^n = t$ of $[0,t]$, where

$$\lim_{n\to\infty} \max_{0\le j\le n-1} (t_{j+1}^n - t_j^n) = 0.$$

Then for each $n \in \mathbb{N}$, consider the simple process $\{\phi_t^n\}$ defined as

$$\phi_s^n \doteq \sum_{j=0}^{n} W_{t_j^n} \mathbf{1}_{[t_j^n, t_{j+1}^n)}(s).$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[\int_0^t (\phi_s^n - W_s)^2 ds\right] &= \mathbb{E}\left[\sum_{j=0}^{n-1} \int_{t_j^n}^{t_{j+1}^n} (\phi_s^n - W_s)^2 ds\right] \\
&= \mathbb{E}\left[\sum_{j=0}^{n-1} \int_{t_j^n}^{t_{j+1}^n} \left(W_{t_j^n} - W_s\right)^2 ds\right] \\
&\overset{1}{=} \mathbb{E}\left[\sum_{j=0}^{n-1} \int_{t_j^n}^{t_{j+1}^n} \left(W_{s-t_j^n}\right)^2 ds\right] \\
&\overset{2}{=} \sum_{j=0}^{n-1} \int_{t_j^n}^{t_{j+1}^n} \mathbb{E}\left[(W_{s-t_j^n})^2\right] ds \\
&\overset{3}{=} \sum_{j=0}^{n-1} \int_{t_j^n}^{t_{j+1}^n} (s - t_j^n) ds \\
&= \frac{1}{2} \sum_{j=0}^{n-1} (t_{j+1}^n - t_j^n)^2 \\
&\le \frac{1}{2} \max_{0\le j\le n-1} (t_{j+1}^n - t_j^n) \sum_{j=0}^{n-1} (t_{j+1}^n - t_j^n) \\
&\overset{4}{=} \frac{t}{2} \max_{0\le j\le n-1} (t_{j+1}^n - t_j^n) \\
&\to 0.
\end{aligned} \tag{44}$$

where $\overset{1}{=}$ uses the fact that $\{W_t\}$ has stationary increments, $\overset{2}{=}$ uses the Fubini-Tonelli theorem to switch the expectation and the integral, $\overset{3}{=}$ uses the fact that $\mathbb{E}[(W_{s-t_j^n})^2] = s - t_j^n$, and $\overset{4}{=}$ uses the fact that

$$\sum_{j=0}^{n-1} (t_{j+1}^n - t_j^n) = t.$$

We have shown that $\{\phi_t^n\}$ converges to $\{W_t\}$ in the sense of (42). From the construction in Section 6.2, this means that

$$\int_0^t \phi_s^n dW_s \xrightarrow{L^2} \int_0^t W_s dW_s. \tag{45}$$

We now evaluate the integral on the left side of the previous display; since $\{\phi_s^n\}$ is a simple process, we have, from (41), that

$$\int_0^t \phi_s^n dW_s \doteq \sum_{j=0}^{n-1} W_{t_j^n}(W_{t_{j+1}^n} - W_{t_j^n}). \tag{46}$$

Furthermore,

$$W_{t_{j+1}^n}^2 - W_{t_j^n}^2 = (W_{t_{j+1}^n} - W_{t_j^n})^2 + 2W_{t_j^n}(W_{t_{j+1}^n} - W_{t_j^n}),$$

which says that

$$W_{t_j^n}(W_{t_{j+1}^n} - W_{t_j^n}) = \frac{1}{2}\left(W_{t_{j+1}^n}^2 - W_{t_j^n}^2\right) - \frac{1}{2}\left(W_{t_{j+1}^n} - W_{t_j^n}\right)^2. \tag{47}$$

Combining (46) and (47), we see that, since $t_n^n = t$,

$$\begin{aligned}\int_0^t \phi_s^n dW_s &= \sum_{j=0}^{n-1} W_{t_j^n}(W_{t_{j+1}^n} - W_{t_j^n}) \\ &= \frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t_{j+1}^n}^2 - W_{t_j^n}^2\right) - \frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t_{j+1}^n} - W_{t_j^n}\right)^2 \\ &= \frac{1}{2}W_{t_n^n}^2 - \frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t_{j+1}^n} - W_{t_j^n}\right)^2 \\ &= \frac{1}{2}W_t^2 - \frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t_{j+1}^n} - W_{t_j^n}\right)^2.\end{aligned} \tag{48}$$

In the final line of (48), the only quantity that depends on n is $\frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t_{j+1}^n} - W_{t_j^n}\right)^2$. Thus, in order to evaluate the limit (in L^2) of $\int_0^t \phi_s^n dW_s$, it suffices to evaluate the limit (in L^2) of this sequence. In particular we will show that as, $n \to \infty$,

$$\frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t_{j+1}^n} - W_{t_j^n}\right)^2 \xrightarrow{L^2} \frac{t}{2}. \tag{49}$$

Suppose that we have shown the convergence in (49). That convergence, along with (48), tells us that,

$$\int_0^t \phi_s^n dW_s \xrightarrow{L^2} \frac{1}{2}W_t^2 - \frac{1}{2}t.$$

However, the sequence $\left\{\int_0^t \phi_s^n dW_s\right\}_{n=1}^{\infty}$ can only have one limit, so it follows from the previous display and (45) that

$$\int_0^t W_s dW_s = \frac{1}{2}W_t^2 - \frac{1}{2}t.$$

Thus, to complete the proof, it suffices to show that the convergence in (49) holds. Towards that end, observe that

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t^n_{j+1}}-W_{t^n_j}\right)^2-\frac{t}{2}\right)^2\right] &= \frac{1}{4}\mathbb{E}\left[\left(\sum_{j=0}^{n-1}\left(W_{t^n_{j+1}}-W_{t^n_j}\right)^2-\sum_{j=0}^{n-1}(t^n_{j+1}-t^n_j)\right)^2\right]\\
&= \frac{1}{4}\mathbb{E}\left[\left(\sum_{j=0}^{n-1}\left[\left(W_{t^n_{j+1}}-W_{t^n_j}\right)^2-(t^n_{j+1}-t^n_j)\right]\right)^2\right]\\
&= \frac{1}{4}\sum_{j=0}^{n-1}\mathbb{E}\left[\left(W_{t^n_{j+1}}-W_{t^n_j}\right)^4-2\left(W_{t^n_{j+1}}-W_{t^n_j}\right)^2(t^n_{j+1}-t^n_j)+(t^n_{j+1}-t^n_j)^2\right]\\
&\overset{1}{=} \frac{1}{4}\sum_{j=0}^{n-1}\left[\mathbb{E}\left[\left(W_{t^n_{j+1}-t^n_j}\right)^4\right]-2(t^n_{j+1}-t^n_j)\mathbb{E}\left[\left(W_{t^n_{j+1}-t^n_j}\right)^2\right]+(t^n_{j+1}-t^n_j)^2\right]\\
&\overset{2}{=} \frac{1}{4}\sum_{j=0}^{n-1}\left[3(t^n_{j+1}-t^n_j)^2-2(t^n_{j+1}-t^n_j)(t^n_{j+1}-t^n_j)+(t^n_{j+1}-t^n_j)^2\right]\\
&= \frac{1}{2}\sum_{j=0}^{n-1}(t^n_{j+1}-t^n_j)^2
\end{aligned}
\tag{50}
$$

where in $\overset{1}{=}$ used the fact that $\{W_t\}$ has stationary increments, and $\overset{2}{=}$ used the fact that if $X \sim \mathcal{N}(0,t)$, then $\mathbb{E}[(X)^2] = t$ and $\mathbb{E}[(X)^4] = 3t^2$. Combining the calculation in (44) with (50), we see that

$$
\mathbb{E}\left[\left(\frac{1}{2}\sum_{j=0}^{n-1}\left(W_{t^n_{j+1}}-W_{t^n_j}\right)^2-\frac{t}{2}\right)^2\right] \to 0.
$$

This shows that the convergence in (49) holds, so the proof is complete. □

Below we define what it means for a stochastic process to solve an SDE.

Definition 6.4. *Suppose that $\{X_t\}$ is a stochastic process satisfying*

$$
X_t = x_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s, \tag{51}
$$

*where $b,\sigma : \mathbb{R} \to \mathbb{R}$ are functions and $\{W_t\}$ is an SBM. Then we say that $\{X_t\}$ is **a solution to the stochastic differential equation (SDE)***

$$
\begin{aligned}
dX_t &= b(X_t)dt + \sigma(X_t)dW_t\\
X_0 &= x_0.
\end{aligned}
\tag{52}
$$

*We refer to b as the drift coefficient and σ as the diffusion coefficient. We refer to the term $b(X_t)dt$ as the **drift term** and the term $\sigma(X_t)dW_t$ as the **diffusion term**. In practice, we use (51) and (52) interchangeably; we refer to a process $\{X_t\}$ satisfying such equations as an **Itô diffusion**.*

6.5. **Itô's Formula.** Here we discuss Itô's formula, which generalized the chain rule from calculus to stochastic integrals. To motivate it's utility, note that in calculus, when we encounter an integral of the form

$$
\int_0^t s ds,
$$

we generally do not rely on the definition of the integral (in terms of Riemann sums) to evaluate it. Recall that the chain rule says that if f is a differentiable function, then

$$\frac{d}{ds}f(s) = f'(s),$$

which can be rewritten as

$$df(s) = f'(s)ds. \tag{53}$$

For example, if

$$f(s) = \frac{s^2}{2},$$

then $\frac{d}{ds}f(s) = f'(s) = s$, meaning that

$$df(s) = f'(s)ds = sds,$$

so

$$\int_0^t sds = \int_0^t df(s) = f(t) - f(0) = f(t) = \frac{t^2}{2}.$$

However, as we saw in Proposition 6.3, if $\{W_t\}$ is an SBM, then

$$\int_0^t W_s dW_s = \frac{W_t^2}{2} - \frac{t}{2},$$

so if we naively tried to apply the chain rule, we would not properly evaluate the integral. The following formula, known as Itô's formula Itô's lemma, establishes an analogue of the chain rule for Itô integrals.

Theorem 6.5. *(Itô's Formula I)*

Let $\{W_t\}$ be an SBM, and suppose that, for some functions $b : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}$, $\{X_t\}$ is a stochastic process satisfying

$$X_t = \int_0^t b(X_t)dt + \int_0^t \sigma(X_t)dW_t, \quad t \geq 0,$$

or, equivalently,

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad t \geq 0.$$

Additionally, define the quadratic variation process of $\{X_t\}$ by

$$\langle X \rangle_t \doteq \int_0^t \sigma^2(X_s)ds, \quad t \geq 0,$$

so that

$$d\langle X \rangle_t = \sigma^2(X_t)dt.$$

Then, if $f : \mathbb{R} \to \mathbb{R}$ is a twice-differentiable function with continuous second derivative, we have

$$\begin{aligned} f(X_t) &= f(X_0) + \int_0^t f'(X_s)dX_s + \frac{1}{2}\int_0^t f''(X_s)d\langle X \rangle_s \\ &= f(X_0) + \int_0^t f'(X_s)b(X_s)ds + \int_0^t f'(X_s)\sigma(X_s)dW_s + \frac{1}{2}\int_0^t f''(X_s)\sigma^2(X_s)ds. \end{aligned} \tag{54}$$

Observe that the first line of (54) *can be rewritten as*

$$\begin{aligned} df(X_t) &= f'(X_t)dX_t + \frac{1}{2}f''(X_t)d\langle X \rangle_t \\ &= f'(X_t)dX_t + \frac{1}{2}f''(X_t)\sigma^2(X_t)dt \\ &= f'(X_t)b(X_t)dt + f'(X_t)\sigma(X_t)dW_t + \frac{1}{2}f''(X_t)\sigma^2(X_t)dt. \end{aligned} \tag{55}$$

The following corollary gives Itô's formula for functions of Brownian motion.

Corollary 6.6. *Let $\{W_t\}$ be an SBM, and let f be a twice-differentiable function with continuous second derivative. Then,*

$$f(W_t) = f(0) + \int_0^t f'(W_s)dW_s + \frac{1}{2}\int_0^t f''(W_s)ds.$$

In the differential form, this says

$$df(W_t) = f'(W_t)dW_t + \frac{1}{2}f''(W_t)dt.$$

Proof. The result follows on noting that $\{W_t\}$ satisfies the SDE

$$dW_t = b(W_t)dt + \sigma(W_t)dW_t = 0dt + 1dW_t = dW_t,$$

where $b(x) \doteq 0$ and $\sigma(x) \doteq 1$. Just apply Theorem 6.5 with these choices of b and σ.

□

6.6. **Examples of How to Apply Itô's Formula.** Before we look at examples of how to apply Itô's formula, to see why it can be interpreted as an analogue of the chain rule for Itô integrals, compare (53) and (55). We begin with a simple example of how to apply Itô's formula.

Example 6.7. *Consider the process $\{X_t\}$ defined by*

$$X_t \doteq \int_0^t W_s dW_s, \quad t \geq 0,$$

where $\{W_t\}$ is an SBM. We begin by rewriting the equation in its differential form:

$$dX_t = W_t dW_t. \tag{56}$$

Since X_t does not appear on right hand side of the equation above, we might expect that we should apply Itô's formula from Corollary 6.6 to a function of W_t; namely, we might consider

$$df(W_t) = f'(W_t)dW_t + \frac{1}{2}f''(W_t)dt,$$

which says that

$$f'(W_t)dW_t = df(W_t) - \frac{1}{2}f''(W_t)dt. \tag{57}$$

Comparing (56) and (57), we see that we should choose a function f such that $f'(W_t) = W_t$; namely, we choose $f(w) = \frac{w^2}{2}$, so that

$$f'(w) = w, \quad f''(w) = 1.$$

Plugging this choice of f into (57), we see that

$$W_t dW_t = d\left(\frac{W_t^2}{2}\right) - \frac{1}{2}dt,$$

which says that

$$X_t = \int_0^t W_s dW_s = \int_0^t d\left(\frac{W_t^2}{2}\right) - \int_0^t \frac{1}{2}dt = \frac{W_t^2}{2} - \frac{t}{2}.$$

Note that Theorem 6.2 tells us that $\{X_t\}$ is a martingale with respect to $\{W_t\}$ (note also that we proved this directly in Example 5.8).

Below we discuss two more approaches that lead us to applying Itô's formula to $f(W_t) = \frac{W_t^2}{2}$. Note that we were interested in calculating

$$\int_0^t W_s dW_s.$$

Note that we can rearrange the result in Corollary 6.6 to obtain

$$\int_0^t f'(W_s)dW_s = f(W_t) - f(0) - \frac{1}{2}\int_0^t f''(W_s)ds.$$

So if we choose the function f so that $f'(W_s) = W_s$, then we will have

$$\int_0^t f'(W_s)dW_s = \int_0^t W_s dW_s = f(W_t) - f(0) - \frac{1}{2}\int_0^t f''(W_s)ds.$$

Thus, since $f(W_t) = \frac{W_t^2}{2}$ has $f'(W_s)dW_s = W_s dW_s$, this is the 'correct' function to use. Of course, we could also have used $f(x) = \frac{x^2}{2}$ to avoid the factor of 2 above.

Another way to think of our approach is as follows; the deterministic analogue of this problem is

$$dx(t) = t dt,$$

which yields $x(t) = \frac{t^2}{2}$, so we might expect that $X_t \overset{?}{\approx} \frac{W_t^2}{2}$, which suggests that we apply Itô's formula to $f(W_t) = \frac{W_t^2}{2}$.

Now we introduce a stochastic process, known as geometric Brownian motion, that has many important applications in finance (i.e., in relation to the Black-Scholes Model).

Definition 6.8. *A stochastic process $\{X_t\}$ is known as a geometric Brownian motion (GBM) if is satisfies the SDE*

$$dX_t = \mu X_t dt + \sigma X_t dW_t \tag{58}$$

$$X_0 = x_0. \tag{59}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are constants and $\{W_t\}$ is an SBM. Equivalently, $\{X_t\}$ is a GBM if

$$X_t = x_0 + \int_0^t \mu X_t dt + \int_0^t \sigma X_t dW_t.$$

The following proposition shows us how to explicitly construct a GBM using an SBM. We use

Proposition 6.9. *Let $\{W_t\}$ be an SBM and define $\{X_t\}$ by*

$$X_t \doteq x_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right), \quad t \geq 0.$$

Then $\{X_t\}$ is a GBM with initial value of x_0.

Proof. We begin by noting that that the diffusion term in

$$dX_t = \mu X_t dt + \sigma X_t dW_t,$$

is of the form

$$\sigma X_t dW_t.$$

Since this terms depends on X_t, we suspect that we will need to apply Itô's formula to some function $f(X_t)$; let us determine the function. Itô's formula says that

$$df(X_t) = f'(X_t)\mu X_t dt + f'(X_t)\sigma X_t dW_t + \frac{1}{2} f''(X_t)\sigma^2 X_t^2 dt. \tag{60}$$

To simplify the computation, we choose a function f such that

$$f'(X_t)\sigma X_t = 1;$$

this is helpful as it will allow us to evaluate the integral of the term

$$f'(X_t)\sigma X_t dW_t.$$

Such a function satisfies

$$f'(x) = \frac{1}{\sigma} x^{-1},$$

which means that

$$f(x) = \frac{1}{\sigma}\log(x).$$

Plugging this choice of f into (60), we obtain

$$\begin{aligned} d\left(\frac{1}{\sigma}\log(X_t)\right) &= \frac{1}{\sigma}\frac{1}{X_t}\mu X_t dt + dW_t + \frac{1}{2}\frac{1}{\sigma}\left(-\frac{1}{X_t^2}\right)\sigma^2 X_t^2 dt \\ &= \frac{1}{\sigma}\mu dt + dW_t + \frac{1}{2}\sigma dt. \end{aligned} \tag{61}$$

Integrating (61) and using the initial condition that $\frac{1}{\sigma}\log(X_0) = \frac{1}{\sigma}\log(x_0)$, we see that

$$\frac{1}{\sigma}\log(X_t) = \frac{1}{\sigma}\log(x_0) + \frac{1}{\sigma}\mu t + W_t - \frac{1}{2}\sigma t,$$

or, equivalently,

$$X_t = x_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right).$$

This completes the proof. Another way to determine the correct function f is outlined below.

The SDE describing $\{X_t\}$ can be rewritten as

$$\frac{dX_t}{X_t} = \mu dt + \sigma dW_t.$$

Since the left side of the previous display is of the form $\frac{dx}{x}$, and we know that

$$d\log(x) = \frac{dx}{x},$$

this suggests that we should apply Theorem 6.5 to the function $f(x) \doteq \log(x)$.

□

Now we present another example.

Example 6.10. *Solve the SDE*

$$\begin{aligned} dX_t &= dt + 2\sqrt{X_t}dW_t \\ X_0 &= x_0. \end{aligned} \tag{62}$$

We begin by noting that the diffusion term

$$2\sqrt{X_t}dW_t,$$

depends on X_t, so we begin by applying Itô's formula to a function of X_t:

$$df(X_t) = f'(X_t)dt + f'(X_t)2\sqrt{X_t}dW_t + \frac{1}{2}f''(X_t)4X_t dt. \tag{63}$$

As in the proof of Proposition 6.9*, it is convenient to consider a function f such that*

$$f'(X_t)2\sqrt{X_t} = 1,$$

or, equivalently,

$$f'(X_t) = \frac{1}{2\sqrt{X_t}}.$$

Note that such a function is given by $f(X_t) = \sqrt{X_t}$, so if we rewrite (63) *with this choice of f, we obtain, using the fact that*

$$f''(X_t) = -\frac{1}{4X_t^{\frac{3}{2}}},$$

$$\begin{aligned} d\sqrt{X_t} &= \frac{1}{2\sqrt{X_t}}dt + dW_t + \frac{1}{2}\left(-\frac{1}{4X_t^{\frac{3}{2}}}\right)4X_t dt \\ &= \frac{1}{2\sqrt{X_t}}dt + dW_t - \frac{1}{2\sqrt{X_t}}dt \\ &= dW_t. \end{aligned}$$

Since $X_0 = x_0$, it follows that

$$\sqrt{X_t} = \sqrt{x_0} + W_t,$$

which says that

$$X_t = (\sqrt{x_0} + W_t)^2,$$

solves (62).

We can also verify directly using Itô's formula that our solution is correct. Plugging

$$\sqrt{X_t} = \sqrt{x_0} + W_t,$$

into (62)*, we see that the SDE can be rewritten as*

$$\begin{aligned} dX_t &= dt + 2(\sqrt{x_0} + W_t)dW_t \\ X_0 &= x_0. \end{aligned} \tag{64}$$

Since the diffusion term of (64) *is*

$$2(\sqrt{x_0} + W_t)dW_t,$$

which does not explicitly contain X_t, we expect that we should apply Itô's formula to some function $g(W_t)$; we need only to choose the function g. Itô's formula says that

$$g(W_t) = g'(W_t)dW_t + \frac{1}{2}g''(W_t)dt, \tag{65}$$

so we should find a function g such that

$$g'(W_t) = 2(\sqrt{x_0} + W_t).$$

Such a function is given by

$$g(W_t) = 2\left(\sqrt{x_0}W_t + \frac{W_t^2}{2}\right).$$

Note that $g''(W_t) = 2$, so (65) *says that*

$$d\left(2\left(\sqrt{x_0}W_t + \frac{W_t^2}{2}\right)\right) = 2(\sqrt{x_0} + W_t)dW_t + \frac{1}{2}2dt.$$

Integrating both sides of the previous display and using the initial condition that $g(W_0) = 0$, we see that

$$\left(2\sqrt{x_0}W_t + W_t^2\right) = \int_0^t 2(\sqrt{x_0} + W_s)dW_s + t.$$

Thus, adding x_0 to both sides of the previous display,

$$(\sqrt{x_0} + W_t)^2 = x_0 + \left(2\sqrt{x_0}W_t + W_t^2\right) = x_0 + \int_0^t 2(\sqrt{x_0} + W_s)dW_s + t. \tag{66}$$

Observe that (66) *says that, with $X_t \doteq (\sqrt{x_0} + W_t)^2$,*

$$\begin{aligned} dX_t &= dt + 2\sqrt{X_t}dW_t \\ X_0 &= x_0, \end{aligned}$$

as claimed.

6.7. **Itô's Formula for Functions Depending on Time.** Now we state a more general version of Itô's formula. The key difference is that this version allows the function f to depend on time.

Theorem 6.11. *(Itô's Formula II)*

Let $\{W_t\}$ be an SBM, and suppose that, for some functions $b, \sigma : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$, $\{X_t\}$ is a stochastic process satisfying

$$X_t = \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s, \quad t \geq 0,$$

or, equivalently,

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \quad t \geq 0.$$

If $f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ is a function with continuous partial derivatives

$$\dot{f}(t,x) \doteq \frac{d}{dt}f(t,x), \quad f'(t,x) \doteq \frac{d}{dx}f(t,x), \quad f''(t,x) \doteq \frac{d^2}{dx^2}f(t,x),$$

then

$$f(t, X_t) = f(0, X_0) + \int_0^t \dot{f}(s, X_s)ds + \int_0^t f'(s, X_s)dX_s + \frac{1}{2}\int_0^t f''(s, X_s)d\langle X\rangle_s,$$

where

$$d\langle X\rangle_t = \sigma^2(t, X_t)dt.$$

Note that this can be rewritten as

$$\begin{aligned} df(t, X_t) &= \dot{f}(t, X_t)dt + f'(t, X_t)dX_t + \frac{1}{2}f''(t, X_t)d\langle X\rangle_t \\ &= \dot{f}(t, X_t)dt + f'(t, X_t)b(t, X_t)dt + f'(t, X_t)\sigma(t, X_t)dW_t + \frac{1}{2}f''(t, X_t)\sigma^2(t, X_t)dt. \end{aligned}$$

In the next example we apply Theorem 6.11 to explicitly solve a stochastic differential equation (SDE).

Example 6.12. *Let $\{W_t\}$ be an SBM. Find a process $\{X_t\}$ such that*

$$\begin{aligned} dX_t &= -\frac{1}{1+t}X_t dt + \frac{1}{1+t}dW_t \\ X_0 &= 0. \end{aligned} \tag{67}$$

We will find a function $f : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ such that if $X_t \doteq f(t, W_t)$, then (67) *holds.*

Since the diffusion term on the right side of (67) *is*

$$\frac{1}{1+t}dW_t,$$

we suspect that $X_t = f(t, W_t)$; we just need to determine the function. Itô's formula says that

$$\begin{aligned} dX_t &= df(t, W_t) \\ &= \dot{f}(t, W_t)dt + f'(t, W_t)dW_t + \frac{1}{2}f''(t, W_t)dt \\ &= \left(\dot{f}(t, W_t) + \frac{1}{2}f''(t, W_t)\right)dt + f'(t, W_t)dW_t. \end{aligned} \tag{68}$$

Comparing (67) *and* (68)*, and recalling that we believe that $X_t = f(t, W_t)$, it suffices to find a function with the following properties:*

$$\dot{f}(t, W_t) = -\frac{1}{1+t}f(t, W_t), \quad f'(t, W_t) = \frac{1}{1+t}, \quad f''(t, W_t) = 0. \tag{69}$$

Observe that the function

$$f(t, x) \doteq \frac{x}{1+t}, \tag{70}$$

satisfies the properties in (69)*. Thus, if we let*

$$X_t \doteq f(t, W_t) = \frac{W_t}{1+t},$$

then (68), (69)*, and* (70) *tells us that*

$$dX_t = -\frac{1}{1+t}X_t dt + \frac{1}{1+t}dW_t.$$

Additionally,

$$X_0 = f(0, W_0) = \frac{W_0}{1+0} = \frac{0}{1} = 0,$$

so we have shown that our choice of $\{X_t\}$ does solve the SDE in (67)*.*

These appendices contain lecture notes from PSTAT 160A. It may be helpful to review these notes before and during PSTAT 160B.

Appendix A. Random Variables

In this section we review some properties of random variables from an introductory probability class.

A.1. **Probability Spaces.**

To formally define a random variable, we begin by introducing the notion of a **probability space**. Probability spaces consist of three parts:

(1) A sample space Ω that contains all possible outcomes ω of some experiment or random trial.
(2) An event space $\mathscr{F}$ that contains all of the events that we can assign probabilities to. Namely, $\mathscr{F}$ is the collection of subsets $A \subseteq \Omega$ that we can assign probabilities to.
(3) A probability measure $\mathbb{P}$ that assigns a probability $\mathbb{P}(A)$ to each event $A \in \mathscr{F}$.

We refer to the triplet $(\Omega, \mathscr{F}, \mathbb{P})$ as a probability space.

Definition A.1. *A function $\mathbb{P} : \mathscr{F} \to [0,1]$ is a* ***probability measure*** *if it satisfies:*

(1) For each $A \in \mathscr{F}$, $\mathbb{P}(A) \in [0,1]$ (the probability of each event is some number between 0 and 1).
(2) $\mathbb{P}(\Omega) = 1$ (the probability that some *outcome takes place is 1).*
(3) If $\{A_n\}_{n=1}^{\infty} = \{A_1, A_2, \dots\}$ is a collection of pairwise disjoint events (namely, $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Some important properties of probability measures that follow from Definition A.1 are:

(1) For each $A \in \mathscr{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(2) For each $A, B \in \mathscr{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.[3]
(3) If $A, B \in \mathscr{F}$ are such that $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
(4) $\mathbb{P}(\emptyset) = 0$.

It is a helpful exercise to remind yourself why each of these properties follows from the definition above.

Example A.2. *You have a coin that is equally likely to land on heads and tails.*

If you flip the coin once, then the probability space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$ is given by $\Omega_1 = \{H, T\}, \mathscr{F}_1 = \{\{H,T\}, \{H\}, \{T\}, \emptyset\}$, and the probability measure $\mathbb{P}$ is defined by $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2$.

If you flip the coin twice, then the probability space $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$ is given by

$$\Omega_2 = \{(H,T), (T,H), (H,H), (T,T)\},$$

and $\mathscr{F}_2$ contains all subsets of Ω_2. For instance, $\mathscr{F}_2$ contains the event that you "flip heads then tails or flip tails then tails". We denote this event by $\{(H,T), (T,T)\}$. Here the probability measure on Ω_2 is given by

$$\mathbb{P}(\{(H,T)\}) = P(\{(T,H)\}) = P(\{(H,H)\}) = P(\{(T,T)\}) = 1/4.$$

Recall that two events A and B are **independent** if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

[3] Recall that $A \cup B \doteq$ "A or B" and $A \cap B \doteq$ "A and B". For a quick review of set notation, see `https://www.purplemath.com/modules/setnotn.htm`.

We often are not interested in the outcomes themselves, but rather in some numerical value derived from the outcomes. For example, when we flip two coins we might be interested in the *number* of coins that land on heads, rather than the order in which the coins land. If we let X denote the number of coins that land on heads, then X assigns a numerical value to each outcome in the sample space. Namely, it is a *function* defined on the sample space. By that we mean that for each $\omega \in \Omega_2$, there is a corresponding value $X(\omega)$ that tells us how many coins landed on heads. Here we can precisely write down how the function $X : \Omega_2 \to \{0,1,2\}$ is defined:

- $X(\{(H,H)\}) = 2$
- $X(\{(H,T))\}) = 1$
- $X(\{(T,H)\}) = 1$
- $X(\{(T,T)\}) = 0$

The function X defined above is an example of a **random variable**. We think of random variables as functions that input an outcome (i.e., some scenario) and output a number. We refer to the possible numbers that the random variable X can take on as the **state space** of X.

Some important notation regarding random variables:

(1) Since we think of X as a function, it has a *domain* and *range*. Its domain is the sample space Ω, and its range is the state space, which we typically denote by $\mathscr{S}_X$. Note that

$$\mathscr{S}_X = X(\Omega) \doteq \{x \in \mathbb{R} : \text{ there is some } \omega \in \Omega \text{ such that } X(\omega) = x\}$$

(2) For a subset $A \subseteq \mathscr{S}_X$, we write

$$\{X \in A\} \doteq \{\omega \in \Omega : X(\omega) \in A\}.$$

Note that $\{X \in A\} \subseteq \Omega$.

(3) For a subset $A \subseteq \mathscr{S}_X$, we write

$$\mathbb{P}(X \in A) \doteq \mathbb{P}(\{X \in A\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

(4) Random variables are generally denote by capital letters such as X, Y, Z, etc.

Example A.3. *You roll a fair four-sided die twice. Let X denote the maximum of the two rolls. Write down the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and the state space $\mathscr{S}_X$, and specify X explicitly as a function from the sample space Ω to the state space $\mathscr{S}_X$.*

You may not have seen random variables formulated this way before. Typically it will suffice to think of a random variable as exactly that, namely a random numerical quantity. However, it will occasionally be helpful for us to think of random variables as functions defined on some sample space.

A.2. **Discrete Random Variables.** A **discrete random variable** is a random variable that can take on only an enumerable number of values. Some common state spaces of discrete random variables are:

(1) $\mathscr{S}_X = \{1,2,3,4,5,6\}$
(2) $\mathscr{S}_X = \{0,1\}$
(3) $\mathscr{S}_X = \mathbb{N} \doteq \{1,2,3,\dots\}$
(4) $\mathscr{S}_X = \mathbb{N}_0 \doteq \{0,1,2,\dots\}$

Discrete random variables are described in terms of a **probability mass function** (p.m.f.), which is a function $p_X : \mathscr{S}_X \to [0,1]$ defined as

$$p_X(x) = \mathbb{P}(X = x).$$

Note that if p_X is the p.m.f. of a random variable X, then

$$\sum_{x \in \mathscr{S}_X} p_X(x) = \sum_{x \in \mathscr{S}_X} \mathbb{P}(X = x) = 1,$$

and for each $A \subseteq \mathscr{S}_X$,

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} p_X(x).$$

Some common discrete random variables are Bernoulli, binomial, Poisson, geometric, etc.

Example A.4. *Consider the random variable X from Example* A.3. *The p.m.f. of X is given by*

$$p_X(x) = \begin{cases} \frac{1}{16} & x = 1 \\ \frac{3}{16} & x = 2 \\ \frac{5}{16} & x = 3 \\ \frac{7}{16} & x = 4. \end{cases}$$

To calculate the probability that X is between 1 and 3, we evaluate

$$\mathbb{P}(1 \le X \le 3) = \sum_{x=1}^{3} p_X(x) = \frac{1}{16} + \frac{3}{16} + \frac{5}{16} = \frac{9}{16}.$$

A.3. **Continuous Random Variables.** A **continuos random variable** is a random variable that can take on an uncountable number of values. Some common state spaces of continuous random variables are:

(1) $\mathscr{S}_X = \mathbb{R} = (-\infty, \infty)$
(2) $\mathscr{S}_X = \mathbb{R}_+ = [0, \infty)$
(3) $\mathscr{S}_X = [0, 1]$

Continuous random variables are described in terms of a **probability density function** (p.d.f.), which is a function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} f_X(x)dx = 1.$$

The idea is that if f_X is relatively large in some region, then it is more likely that X will take on a value in that region. In particular, for a continuous random variable X, to compute probabilities we look at

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx.$$

Note that for each $c \in \mathbb{R}$,

$$\mathbb{P}(X = c) = \int_c^c f_X(x)dx = 0,$$

so $f_X(c) \neq \mathbb{P}(X = c)$. Generally we will compute probabilities of the form

$$\mathbb{P}(X \in (a,b)) = \mathbb{P}(a < X < b) = \mathbb{P}(X \in [a,b]) = \mathbb{P}(a \le X \le b) = \int_a^b f_X(x)dx.$$

Some common example of continuous random variables are uniform, exponential, normal, etc.

Example A.5. *For $\lambda > 0$, we say that $X \sim Exp(\lambda)$ if*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

Note that $\mathscr{S}_X = \mathbb{R}_+$. To calculate $\mathbb{P}(X \in (a,b))$, where $0 \le a \le b$, we use to change of variables $u \doteq \lambda x$, $du \doteq \lambda dx$, to see

$$\begin{aligned}\mathbb{P}(X \in (a,b)) &= \int_a^b f_X(x)dx \\ &= \int_a^b \lambda e^{-\lambda x} dx \\ &= \int_{\lambda a}^{\lambda b} e^{-u} du \\ &= -e^{-u}\Big|_{\lambda a}^{\lambda b} \\ &= e^{-\lambda a} - e^{-\lambda b}\end{aligned}$$

A.4. **Cumulative Distribution Functions.** The **probability distribution** of a random variable describes how likely the different values of a random variable are. Two discrete random variables have the same p.m.f. if and only if they have the same probability distribution. However, the same is not true for continuous random variables. For example, consider $X \sim \text{Exp}(1)$, and consider the random variable Y whose p.d.f. is given by

$$f_Y(y) = \begin{cases} e^{-y} & y \in \mathbb{R}_+ \setminus \{1\} \\ 10 & y = 1 \\ 0 & y < 0. \end{cases}$$

Then $\mathscr{S}_X = \mathscr{S}_Y = \mathbb{R}_+$, and for any $0 \le a < b$ we have that

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a \le Y \le b) = e^{-a} - e^{-b}.$$

This should tell us that X and Y have the same probability distribution, even though their p.m.f.'s are different. This motivates the definition of another function, defined in terms of a random variable, that completely describes its probability distribution.

Definition A.6. *For a random variable X, the function $F_X : \mathbb{R} \to [0,1]$ defined as*

$$F_x(x) \doteq \mathbb{P}(X \le x)$$

*is the **cumulative distribution function** (c.d.f.) of X.*

Below we list several important properties of c.d.f.'s.

Remark A.7. *(1) If X is a discrete random variable with p.m.f. p_X, then*

$$F_X(x) \doteq \sum_{y \le x} p_X(y). \tag{71}$$

(2) If X is a continuous random variable with p.d.f. f_X, then

$$F_X(x) \doteq \int_{-\infty}^{x} f_X(y)dy. \tag{72}$$

(3) For any random variable X, the function F_X is non-decreasing and right-continuous, namely the following properties hold:

(a) If $x \le y$, then $F_X(x) \le F_Y(y)$.

   (b) *For each* $x \in \mathbb{R}$, $\lim_{z \to x^+} F_X(z) = F_X(x)$.

(4) *The following limits hold:*

   (a) *As* $x \to -\infty$, $F_X(x) \to 0$.

   (b) *As* $x \to \infty$, $F_X(x) \to 1$.

(5) *The c.d.f. of a random variable characterizes the distribution/probability law of a random variable uniquely. Namely, if* X *and* Y *are random variables with c.d.f.'s* F_X *and* F_Y*, respectively, then we say that* X *and* Y *have the same probability distribution if* $F_X(z) = F_Y(z)$ *for all* $z \in \mathbb{R}$. *In that case, we write* $X \overset{d}{=} Y$ *or* $X \overset{\mathscr{L}}{=} Y$.

Example A.8. *Let* $X \sim Exp(\lambda)$ *be as in Example* A.5. *Note that the state space of* X *is* $\mathscr{S}_X = \mathbb{R}_+$*, so if* $x < 0$*, then*

$$F_X(x) = \mathbb{P}(X \le x) = 0,$$

and if $x \ge 0$*, then*

$$F_X(x) \doteq \mathbb{P}(X \le x) = \mathbb{P}(0 \le X \le x) = \int_0^x \lambda e^{-\lambda y} dy = \int_0^{\lambda x} e^{-u} du = -e^{-u}\Big|_0^{\lambda x} = 1 - e^{-\lambda x}.$$

Above we used the change of variables of $u \doteq \lambda y$. *Therefore, the c.d.f. of* X *is the function* $F_X : \mathbb{R} \to [0,1]$ *given by*

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

The next example is slightly more involved. It will be important when you study continuous time Markov chains. Recall that we say that two random variables X and Y are **independent** if for all events A and B we have

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Example A.9. *Helen and Teddy are waiting to be fed after I wake up in the morning. However, they don't always get fed at the same time. Their waiting times* X *and* Y *(in hours) are modeled as independent,* $Exp(\lambda)$ *random variables.*

Let M *denote the amount of time that it takes for the* first *cat to be fed, namely* $M \doteq \min\{X, Y\}$. *What is the distribution/probability law of* M.

In order to solve this problem, note that we can write

$$\{M > x\} = \{\min\{X, Y\} > x\} = \{X > x, Y > x\}.$$

Since X *and* Y *are independent, we know from Example* A.8 *that*

$$\mathbb{P}(X > x, Y > x) = \mathbb{P}(X > x)\mathbb{P}(Y > x) = (1 - F_X(x))(1 - F_Y(x)) = e^{-\lambda x} e^{-\lambda x} = e^{-2\lambda x}$$

Therefore,

$$\begin{aligned} F_M(x) &= \mathbb{P}(M \le x) \\ &= \mathbb{P}(\min\{X, Y\} \le x) \\ &= 1 - \mathbb{P}(\min\{X, Y\} > x) \\ &= 1 - \mathbb{P}(X > x, Y > x) \\ &= 1 - e^{-2\lambda x}. \end{aligned}$$

Since the c.d.f. uniquely characterizes a random variable's probability distribution, it follows that $M \sim Exp(2\lambda)$.

If instead of two cats, I had n cats, and their feeding times in the morning, denoted by $X_1, X_2, \dots, X_n$, were all independent $Exp(\lambda)$ random variables, then what would the distribution of the first feeding time $M \doteq \min\{X_1, X_2, \dots, X_n\}$ be?

The main takeaway of this section is that the c.d.f. of a random variable is a function that fully describes the random variable's probability distribution.

A.5. **Expected Value.** The expected value of a random variable describes its average value or mean.

Definition A.10. *A discrete random variable X with p.m.f. p_X and state space $\mathscr{S}_X$ is* ***integrable*** *if*

$$\sum_{x \in \mathscr{S}_X} |x| p_X(x) < \infty.$$

If X is integrable, then its ***expected value*** *is defined as*

$$\mathbb{E}(X) \doteq \sum_{x \in \mathscr{S}_X} x p_X(x).$$

A continuous random variable X with p.d.f. f_X is ***integrable*** *if*

$$\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty.$$

If X is integrable, then its ***expected value*** *is defined as*

$$\mathbb{E}(X) \doteq \int_{-\infty}^{\infty} x f_X(x) dx.$$

If X is ***non-negative*** *(i.e., if $\mathbb{P}(X \ge 0) = 1$), then we define $\mathbb{E}(X)$ as above regardless of whether X is integrable; in this case it is possible to have $\mathbb{E}(X) = \infty$.*

In Definition A.10 above, the expected value of a random variable can be interpreted as a weighted average of the possible values that the random variable can take on. The more 'likely' an outcome is, the more heavily it is weighted.

Example A.11. *We say that $X \sim Cauchy(0, 1)$ if its p.d.f. is given by*

$$f_X(x) \doteq \frac{1}{\pi(1 + x^2)}, \qquad x \in \mathbb{R}.$$

The p.d.f. of the Cauchy distribution resembles that of the normal distribution, but it goes to 0 less quickly as $x \to \pm\infty$. A consequence of this is that X is not integrable, so its expected value is not *defined. In order to see this, note that*

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx \doteq \lim_{a,b\to\infty} \int_{-a}^{b} |x| f_X(x)\,dx,$$

and observe that for $a, b > 0$,

$$\begin{aligned}\int_{-a}^{b} |x| f_X(x)\,dx &= \int_{-a}^{b} \frac{|x|}{\pi(1+x^2)}\,dx \\ &= \int_{-a}^{0} \frac{|x|}{\pi(1+x^2)}\,dx + \int_{0}^{b} \frac{|x|}{\pi(1+x^2)}\,dx \\ &= \int_{0}^{a} \frac{x}{\pi(1+x^2)}\,dx + \int_{0}^{b} \frac{x}{\pi(1+x^2)}\,dx\end{aligned}$$

Using the change of variable $u \doteq x^2, du \doteq 2x dx$, we can evaluate

$$\begin{aligned}\int_{0}^{a} \frac{x}{\pi(1+x^2)}\,dx &= \frac{1}{2\pi}\int_{0}^{a^2} \frac{1}{1+u}\,du \\ &= \frac{1}{2\pi}\log(1+u)\Big|_{0}^{a^2} \\ &= \frac{1}{2\pi}\left(\log(1+a^2) - \log(1+0)\right) \\ &= \frac{1}{2\pi}\log(1+a^2),\end{aligned}$$

so it follows that

$$\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx \doteq \lim_{a\to\infty}\int_{0}^{a} \frac{x}{\pi(1+x^2)}\,dx = \lim_{a\to\infty} \frac{1}{2\pi}\log(1+a^2) = \infty.$$

Similarly,

$$\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}\,dx = \infty,$$

so X is not integrable.

Nearly every random variable we study in this course will be integrable, and so will have a well-defined expected value. Accordingly, in the rest of this section it will be assumed that all random variables are integrable. The following result will be useful throughout this course.

Theorem A.12. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function.*

If X is a discrete random variable with p.m.f. p_X and state space $\mathscr{S}_X$, then

$$\mathbb{E}(g(X)) = \sum_{x\in\mathscr{S}_X} g(x)p_X(x).$$

If X is a continuous random variable with p.d.f. f_X, then

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

Theorem A.12 tells us exactly how to calculate the expected value of a function of a random variable. For example, it allows us to compute the moments of X.

Definition A.13. *The n-**th moment** ($n \in \mathbb{N}$) of a random variable X is the quantity* $\mathbb{E}(X^n)$.

*The **variance** of a random variable is defined as*

$$Var(X) \doteq \mathbb{E}((X-\mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

*The **covariance** of two random variables X and Y is defined as*

$$Cov(X,Y) \doteq \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Note that in order to calculate $Cov(X,Y)$, we need to know the joint distribution *of X and Y. We will come to this shortly.*

Remark A.14. *In order to calculate the n-th moment of a random variable, we take the function g in Theorem* A.12 *to be $g(x) \doteq x^n$.*

In order to calculate the variance of X, we can take $g(x) \doteq (x-\mathbb{E}(X))^2$ in Theorem A.12.

Some important properties of expectation are below. Here X and Y are random variables and $a, b \in \mathbb{R}$ are (non-random) constants.

(1) $\mathbb{E}(aX+bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ *(linearity)*
(2) *If* $\mathbb{P}(X \le Y) = 1$, *then* $\mathbb{E}(X) \le \mathbb{E}(Y)$ *(monotonicity)*
(3) $\mathbb{E}(X+a) = \mathbb{E}(X) + a$
(4) $Var(aX) = a^2 Var(X)$
(5) $Var(X+a) = Var(X)$
(6) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.
(7) *If X and Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$.*

The following example shows how to calculate the expected value and variance of an exponential random variable.

Example A.15. *Let $X \sim Exp(\lambda)$. Then*

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx\end{aligned}$$

Integrating by parts with $u \doteq \lambda x$, $dv \doteq e^{-\lambda x}$, so that $du = \lambda dx$, $v = -\frac{1}{\lambda}e^{-\lambda x}$, we obtain

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x\cdot\lambda e^{-\lambda x}dx\\ &= \int_0^\infty u dv\\ &= uv\Big|_0^\infty - \int_0^\infty v du\\ &= -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}dx\\ &= 0 + \left(-\frac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty\right)\\ &= \frac{1}{\lambda}.\end{aligned}$$

The second moment of X can be calculated by integrating by parts with $u \doteq \lambda x^2$ and $dv \doteq e^{-\lambda x}$. Then we have $du = 2\lambda x dx$, $v \doteq -\frac{1}{\lambda}e^{-\lambda x}$, so that

$$\begin{aligned}\mathbb{E}(X^2) &= \int_0^\infty x^2 f_X(x)dx\\ &= \int_0^\infty \lambda x^2 e^{-\lambda x}dx\\ &= uv\Big|_0^\infty - \int_0^\infty v du\\ &= \frac{2}{\lambda^2}.\end{aligned}$$

It follows that

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Below we derive the expected value and variance of a geometric random variable.

Example A.16. *Let $X \sim Geometric(p)$ for some $p \in (0,1)$, so that the p.m.f. of X is given by*

$$p_X(x) = (1-p)^{x-1}p, \qquad x \in \mathbb{N}.$$

Calculate $\mathbb{E}(X)$ and $Var(X)$.

We begin by recalling that if $|r| < 1$, then

$$h(r) \doteq \sum_{x=1}^\infty r^x = \frac{1}{1-r}.$$

Therefore,

$$h'(r) = \sum_{x=1}^\infty x r^{x-1} = \frac{d}{dr}\left(\frac{1}{1-r}\right) = \frac{1}{(1-r)^2},$$

and

$$h''(r) = \sum_{x=1}^\infty x(x-1)r^{x-2} = \frac{2}{(1-r)^3}.$$

Note that

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=1}^{\infty} x p_X(x) \\ &= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= ph'(1-p) \\ &= p\frac{1}{(1-(1-p))^2} \\ &= \frac{1}{p}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{x=1}^{\infty} x^2 p_X(x) \\ &= p\sum_{x=1}^{\infty} x^2(1-p)^{x-1} \\ &= p\sum_{x=1}^{\infty} (x^2-x+x)(1-p)^{x-1} \\ &= p(1-p)\sum_{x=1}^{\infty} x(x-1)(1-p)^{x-2} + p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= p(1-p)h''(1-p) + ph'(1-p) \\ &= \frac{2p(1-p)}{p^3} + \frac{p}{p^2} \\ &= \frac{2-p}{p^2}.\end{aligned}$$

It follows that

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

A.6. **Joint Distribution of Random Variables.** The joint distribution of two random variables X and Y describes how the two random variables behave when they are viewed *together*. For an example of the subleties that can arise with joint distributions, let $X \sim \mathcal{N}(0,1)$ be a standard normal random variable, and let $Y \doteq -X$. Then $Y \sim \mathcal{N}(0,1)$ as well, so

$$\mathbb{P}(X \geq 0) = \mathbb{P}(Y \geq 0) = 1/2.$$

However,

$$\mathbb{P}(X \geq 0, Y \geq 0) = \mathbb{P}(X \geq 0, -X \geq 0) = 0,$$

so when we view X and Y together they behave very differently than they do on their own.

To understand the joint distribution of two random variables, we begin by introducing the notion of a bivariate random variable. A **bivariate random variable** on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is a random variable of the form $(X,Y) : \Omega \to \mathbb{R}^2$, meaning that for each $\omega \in \Omega$, there is some $(x,y) \in \mathbb{R}^2$ such that $(X(\omega), Y(\omega)) = (x,y) \in \mathbb{R}^2$.

If X and Y are jointly discrete, then there is some p.m.f. $p_{(X,Y)} : \mathbb{R}^2 \to [0,1]$ such that for each pair of events $A, B \subseteq \mathbb{R}$,

$$\mathbb{P}((X,Y) \in A \times B) = \sum_{x,y \in A \times B} p_{(X,Y)}(x,y) \doteq \sum_{x \in A} \sum_{y \in B} p_{(X,Y)}(x,y)$$

We refer to $p_{(X,Y)}$ as the **joint p.m.f.** of (X,Y).

Similarly, if (X,Y) are jointly continuous, then there is a **joint p.d.f.** $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ such that for each pair of events $A, B \subseteq \mathbb{R}$,

$$P((X,Y) \in A \times B) = \int_{A \times B} f_{(X,Y)}(x,y) d(x,y) = \int_A \int_B f_{(X,Y)}(x,y) dy dx = \int_B \int_A f_{(X,Y)}(x,y) dx dy.$$

Note that we can exchange the order of integration above due to Fubini's theorem. [4]

The joint p.m.f. and joint p.d.f. also give rise to a joint c.d.f. , which, as in the univariate case, fully characterizes the joint distribution of (X,Y). The **joint c.d.f.** of (X,Y) is the function $F_{(X,Y)} : \mathbb{R}^2 \to [0,1]$ defined as

$$F_{(X,Y)}(x,y) \doteq \mathbb{P}(X \le x, Y \le y).$$

Note that if X and Y are jointly discrete with joint p.m.f. $p_{(X,Y)}$, then their joint c.d.f. is given by

$$F_{(X,Y)}(x,y) \doteq \mathbb{P}(X \le x, Y \le y) = \sum_{i \le x} \sum_{j \le y} p_{(X,Y)}(i,j).$$

Similarly, if X and Y are jointly continuous with joint p.d.f. $f_{(X,Y)}$, then their joint c.d.f. is given by

$$\begin{aligned} F_{(X,Y)}(x,y) &\doteq \mathbb{P}(X \le x, Y \le y) \\ &= \int_{(-\infty,x] \times (-\infty,y]} f_{(X,Y)}(u,v) d(u,v) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{(X,Y)}(u,v) dv du \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{(X,Y)}(u,v) du dv, \end{aligned}$$

where we once more use Fubini's theorem to justify the equivalences.

Proposition A.17. *If we know the joint distribution of (X,Y), we can also recover their individual (i.e., marginal) distributions. In the discrete case we have*

$$p_X(x) = \sum_{y \in \mathscr{S}_Y} p_{(X,Y)}(x,y), \qquad p_Y(y) = \sum_{x \in \mathscr{S}_X} p_{(X,Y)}(x,y).$$

and in the continuous case we have

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

Additionally, recall that two discrete random variables X and Y are independent if and only if for all $x \in \mathscr{S}_X, y \in \mathscr{S}_Y$ we have

$$p_{(X,Y)}(x,y) = p_X(x) p_Y(y).$$

Similarly, two continuous random variables X and Y are independent if and only if for all $x, y \in \mathbb{R}$ we have

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

[4] See `https://en.wikipedia.org/wiki/Fubini's_theorem`

Recall from Definition A.13 that in order to calculate the covariance of two random variables X and Y, we need to calculate $\mathbb{E}(XY)$. The following formula allows us to compute this as well as other quantities such as $\mathbb{E}(X^Y)$, $\mathbb{E}(X^2Y^2)$, and so on.

Let $g:\mathbb{R}^2\to\mathbb{R}$ be a function. If (X,Y) is jointly discrete with joint p.m.f. $p_{(X,Y)}$, then

$$\mathbb{E}(g(X,Y)) = \sum_{(x,y)\in\mathscr{S}_X\times\mathscr{S}_Y} g(x,y)p_{(X,Y)}(x,y) = \sum_{x\in\mathscr{S}_X}\sum_{y\in\mathscr{S}_Y} g(x,y)p_{(X,Y)}(x,y).$$

Similarly, if (X,Y) is jointly continuous with joint p.d.f. $f_{(X,Y)}$, then

$$\mathbb{E}(g(X,Y)) = \int_{\mathbb{R}^2} g(x,y)f_{(X,Y)}(x,y)d(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{(X,Y)}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{(X,Y)}(x,y)dydx.$$

If we take $g(x,y)\doteq xy$, then in the discrete case this yields

$$\mathbb{E}(XY) = \sum_{x\in\mathscr{S}_X}\sum_{y\in\mathscr{S}_Y} xyp_{(X,Y)}(x,y).$$

In the continuous case we have

$$\mathbb{E}(XY) = \int_{\mathbb{R}^2} xyf_{(X,Y)}(x,y)d(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{(X,Y)}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{(X,Y)}dydx.$$

Using this, if we know the joint distribution of X and Y, then we can calculate their covariance:

$$\text{Cov}(X,Y) = \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) = \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y).$$

In the following example we use the joint p.d.f. of a pair of random variables to calculate their covariance.

Example A.18. *Let X and Y be continuous random variables with joint p.d.f.*

$$f_{(X,Y)}(x,y)\doteq 3x, \qquad 0\le y\le x\le 1.$$

Calculate the marginal densities of X and Y. Determine whether X and Y are independent and calculate Cov(X,Y).

The marginal density of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dy = \int_0^x 3xdy = 3x^2, \qquad 0\le x\le 1,$$

and the marginal density of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx = \int_y^1 3xdx = \frac{3x^2}{2}\Big|_y^1 = \frac{3(1-y^2)}{2}, \qquad 0\le y\le 1.$$

Since the joint density $f_{(X,Y)}$ is not the product of the marginal densities f_X and f_Y, it follows that X and Y are not independent. Therefore,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf_X(x)dx = \int_0^1 x\cdot 3x^2dx = \frac{3}{4},$$

and

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy = \int_0^1 y\frac{3(1-y^2)}{2}dy = \frac{3}{8}.$$

Finally,

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{(X,Y)}(x,y)dydx\\ &= \int_0^1\int_0^x xy\cdot 3xdydx\\ &= \int_0^1 3x^2\left(\int_0^x ydy\right)dx\\ &= \int_0^1 \frac{x^2}{2}\cdot 3x^2dx\\ &= \frac{3}{10}.\end{aligned}$$

Therefore,

$$Cov(XY) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{3}{10} - \frac{3}{4}\times\frac{3}{8} = \frac{3}{160}.$$

A.7. **Conditional Probability and Independence.** Conditional probability allows us to understand the how the outcome of one event (i.e., did the event happen or not) affects the outcome of another event. For instance, given two events A and B, it allows us to calculate the probability that A happens given that B has already happened (or will definitely happen, depending on the situation).

Definition A.19. *Given two events A and B, the conditional probability of A given B is defined as*

$$\mathbb{P}(A|B) \doteq \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)},$$

as long as $\mathbb{P}(B) > 0$.

Conditional probability has similar properties to regular probability, namely for events A, B, and C, the following hold whenever $\mathbb{P}(B) > 0$:

(1) $\mathbb{P}(A|B) = 1 - \mathbb{P}(A^c|B)$
(2) $\mathbb{P}(A\cup C|B) = \mathbb{P}(A|B) + \mathbb{P}(C|B) - \mathbb{P}(A\cap C|B)$
(3) If $A \subseteq C$, then $\mathbb{P}(A|B) \le \mathbb{P}(C|B)$
(4) $\mathbb{P}(\emptyset|B) = 0$

The following proposition illustrates how one of the many uses of conditional probability. Recall that a **partition** of the sample space Ω is a (finite or infinite) collection of pairwise disjoint (i.e., $B_i \cap B_j = \emptyset$ whenever $i \neq j$) sets $\{B_1, B_2, \ldots\}$ such that $\bigcup\limits_{n=1}^{\infty} B_n = \Omega$ and $\mathbb{P}(B_i) > 0$ for all $i \geq\in \mathbb{N}$.

Proposition A.20. *(Law of Total Probability) Let $\{B_1, B_2, \ldots\}$ be a partition of Ω. Then for each event $A \subseteq \Omega$,*

$$\mathbb{P}(A) = \sum_{i\geq 1}\mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

A simple consequence of this law is that for any events A, B satisfying $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

The following example illustrates how one can apply the Law of Total Probability to simplify some calculations.

Example A.21. *You have three jars of marbles. Jar 1 contains 75 red and 25 blue marbles, jar 2 contains 60 red and 40 blue marbles, and jar 3 contains 45 red and 55 blue marbles. You choose one fo the jars at random then randomly draw a marble from that jar. What is the probability that you draw a red marble?*

Let A denote the event that you draw a red marble, and for $i \in \{1,2,3\}$, let B_i denote the event that you choose jar i. Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \mathbb{P}(A|B_3)\mathbb{P}(B_3)\\ &= \frac{1}{3}\left(\frac{75}{100} + \frac{60}{100} + \frac{45}{100}\right)\\ &= \frac{3}{5}\end{aligned}$$

The following result is known as Bayes' formula. It follows immediately from the definition of conditional probability.

Proposition A.22. *(Bayes' Formula) For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, we have*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

The following example illustrates how one can apply Bayes' formula.

Example A.23. *You have a bag with 100 coins. Of these coins, 99 are real coins, namely they are fair and have heads on one side and tails on the other, but one of the coins has heads on both sides.*

You pick a coin at random, and do not check whether it is real or fake. You flip the coin $n \in \mathbb{N}$ times in a row, and it lands on heads all n times. What is the probability that you picked the fake *coin?*

Let A be the event that you picked a fake coin, and let B be the event that a coin lands on heads for all n flips. Then

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)\\ &= 1\cdot\frac{1}{100} + \left(\frac{1}{2}\right)^n \cdot \frac{99}{100}\\ &= \frac{1}{100}\left(1 + 99\left(\frac{1}{2}\right)^n\right).\end{aligned}$$

Consequently,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{100}\cdot 1}{\frac{1}{100}\left(1+99\left(\frac{1}{2}\right)^n\right)} = \frac{1}{1+99\left(\frac{1}{2}\right)^n}.$$

Since $\mathbb{P}(A|B)$ increases towards 1 as $n \to \infty$, it follows that the larger n is, the more likely it is that you drew the fake coin.

Finally, we recall the notion of independent random variables once more.

Definition A.24. *Two random variables X and Y are* ***independent*** *if and only if for all events A and B,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Similarly, random variables $X_1, \dots, X_n$ are ***independent*** *if and only if for all events $A_1, \dots, A_n$, we have*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) \doteq \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2)\cdots\mathbb{P}(X_n \in A_n).$$

The following proposition gives several different criteria that can be used to check for independence of random variables. It is stated only for a pair of random variables, but the analogous result holds for a collection of $n > 2$ random variables as well.

Proposition A.25. *Let X and Y be random variables with c.d.f. 's F_X and F_Y, respectively. Let $F_{(X,Y)}$ denote the joint c.d.f. of (X,Y). Then the following are equivalent:*

(1) X and Y are independent.

(2) For all $(x, y) \in \mathbb{R}^2$, $F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$.

(3) If X and Y are discrete, then for all $(x, y) \in \mathscr{S}_X \times \mathscr{S}_y$, $p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$. Similarly, if X and Y are continuous, then for all $(x, y) \in \mathbb{R}^2$, $f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$.

Additionally, if X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

This brings us to the end of our review of the material from PSTAT 120A. If you are feeling uncomfortable with any of this material, please let me know as soon as possible, as these formulas, definitions, and ideas will be used regularly throughout this course.

APPENDIX B. CONDITIONAL EXPECTATION

In Sections A.6 and A.7 we discussed jointly distributed random variables and conditional probability, respectively. In this section we introduce the notion of conditional expectation, which gives us further insight into the relationship between random variables. In particular, we will define the following quantities:

(1) $\mathbb{P}[X \in A | Y = y]$: the probability that X belongs to a set A given the outcome $Y = y$.
(2) $\mathbb{E}[X|Y = y]$: the conditional expectation of X given the outcome $Y = y$.
(3) $\mathbb{E}[X|Y]$: the conditional expectation of X given that we observe Y.
(4) $\mathbb{E}[X|A]$: the expected value of a random variable X given that some event A occurred.

To see some of the subtlety of defining such quantities, let $Y, Z \overset{\text{iid}}{\sim} \text{Exp}(1)$, and let $X = Y + Z$. Then clearly X and Y are not independent. Suppose that we want to calculate $\mathbb{P}(X \geq 2|Y = 1)$. Since Y is a continuous random variable, $\mathbb{P}(Y = 1) = 0$. Accordingly, if we let $A \doteq \{X \geq 2\}$, $B \doteq \{Y = 1\}$, then we have

$$\mathbb{P}(X \geq 2|Y = 1) = \mathbb{P}(A|B),$$

but $\mathbb{P}(B) = 0$, so we can't use the definition of conditional probability given in Definition A.19. However, intuitively we might expect that

$$\mathbb{P}(X \geq 2|Y = 1) = \mathbb{P}(Y + Z \geq 2|Y = 1) = \mathbb{P}(Z \geq 1) = e^{-1}.$$

Note that the issue above arose only because Y was continuous.

B.1. **Conditional Expectation Given a Particular Outcome.**

We begin with the somewhat simpler definition of the conditional p.m.f. of two discrete random variables.

Definition B.1. *Let X and Y be discrete random variables. Then for each $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$ (i.e., for each $x \in \mathscr{S}_X$), the* ***conditional p.m.f.*** *of Y given that $X = x$ is the function $p_{Y|X=x}(\cdot|x) : \mathbb{R} \to [0,1]$ defined as*

$$p_{Y|X=x}(y|x) \doteq \mathbb{P}[Y = y|X = x] = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)} = \frac{p_{(X,Y)}(x, y)}{p_X(x)},$$

where $p_{(X,Y)}$ denotes the joint p.m.f. of (X, Y) and p_X denotes the marginal p.m.f. of X. Additionally, we have

$$\mathbb{P}(Y \in A|X = x) = \sum_{y \in A} p_{Y|X=x}(y|x).$$

For notational convenience we often write $p_{Y|X}(y|x) \doteq p_{Y|X=x}(y|x)$.

The conditional p.d.f. of two continuous random variables is defined similarly.

Definition B.2. *Let X and Y be continuous random variables with joint p.d.f. $f_{(X,Y)}$, and marginal p.d.f. 's f_X and f_Y. For each $x \in \mathbb{R}$ such that $f_X(x) > 0$,[a] the* ***conditional p.d.f.*** *of Y given that $X = x$ is the function $f_{Y|X=x}(\cdot\ x) : \mathbb{R} \to \mathbb{R}_+$ defined by*

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

Then

$$\mathbb{P}(Y \in A|X = x) = \int_A f_{Y|X=x}(y|x)dy.$$

As in the discrete case, for notational convenience we often write $f_{Y|X}(y|x) \doteq f_{Y|X=x}(y|x)$.

[a] Note that $f_X(x) > 0$ does not imply that $\mathbb{P}(X = x) > 0$, since $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

Conditional expectation allows us to calculate the expected value of a random variable given that another random variable has taken on a particular value.

Definition B.3. *If X and Y are discrete random variables, then the conditional expectation of Y given that $X = x$ is defined as*

$$\mathbb{E}[Y|X = x] \doteq \sum_{y\in\mathcal{S}_Y} y p_{Y|X=x}(y|x).$$

Similarly, if X and Y are continuous random variables, then the conditional expectation of Y given that $X = x$ is defined as

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

In both cases, the conditional expectation is interpreted as calculating the expectation with respect to the conditional probability distribution.

The following observation will be useful for making sense of things like $\mathbb{E}[Y|X]$.

Observation B.4. *In the definition above, the quantity $\mathbb{E}[Y|X = x]$ depends on our choice of x and nothing else. Accordingly, we can view the conditional expectation of Y given that $X = x$ as a function h defined by*

$$h(x) \doteq \mathbb{E}[Y|X = x].$$

The next result is an analogue of Theorem A.12 in the context of conditional expectation.

Theorem B.5. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function. If X and Y are discrete random variables with conditional p.m.f. $p_{Y|X}$, then*

$$\mathbb{E}[g(Y)|X = x] = \sum_{y\in\mathcal{S}_Y} g(y)p_{Y|X}(y|x).$$

Similarly, if X and Y are continuous random variables with conditional p.d.f. $f_{Y|X}$, then

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

The follow example demonstrates conditional probability in the discrete setting.

Example B.6. *Fix $p \in (0,1)$ and consider random variables X and Y with joint p.m.f.*

$$p_{(X,Y)}(x,y) = \frac{y^x e^{-y}}{x!}(1-p)^{y-1}p, \qquad x \in \mathbb{N}_0, y \in \mathbb{N}.$$

Calculate the conditional p.m.f. of X given that $Y = y$

We begin by checking that $p_{(X,Y)}$ *is in fact a p.m.f. :*

$$\begin{aligned}\sum_{y=1}^{\infty}\sum_{x=0}^{\infty} p_{(X,Y)}(x,y) &= p\sum_{y=1}^{\infty}(1-p)^{y-1}e^{-y}\sum_{x=0}^{\infty}\frac{y^x}{x!}\\ &= p\sum_{y=1}^{\infty}(1-p)^{y-1}e^{-y}e^{y}\\ &= \frac{p}{1-(1-p)}\\ &= 1.\end{aligned}$$

The marginal p.m.f. of Y *is given by*

$$p_Y(y) = \sum_{x=0}^{\infty}\frac{y^x e^{-y}}{x!}(1-p)^{y-1}p = (1-p)^{y-1}pe^{-y}\sum_{x=0}^{\infty}\frac{y^x}{x!} = (1-p)^{y-1}p.$$

Therefore, $Y \sim$ *Geometric*(p). *The conditional p.m.f. of* X *given that* $Y = y$ *is given by*

$$\begin{aligned}p_{X|Y}(x|y) &= \frac{p_{(X,Y)}(x,y)}{p_Y(y)}\\ &= \frac{\frac{y^x e^{-y}}{x!}(1-p)^{y-1}p}{(1-p)^{y-1}p}\\ &= \frac{y^x e^{-y}}{x!}.\end{aligned}$$

Recall that we say that $U \sim$ *Poisson*(λ) *if the p.m.f. of* U *is given by*

$$p_U(u) = \frac{\lambda^u e^{-\lambda}}{u!}, \qquad u \in \mathbb{N}_0,$$

and that $\mathbb{E}(U) = \lambda$. *Observe that the conditional p.m.f. of* X *given that* $Y = y$ *is exactly the p.m.f. of a Poisson*(y) *distribution, so it follows that* $\mathbb{E}[X|Y = y] = y$. *In a situation like this, we often write* $X \sim$ *Poisson*(Y), $Y \sim$ *Geometric*(p) *or* $(X|Y = y) \sim$ *Poisson*(y).

Note that we are interpreting the distribution of X in Example B.6 as depending on a random parameter. In order to sample from the probability distribution of X one would first sample Y, then use the resulting value of Y to choose the distribution of X. The following code illustrates how one can use Python to generate samples from the distribution of X.

```
import numpy as np
import matplotlib.pyplot as plt

p = 0.35
n = 100000
Y = np.random.geometric(p=0.35, size=n)
X = np.random.poisson(lam = Y, size = n)

plt.hist(x, density=True, bins=100)
plt.ylabel(’Probability’)
plt.xlabel(’Data’);
```

We now consider an example of conditional probability in the continuous setting.

Example B.7. *Let (X,Y) be uniformly distributed on the closed unit disk, namely the set $\mathscr{D} \doteq \{(x,y) \in \mathbb{R}^2 : x^2+y^2 \le 1\}$. The joint p.d.f. of (X,Y) is then given by*

$$f_{(X,Y)}(x,y) = \begin{cases} \frac{1}{\pi}, & (x,y) \in \mathscr{D} \\ 0, & \textit{otherwise.} \end{cases}$$

Find the conditional density function $f_{X|Y}(\cdot|y)$ and the conditional expectation $\mathbb{E}[X|Y=y]$ for $y \in \mathscr{D}$.

We can calculate

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{1}{\pi}dx = \frac{2}{\pi}\sqrt{1-y^2},$$

so the marginal density of Y is given by

$$f_Y(y) = \begin{cases} \frac{2}{\pi}\sqrt{1-y^2}, & -1 \le y \le 1 \\ 0, & \textit{otherwise.} \end{cases}$$

For $(x,y) \in \mathscr{D}^1$ we have

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{1}{\pi}\frac{\pi}{2}\frac{1}{\sqrt{1-y^2}} = \frac{1}{2\sqrt{1-y^2}}.$$

Thus, for each $y \in [-1,1]$, the conditional density of X given that $Y=y$ is

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{2\sqrt{1-y^2}}, & -\sqrt{1-y^2} \le x \le \sqrt{1-y^2} \\ 0, & \textit{otherwise.} \end{cases}$$

Therefore, we write $(X|Y=y) \sim$ Unif$(-\sqrt{1-y^2}, \sqrt{1-y^2})$.[a] For each $y \in [-1,1]$ we can calculate

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx = \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} x\frac{1}{2\sqrt{1-y^2}}dx = 0.$$

[a] Recall that we say $U \sim \text{Uniform}(a,b)$ if the p.d.f. of U is given by $f_U(u) = \frac{1}{b-a},\ a \le u \le b$.

Below is another example of calculating conditional probability distributions.

Example B.8. *Let $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ) be independent. Let $Z \doteq X+Y$ and find the conditional expectation of Y given that $Z=z$ for some $z \in \mathbb{N}_0$.*

The joint p.m.f. of (Z,Y) is given by

$$p_{(Z,Y)}(z,y) = \mathbb{P}(Z=z, Y=y) = \mathbb{P}(X+Y=z, Y=y) = \mathbb{P}(X=z-y, Y=y) = \mathbb{P}(X=z-y)\mathbb{P}(Y=y),$$

which is given by

$$p_{(Z,Y)}(z,y) = \begin{cases} \frac{\lambda^{z-y}e^{-\lambda}}{(z-y)!}\frac{\mu^y e^{-\mu}}{y!}, & z \ge y \\ 0, & \textit{otherwise.} \end{cases}$$

The marginal p.m.f. of Z *is given by*

$$\begin{aligned} p_Z(z) &= \sum_{y=0}^{\infty} p_{(Z,Y)}(z,y) \\ &= \sum_{y=0}^{z} \frac{\lambda^{z-y}e^{-\lambda}}{(z-y)!}\frac{\mu^y e^{-\mu}}{y!} \\ &= e^{-(\lambda+\mu)} \sum_{y=0}^{z} \frac{z!}{(z-y)!y!}\frac{\lambda^{z-y}\mu^y}{z!} \\ &= \frac{e^{-(\lambda+\mu)}}{z!} \sum_{y=0}^{z} \binom{z}{y}\lambda^{z-y}\mu^y \\ &= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^z}{z!}, \end{aligned}$$

which shows that $Z \sim$ *Poisson*$(\lambda+\mu)$.[a] *The conditional distribution of* Y *given that* $Z = z$ *is given by*

$$p_{Y|Z}(y|z) = \frac{p_{(Z,Y)}(z,y)}{p_Z(z)} = \frac{\lambda^{z-y}\mu^y z!}{(\lambda+\mu)^z(z-y)!y!} = \binom{z}{y}\left(\frac{\mu}{\lambda+\mu}\right)^y\left(1-\frac{\mu}{\lambda+\mu}\right)^{z-y}$$

It follows that $(Y|Z=z) \sim$ *Binomial*$(z, p_{\mu,\lambda})$, *where* $p_{\mu,\lambda} \doteq \frac{\mu}{\lambda+\mu}$. *Consequently,* $\mathbb{E}[Y|Z=z] = zp_{\mu,\lambda}$.

A similar calculation shows that $(X|Z=z) \sim$ *Binomial*$(z, p_{\lambda,\mu})$, *where* $p_{\lambda,\mu} \doteq \frac{\lambda}{\lambda+\mu}$, *which ensures that* $\mathbb{E}[X|Z=z] = zp_{\lambda,\mu}$.

[a] The final identity follows from the Binomial Theorem, which says that $(a+b)^n = \sum_{i=0}^{n}\binom{n}{i}a^{n-i}b^i$.

B.2. **Conditional Expectation as a Random Variable.** In Example B.6 we had that $\mathbb{E}[X|Y=y] = y$, in Example B.7 we had that $\mathbb{E}[X|Y=y] = 0$, and in Example B.8 we had that $\mathbb{E}[X|Z=z] = zp_{\lambda,\mu}$. In all three of these examples, we could express the conditional expectation of a random variable given the outcome of some other random variable as a function of the outcome of that second random variable. This suggests a sensible definition for quantities of the form $\mathbb{E}[X|Y]$.

Definition B.9. *Let* X *and* Y *be random variables. For each* $y \in \mathbb{R}$ *such that the conditional expectation of* X *given that* $Y = y$ *is well defined, we can define the map* $h : \mathbb{R} \to \mathbb{R}$ *by*

$$h(y) \doteq \mathbb{E}[X|Y=y].$$

The ***conditional expectation of*** X ***given*** Y *is the random variable* $h(Y)$, *namely we define*

$$\mathbb{E}[X|Y] \doteq h(Y).$$

It is important to distinguish between $\mathbb{E}[X|Y]$ and $\mathbb{E}[X|Y=y]$. The quantity $\mathbb{E}[X|Y=y]$ is a real number describing the expected value of X given the *particular* outcome $Y=y$. On the other hand $\mathbb{E}[X|Y]$ is a random variable. Just as we think of a real number x as a particular realization of a random variable X, we can think of the real number $\mathbb{E}[X|Y=y]$ as a particular realization of the random variable $\mathbb{E}[X|Y]$. Note that $\mathbb{E}[X|Y]$ takes on the value $\mathbb{E}[X|Y=y]$ exactly when Y takes on the value y, namely

$$\left\{\omega \in \Omega : \mathbb{E}[X|Y](\omega) = \mathbb{E}[X|Y=y]\right\} = \left\{\omega \in \Omega : Y(\omega) = y\right\}.$$

The following proposition summarizes some important properties of conditional expectation.

Proposition B.10. *(1) (linearity) For $a, b \in \mathbb{R}$ and random variables X, Y, and Z,*

$$\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z].$$

(2) (independence) If X and Y are independent, then

$$\mathbb{E}[X|Y] = \mathbb{E}(X).$$

(3) If $Y = g(X)$ for some function g, then

$$\mathbb{E}[Y|X] = \mathbb{E}[g(X)|X] = g(X).$$

To get a clearer picture of property (3) above, note that it ensures that $\mathbb{E}[X^2|X] = X^2$, $\mathbb{E}[\sin(X)|X] = \sin(X)$, and so on.

Example B.11. *Let $X, Y \sim$ Unif$(0,1)$, and let U be independent of X and Y. Find*

$$\mathbb{E}[UX^2 + (1-U)Y^2|U].$$

Using the properties above we have

$$\mathbb{E}[UX^2 + (1-U)Y^2|U] = U\mathbb{E}[X^2|U] + (1-U)\mathbb{E}[Y^2|U] = U\mathbb{E}(X^2) + (1-U)\mathbb{E}(Y^2),$$

and we can calculate

$$\mathbb{E}(Y^2) = \mathbb{E}(X^2) = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3},$$

which yields

$$\mathbb{E}[UX^2 + (1-U)Y^2|U] = U \cdot \frac{1}{3} + (1-U)\cdot\frac{1}{3} = \frac{U}{3}.$$

The following example provides some valuable insight into conditional expectation. You will complete it in the special case of exponentially distributed random variables on your homework.

Example B.12. *Let X_1 and X_2 be i.i.d. random variables, and define $S_2 \doteq X_1 + X_2$. Then $\mathbb{E}[X_1|S_2] = \frac{S_2}{2} = \bar{X}_2$.*

B.3. **Conditional Expectation with Respect to an Event.** Suppose that we are interested in calculating the conditional expectation of a random variable given that a particular event occurs. For example, suppose that we model the amount of time that a customer must wait to be served in a store as an Exp(1) random variable. If the customer has already waited 10 minutes, how long in total should they expect to wait before being served?

We can state this question in concise probabilistic notation. Let $X \sim \text{Exp}(1)$ and let $A \doteq \{X \geq 10\}$. Then we would like to calculate $\mathbb{E}[X|A]$, so we need to make sense of this quantity. We begin by introducing the notion of an indicator function, which is a way to keep track of whether a particular event occurs or not.

Definition B.13. *Let $A \subseteq \Omega$ be some event. The function $1_A : \Omega \to \{0,1\}$ defined by*

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^c, \end{cases}$$

is known as the ***indicator function*** *of* A. *Another way to think of* 1_A *is that*

$$1_A = \begin{cases} 1, & \textit{if } A \textit{ occurs} \\ 0, & \textit{if } A \textit{ does not occur.} \end{cases}$$

This says that 1_A *is simply a random variable that takes on a value of* 0 *or* 1 *depending on whether the event* A *occurs.*

The following example illustrates how indicator functions behave.

Example B.14. *Roll a fair six-sided die. Let* A *denote the event that it lands on a* 6. *Then* 1_A *is defined as*

$$1_A = \begin{cases} 1, & \textit{if the die lands on } 6 \\ 0, & \textit{if the die does not land on } 6. \end{cases}$$

Since 1_A *is a random variable, we can calculate its expected value;*

$$\mathbb{E}[1_A] = 1 \cdot \mathbb{P}(1_A = 1) + 0 \cdot \mathbb{P}(1_A = 0) = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = 1 \cdot \frac{1}{6} + 0 \cdot \frac{5}{6} = \frac{1}{6} = \mathbb{P}(A).$$

The fact that $\mathbb{E}[1_A] = \mathbb{P}(A)$ in the example above is no coincidence, as $\mathbb{E}[1_A]$ simply measures what proportion of the time we expect 1_A to be 1, which is just given by the probability of A. In particular, for any event A, we can always write the probability of A as the expected value of its indicator function, namely

$$\mathbb{P}(A) = \mathbb{E}[1_A].$$

The following observation will be useful throughout the course.

Observation B.15. *Let* X *be a random variable. For a set* $B \subseteq \mathbb{R}$, *we have*

$$\{X \in B\} \doteq \{\omega \in \Omega : X(\omega) \in B\}.$$

Therefore, we can evaluate the function $1_{\{X \in B\}} : \Omega \to \{0,1\}$ *by noting that*

$$1_{\{X\in B\}}(\omega) = \begin{cases} 1, & \textit{if } X \in B \\ 0, & \textit{if } X \in B^c \end{cases} = \begin{cases} 1, & \textit{if } \omega \in \{X \in B\} \\ 0, & \textit{if } \omega \in \{X \in B\}^c \end{cases} = 1_B(X(\omega)).$$

Additionally, $\mathbb{P}(X \in B) = \mathbb{E}(1_{\{X \in B\}})$.

We are now ready to define the conditional expectation of a random variable X given an event.

Definition B.16. *Let* X *be a random variable and* A *be an event such that* $\mathbb{P}(A > 0)$. *Then the* ***conditional expectation of*** X ***given*** A *is given by*

$$\mathbb{E}[X|A] \doteq \frac{\mathbb{E}[X 1_A]}{\mathbb{P}(A)}.$$

Similarly, the ***conditional probability of*** A ***given*** X *is defined as*

$$\mathbb{P}[A|X] = \mathbb{E}[1_A|X].$$

In light of Observation B.15 and Definition B.16, we note the following.

Observation B.17. *If we are interested in calculating* $\mathbb{E}[X1_{\{X\in B\}}]$ *for some random variable* X *and some* $B\subseteq\mathbb{R}$*, then if we let* $g(x)\doteq x1_B(x)$*, Thoeorem* A.12 *ensures that the following hold:*

(1) If X *is discrete with p.m.f.* p_X*, then*

$$\mathbb{E}[X1_{\{X\in B\}}]=\mathbb{E}[X1_B(X)]=\mathbb{E}[g(X)]=\sum_{x\in\mathscr{S}_X}g(x)p_X(x)=\sum_{x\in\mathscr{S}_X}x1_B(x)p_X(x)=\sum_{x\in\mathscr{S}_X\cap B}xp_X(x).$$

(2) Similarly, if X *is continuous with p.d.f.* f_X*, then*

$$\mathbb{E}[X1_{\{X\in B\}}]=\mathbb{E}[X1_B(X)]=\mathbb{E}[g(X)]=\int_{-\infty}^{\infty}g(x)f_X(x)dx=\int_{-\infty}^{\infty}x1_B(x)f_X(x)dx=\int_B xf_X(x)dx.$$

Note that we can also express the conditional probabilities from Section A.7 as conditional expectations: for events A and B,

$$\mathbb{P}(A|B)=\mathbb{E}[1_A|B]=\frac{\mathbb{E}[1_A1_B]}{\mathbb{P}(B)}=\frac{\mathbb{E}[1_{A\cap B}]}{\mathbb{P}(B)}=\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}.$$

Remark B.18. *It will often be useful to use indicator functions when writing down the various distribution functions of random variables. For example, if* $X\sim Exp(\lambda)$*, then the p.d.f. of* X *is given by*

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

A simple way to write this is

$$f_X(x)=\lambda e^{-\lambda x}1_{[0,\infty)}(x).$$

We now return to the example from the beginning of this section.

Example B.19. *Suppose that the amount of time that a customer must wait to be served in a store is an* $Exp(1)$ *random variable. If the customer has already waited 10 minutes, how long in total should they expect to wait before being served?*

Let $X\sim Exp(1)$*, and let* $A\doteq\{X\geq 10\}$*. Then with* $g(x)\doteq x1_{[10,\infty)}(x)$*, we have*

$$\mathbb{E}[X|A]=\frac{\mathbb{E}[X1_A]}{\mathbb{P}(A)}=\frac{\mathbb{E}(g(X))}{\mathbb{P}(A)}.$$

Recall that $\mathbb{P}(A)=e^{-10}$ *and calculate*

$$\mathbb{E}[g(X)]=\int_{-\infty}^{\infty}g(x)f_X(x)dx=\int_0^{\infty}x1_{[10,\infty)}(x)e^{-x}dx=\int_{10}^{\infty}xe^{-x}dx=11e^{-10}.$$

Thus

$$\mathbb{E}[X|A]=\frac{11e^{-10}}{e^{-10}}=11,$$

so the customer should expect their total *wait to be 11 minutes (i.e., they will have to wait one more minute, on average).*

The following example is a good exercise.

Example B.20. *Let $X \sim Exp(2)$ and calculate the expected value of X^2 given that X is at least 1. Namely, calculate*

$$\mathbb{E}[X^2|A],$$

where $A \doteq \{X \geq 1\}$.

Clearly $\mathbb{P}(A) = e^{-2}$, and with $g(x) \doteq x^2 1_{[1,\infty)}(x)$ we have[a]

$$\mathbb{E}[X^2 1_A] = \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} x^2 1_{[1,\infty)}(x) 2e^{-2x}dx = 2\int_1^{\infty} x^2 e^{-2x}dx.$$

Integrating by parts with $u = x^2$ and $dv = e^{-2x}dx$, we have $du = 2xdx$ and $v = -\frac{1}{2}e^{-2x}$, which yields

$$\begin{aligned}\int_1^{\infty} x^2 e^{-2x}dx &= -\frac{1}{2}x^2 e^{-2x}\Big|_1^{\infty} + \frac{1}{2}\int_1^{\infty} e^{-2x}2xdx \\ &= -\frac{1}{2}\left(0 - e^{-2}\right) + \frac{1}{2}\int_1^{\infty} e^{-2x}2xdx \\ &= \frac{1}{2}e^{-2} + \frac{1}{2}\int_1^{\infty} e^{-2x}2xdx.\end{aligned} \tag{73}$$

In order to calculate the remaining integral, we integrate by parts with $u \doteq 2x$ and $dv = e^{-2x}dx$, giving $du = 2dx$ and $v = -\frac{1}{2}e^{-2x}$, which yields

$$\begin{aligned}\int_1^{\infty} e^{-2x}2xdx &= -xe^{-2x}\Big|_1^{\infty} - \int_1^{\infty} e^{-2x}dx \\ &= e^{-2} - \left(-\frac{1}{2}e^{-2x}\Big|_1^{\infty}\right) \\ &= e^{-2} + \frac{1}{2}e^{-2} \\ &= \frac{3}{2}e^{-2},\end{aligned} \tag{74}$$

so if we combine (73) and (74), we see that

$$\mathbb{E}[X^2 1_A] = 2\left(\frac{1}{2}e^{-2} + \frac{1}{2}\cdot\frac{3}{2}e^{-2}\right) = \frac{5}{2}e^{-2},$$

and therefore that

$$\mathbb{E}[X^2|A] = \frac{\frac{5}{2}e^{-2}}{e^{-2}} = \frac{5}{2}.$$

[a] See Observation B.17.

Below we consider another example of calculating conditional expectation with respect to an event.

Example B.21. *Let X and Y be jointly continuous random variables with joint p.d.f. $f_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}_+$ given by*

$$f_{(X,Y)}(x,y) = \begin{cases} x+y, & 0 < y < x < 1 \\ 0, & \textit{otherwise.}\end{cases}$$

Find the conditional expectation of Y given $X < 1/2$.

We begin by noting that

$$\mathbb{E}\left[Y \middle| X < \frac{1}{2}\right] = \frac{\mathbb{E}\left[Y1_{\{X<\frac{1}{2}\}}\right]}{\mathbb{P}\left(X<\frac{1}{2}\right)} = \frac{\mathbb{E}\left[Y1_{(0,\frac{1}{2})}(X)\right]}{\mathbb{P}\left(X<\frac{1}{2}\right)}.$$

In order to calculate the numerator of the previous expression, consider the function $g : \mathbb{R}^2 \to \mathbb{R}$ *given by*

$$g(x,y) \doteq y1_{(0,\frac{1}{2})}(x),$$

and note that

$$\begin{aligned}\mathbb{E}\left[Y1_{(0,\frac{1}{2})}(X)\right] &= \mathbb{E}(g(X,Y))\\ &= \int_0^1\int_0^1 g(x,y)f_{(X,Y)}(x,y)dydx\\ &= \int_0^{\frac{1}{2}}\int_0^1 yf_{(X,Y)}(x,y)dydx\\ &= \int_0^{\frac{1}{2}}\int_0^x y(x+y)dydx\\ &= \int_0^{\frac{1}{2}}\left(\frac{x^3}{2}+\frac{x^3}{3}\right)dx\\ &= \frac{5}{384}.\end{aligned}$$

Additionally,

$$\mathbb{P}(X<1/2) = \int_0^{1/2}\int_0^x f_{(X,Y)}(x,y)dydx = \int_0^{1/2}\int_0^x (x+y)dydx = \frac{1}{16},$$

from which we obtain

$$\mathbb{E}\left[Y \middle| X < \frac{1}{2}\right] = \frac{5}{24}.$$

The following theorem is a generalization of the law of total probability in Proposition A.20. It explains how one can use conditioning to calculate expected values.

Proposition B.22. *(Law of total expectation/Adam's law) Let* X *be a random variable and let* $\{A_i\}$ *be a (finite or countably infinite) collection of events that partition the sample space. Then*

$$\mathbb{E}(X) = \sum_i \mathbb{E}[X|A_i]\mathbb{P}(A_i).$$

The following examples illustrate how the law of total expectation can be used to simplify some problems.

Example B.23. *Roll a fair die repeatedly. How many fives do you expect to see on average before seeing a six?*

Let N *denote the number of fives that you see before seeing a six and for* $i \in \mathbb{N}$*, let* A_i *denote the event that you see your first six on the* i*-th roll. Then* $\mathbb{E}[N|A_i] = \frac{i-1}{5}$*, so*

$$
\begin{aligned}
\mathbb{E}(N) &= \sum_{i=1}^{\infty} \mathbb{E}[N|A_i]\mathbb{P}(A_i) \\
&= \sum_{i=1}^{\infty} \left(\frac{i-1}{5}\right)\left(\frac{5}{6}\right)^{i-1}\left(\frac{1}{6}\right) \\
&= \left(\frac{1}{30}\right)\sum_{i=1}^{\infty}(i-1)\left(\frac{5}{6}\right)^{i-1} \\
&= \left(\frac{1}{30}\right)\left[\sum_{i=1}^{\infty} i\left(\frac{5}{6}\right)^{i-1} - \sum_{i=1}^{\infty}\left(\frac{5}{6}\right)^{i-1}\right] \\
&= \left(\frac{1}{30}\right)\left[\sum_{i=1}^{\infty} i\left(\frac{5}{6}\right)^{i-1} - \sum_{i=0}^{\infty}\left(\frac{5}{6}\right)^{i}\right]
\end{aligned}
$$

Recall that with $f:(0,1)\to\mathbb{R}$ defined by

$$f(p) \doteq \sum_{i=1}^{\infty} p^i = \frac{p}{1-p},$$

we have

$$f'(p) = \sum_{i=1}^{\infty} i p^{i-1} = \frac{1}{(1-p)^2},$$

so

$$\mathbb{E}(N) = \left(\frac{1}{30}\right)\left[\frac{1}{\left(1-\frac{5}{6}\right)^2} - \frac{1}{1-\frac{5}{6}}\right] = 1.$$

The law of total expectation can be stated in terms of the conditional expectation between two random variables as well.

B.4. **Conditional Expectation Continued.**

Proposition B.24. *(Law of total expectation) Let X and Y be two random variables. Then*

$$\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}(X).$$

The following remark will help make sense of the proposition above.

Remark B.25. *Recall that if X and Y are random variables then there is some function h such that $\mathbb{E}[X|Y] = h(Y)$. Accordingly, if X and Y are discrete, then*

$$\mathbb{E}(X) = \mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}(h(Y)) = \sum_{y\in\mathscr{S}_Y} h(y)\mathbb{P}(Y=y) = \sum_{y\in\mathscr{S}_Y}\mathbb{E}[X|Y=y]\mathbb{P}(Y=y).$$

Similarly, if X and Y are continuous, then

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

In the case when $X = 1_A$ for some event A, we obtain

$$\mathbb{P}(A) = \mathbb{E}\left[\mathbb{P}[A|Y]\right] = \sum_{y\in\mathscr{S}_Y}\mathbb{P}[A|Y=y]\mathbb{P}(Y=y),$$

when Y is discrete, and

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{P}[A|Y]] = \int_{-\infty}^{\infty} \mathbb{P}[A|Y = y] f_Y(y)dy,$$

when Y is continuous.

Example B.26. *There are two tellers working at a bank. The amount of time that it takes the first teller to finish serving a customer follows an Exp(λ) distribution, and the amount of time that it takes the second teller to finish serving a customer follows an Exp(μ) distribution. Additionally, the time it takes for each teller to finish serving a customer is independent of the other teller. You are at the front of the line; what is the probability that you get served by the first teller?*

Let $X \sim Exp(\lambda)$ and $Y \sim Exp(\mu)$. Then we are interested in calculating $\mathbb{P}(A)$, where $A \doteq \{X < Y\}$. Using Remark B.25, *we see that*

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{E}(P(A|Y)) \\
&= \int_{-\infty}^{\infty} \mathbb{P}[A|Y = y] f_Y(y)dy \\
&= \int_{0}^{\infty} \mathbb{P}(X < y) f_Y(y)dy \\
&= \int_{0}^{\infty} (1 - e^{-\lambda y}) f_Y(y)dy \\
&= \int_{0}^{\infty} f_Y(y)dy - \mu \int_{0}^{\infty} e^{-(\lambda+\mu)y}dy \\
&= 1 - \mu\left(\frac{1}{\lambda+\mu}\right) \\
&= \frac{\lambda}{\lambda+\mu}
\end{aligned}$$

We now introduce the notion of conditional variance.

Definition B.27. *Let X and Y be random variables. The* ***conditional variance of*** *Y* ***given*** *X is defined as*

$$Var[Y|X] \doteq \mathbb{E}\left[(Y - \mathbb{E}[Y|X])^2 |X\right].$$

Two key properties of conditional variance are listed below.

Proposition B.28. *Let X and Y be random variables. The following hold:*

(1) $Var[Y|X] = \mathbb{E}[Y^2|X] - (\mathbb{E}[Y|X])^2$

(2) $Var(Y) = \mathbb{E}[Var[Y|X]] + Var(\mathbb{E}[Y|X])$

Property (2) above is known as the law of total variance.

The following example illustrates how one can apply the law of total variance.

Example B.29. *Let $Y \sim$ Geometric(p), and let*

$$(X|Y) \sim \textit{Poisson}(\mu Y),$$

where $\mu > 0$. In particular, suppose that the conditional p.m.f. of $(X|Y)$ is given by

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{e^{-\lambda y}(\lambda y)^x}{x!}.$$

Calculate $\mathbb{E}(X)$ and Var(X). Note that

$$\begin{aligned}\mathbb{E}(X|Y = y) &= \sum_{x=0}^{\infty} x p_{X|Y}(x|y)\\ &= 0 \cdot \frac{e^{-\lambda y}(\lambda y)^0}{0!} + \sum_{x=1}^{\infty} x \frac{e^{-\lambda y}(\lambda y)^x}{x!}\\ &= \sum_{x=1}^{\infty} \frac{e^{-\lambda y}(\lambda y)^x}{(x-1)!}\\ &= e^{-\lambda y}(\lambda y) \sum_{x=1}^{\infty} \frac{(\lambda y)^{x-1}}{(x-1)!}\\ &= e^{-\lambda y}(\lambda y) \sum_{x=0}^{\infty} \frac{(\lambda y)^x}{x!}\\ &= \lambda y.\end{aligned}$$

Therefore, $\mathbb{E}[X|Y] = \lambda Y$, so $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}[X|Y]) = \mathbb{E}(\lambda Y) = \frac{\lambda}{p}$. A similar calculation shows that

$$\textit{Var}(X|Y) = \lambda Y,$$

so the law of total variance tells us that

$$\textit{Var}(X) = \mathbb{E}(\textit{Var}(X|Y)) + \textit{Var}(E[X|Y]) = \mathbb{E}(\lambda Y) + \textit{Var}(\lambda Y) = \frac{\lambda}{p} + \lambda^2 \frac{1-p}{p^2} = \frac{\lambda^2 + \lambda p - \lambda^2 p}{p^2}.$$

The following proposition is known as Wald's Identity.

Proposition B.30. *Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables with $\mu_X \doteq \mathbb{E}(X_i)$ and $\sigma_X^2 \doteq$ Var(X_i), and let N be a random variable which takes values in $\mathbb{N}$ and which is independent of $X_1, X_2, \ldots$. Let $\mu_N \doteq \mathbb{E}(N)$ and $\sigma_N^2 \doteq$ Var(N). Then*

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}(X_1)\mathbb{E}(N) = \mu_N \mu_X$$

and

$$\textit{Var}\left(\sum_{i=1}^{N} X_i\right) = \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] &= \sum_{n=1}^{\infty} \mathbb{E}\left[\sum_{i=1}^{n} X_i \middle| N = n\right] \mathbb{P}(N = n) \\
&= \sum_{n=1}^{\infty} \sum_{i=1}^{n} \mathbb{E}[X_i | N = n] \mathbb{P}(N = n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(N = n) \sum_{i=1}^{n} \mathbb{E}(X_i) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(N = n) n \mu_X \\
&= \mu_X \left[\sum_{n=1}^{\infty} n \mathbb{P}(N = n)\right] \\
&= \mu_X \mathbb{E}(N) \\
&= \mu_N \mu_X
\end{aligned}$$

Additionally, the law of total variance tells us that

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}\left(\mathrm{Var}\left[\sum_{i=1}^{N} X_i \middle| N\right]\right) + \mathrm{Var}\left(\mathbb{E}\left[\sum_{i=1}^{N} X_i \middle| N\right]\right).$$

Using the definition of conditional variance, we can check that

$$\mathrm{Var}\left[\sum_{i=1}^{N} X_i \middle| N\right] = N \mathrm{Var}(X_1) = N \sigma_X^2,$$

so

$$\mathbb{E}\left(\mathrm{Var}\left[\sum_{i=1}^{N} X_i \middle| N\right]\right) = \mathbb{E}[N \sigma^2] = \mu_N \sigma_X^2.$$

Additionally,

$$\mathrm{Var}\left(\mathbb{E}\left[\sum_{i=1}^{N} X_i \middle| N\right]\right) = \mathrm{Var}(N \mu_X) = \mu_X^2 \mathrm{Var}(N) = \mu_X^2 \sigma_N^2.$$

It follows that

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2.$$

as claimed. □

In order to provide some more intuition into conditional expectation, we begin by recalling an important notion from statistics. For a given random variable X, what is the best (deterministic) predictor/estimate of the outcome of X? There are many ways to define the 'best' predictor, but you may recall from other statistics courses that one if often interested in minimizing the mean-squared error. That is to say that one is interested in finding the (deterministic) quantity $c_* \in \mathbb{R}$ that minimizes the function $MSE : \mathbb{R} \to \mathbb{R}_+$ given by

$$MSE(c) \doteq \mathbb{E}\left((X - c)^2\right).$$

As expected, our best guess for the outcome of X is its expected value, namely we have that $c_* = \mathbb{E}(X)$, which yields $MSE(c_*) = \mathrm{Var}(X)$.

In the setting above we were not given any information about the random variable X. Suppose that we are given some additional information about X through some other random variable Y. For instance, we might have that X denotes the amount of ice cream sold on some day, and Y denotes the temperature outside that day. In this setting we expect that if Y is large, then X will be large as well.

We can think of estimates that use information about Y to predict X as functions of the random variable Y. For example, a simple such function is $h(Y) \doteq \mathbb{E}(X)$. Note that this function is constant and therefore doesn't actually use any information about Y. The following theorem states that the best estimate (in terms of mean-squared error) of X that we can construct using Y is simply the conditional expectation of X given Y. Consequently, we can interpret $\mathbb{E}[X|Y]$ as the best estimate for X given that we observe Y.

Theorem B.31. *For any function h we have*

$$\mathbb{E}\left[(X - h(Y))^2\right] \geq \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\right],$$

and equality holds in the expression above if and only if $h(Y) \doteq \mathbb{E}[X|Y]$.

The field of **stochastic filtering** is concerned with questions such as these.

APPENDIX C. LIMIT THEOREMS

C.1. **Law of Large Numbers.** We begin by recalling several important results and definitions from undergraduate probability.

Definition C.1. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables, and let X be another random variable. We say that X_n* ***converges to*** *X* ***in probability*** *and write $X_n \xrightarrow{\mathbb{P}} X$ if for each $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

The weak law of large numbers is a fundamental example of convergence in probability. In order to state it we recall Markov's inequality.

Proposition C.2. *(Markov's inequality) Let X be a nonnegative random variable. Then for each $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. Note that $a1_{\{X\geq a\}} \leq X$, so

$$\mathbb{E}(a1_{\{X\geq a\}}) = a\mathbb{E}(1_{\{X\geq a\}}) = a\mathbb{P}(X \geq a) \leq \mathbb{E}(X).$$

The result follows on dividing both sides of the inequality of a. □

The following result is known as Chebyshev's inequality.

Corollary C.3. *Let X be a random variable with $\mathbb{E}(X) = \mu$ and Var$(X) = \sigma^2 < \infty$. Then for each $a > 0$,*

$$\mathbb{P}(|X - \mu| > a) \leq \frac{\sigma^2}{a^2}.$$

Proof. Applying Markov's inequality to the non-negative random variable $(X-\mu)^2$, we have

$$\mathbb{P}(|X-\mu| \geq a) = \mathbb{P}((X-\mu)^2 \geq a^2) \leq \frac{\mathbb{E}((X-\mu)^2)}{a^2} = \frac{\sigma^2}{a^2}.$$

□

The following result shows that the sample average of i.i.d. random variables converges in probability to their expected value.

Theorem C.4. *Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables such that $\mathbb{E}(X_n) = \mu$ and Var$(X_n) = \sigma^2 < \infty$. Then $\bar{X}_n \xrightarrow{\mathbb{P}} \mu$.*

Proof. Fix $\epsilon > 0$. Then

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\sigma^2}{n},$$

so Chebyshev's inequality ensures that

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \le \frac{\frac{\sigma^2}{n}}{\epsilon} = \frac{1}{n} \cdot \frac{\sigma^2}{\epsilon},$$

which tends to 0 as $n \to \infty$. □

Below we introduce the notion of almost sure convergence.

Definition C.5. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables, and let X be another random variable. We say that X_n converges to X almost surely (or with probability one) and write $X_n \xrightarrow{a.s.} X$ if*

$$\mathbb{P}\left(\lim_{n\to\infty} X_n \to X\right) = 1.$$

Note that convergence in probability is weaker than almost sure convergence, namely, if $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$, but $X_n \xrightarrow{\mathbb{P}} X$ does not necessarily tell us that $X_n \xrightarrow{a.s.} X$.

The following is an example of a sequence of random variables that converge in probability but not almost surely.

Example C.6. *Consider a sequence $\{X_n\}_{n=1}^{\infty}$ of independent random variables, where the probability distribution of X_n is given by*

$$p_n \doteq \mathbb{P}(X_n = 1), \qquad q_n \doteq \mathbb{P}(X_n = 0).$$

Suppose that $p_n \to 0$ as $n \to \infty$. Let $X = 0$ be a random variable that is always 0. Then for each $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon) \le \lim_{n\to\infty} \mathbb{P}(X_n > 0) = \lim_{n\to\infty} p_n = 0,$$

so $X_n \xrightarrow{\mathbb{P}} 0$.

A natural question is whether $X_n \xrightarrow{a.s.} 0$. It turns out that this depends on how quickly the sequence $\{p_n\}$ converges to 0. In particular, if

$$\sum_{n=1}^{\infty} p_n < \infty,$$

then $X_n \xrightarrow{a.s.} 0$. This holds, for example, if $p_n \doteq \frac{1}{n^2}$. On the other hand, if

$$\sum_{n=1}^{\infty} p_n = \infty,$$

such as when $p_n = \frac{1}{n}$, then the sequence $\{X_n\}$ does not converge almost surely to 0, even though it does converge to 0 in probability.

The proof that X_n does not converge almost surely to X when $\{p_n\}$ is not summable relies on the Borel-Cantelli Lemma. Intuitively, even though the sequence $\{p_n\}$ converges to 0, it does so at a 'slow' rate. This means that each X_n has a 'large enough' change of being equal to 1 that infinitely many of the X_n's will always equal 1, which of course means that the sequence cannot converge to 0.

Convergence almost surely says that the sequence X_n *always* converges to X, while convergence in probability says only that if n is large, then there is a small change that X_n is far from X. In practice, almost sure convergence is a more desirable property.

For example, the weak law of large numbers stated in Theorem C.4 ensures that if we take a large enough sample from the population, then there is a high probability that our sample average is close to the population mean. The strong law of large numbers (stated below) says that if we take a large enough sample, then the sample average will *definitely* be close to the population mean. The strong law of large number is stated below.

Theorem C.7. *(Strong law of large numbers) Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables such that $\mathbb{E}(X_n) = \mu$ and $Var(X_n) = \sigma^2 < \infty$. Then $\bar{X}_n \overset{a.s.}{\to} \mu$.*

The proof of the strong law of large numbers is much more difficult than the proof of the weak law.

Remark C.8. *While the strong law of large numbers tells us that the sample average of i.i.d. random variables will eventually converge to the population mean, there are some things that it doesn't tell us. For instance, suppose that we wanted to calculate or estimate*

$$\mathbb{P}(\bar{X}_n - \mu > 10),$$

namely the probability that the sample average of n observations is at least 10 *units larger than the population mean. The strong and weak laws tells us that this probability goes to* 0 *as $n \to \infty$, but they don't say anything about how quickly it goes to* 0.

It turns out that the central limit theorem from introductory statistics allows us to estimate the probability of events like these, even if we don't know the probability distribution of the X_n's.

Before we introduce the central limit theorem we first recall some important properties of moment generating functions.

C.2. **Moment Generating Functions.** We begin by recalling the definition of the moment generating function of a random variable.

Definition C.9. *Let X be a random variable and suppose that there is some $a > 0$ such that for all $t \in [-a, a]$, $\mathbb{E}(e^{tX}) < \infty$. Then we say that the* ***moment generating function (MGF)*** *of X exists; it is the function defined as*

$$m_X(t) \doteq \mathbb{E}[e^{tX}], \quad t \in [-a, a].$$

The MGF of a random variable can be used to compute its moments.

Theorem C.10. *Let X be a random variable with MGF m_X. Let*

$$m_X^{(n)}(t) \doteq \frac{\partial^n}{\partial t^n} m_X(t),$$

denote the n-th derivative of m_X with respect to t. Then for each $n \in \mathbb{N}$,

$$m_X^{(n)}(0) = \mathbb{E}(X^n).$$

To get some intuition for where Theorem C.10 comes from, use the Taylor series expansion of e^{tx} to write

$$m_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right).$$

If we can exchange the expected value and the summation, then this becomes

$$m_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbb{E}(X^n).$$

Then, assuming that we can exchange differentiation with the summation, we have

$$\begin{aligned}\frac{\partial}{\partial t} m_X(t) &= \frac{\partial}{\partial t}\left(1 + \frac{t^1}{1!}\mathbb{E}(X) + \frac{t^2}{2!}\mathbb{E}(X^2) + \frac{t^3}{3!}\mathbb{E}(X^3)\cdots\right)\\ &= 0 + \mathbb{E}(X) + \frac{2t}{2!}\mathbb{E}(X^2) + \frac{3t^2}{3!}\mathbb{E}(X^3) + \cdots\\ &= \mathbb{E}(X) + t\mathbb{E}(X^2) + \frac{t^2}{2!}\mathbb{E}(X^3) + \cdots,\end{aligned}$$

and if we plug in $t = 0$, we obtain

$$\frac{\partial}{\partial t} m_X(0) = 0 + \mathbb{E}(X) + \frac{2\cdot 0}{2}\mathbb{E}(X^2) + \frac{3\cdot 0^2}{3!}\mathbb{E}(X^3) + \cdots = \mathbb{E}(X).$$

Additionally,

$$\begin{aligned}\frac{\partial^2}{\partial t^2} m_X(t) &= \frac{\partial}{\partial t}\left(\frac{\partial}{\partial t} m_X(t)\right)\\ &= \frac{\partial}{\partial t}\left(\mathbb{E}(X) + t\mathbb{E}(X^2) + \frac{t^2}{2!}\mathbb{E}(X^3) + \cdots\right)\\ &= 0 + \mathbb{E}(X^2) + \frac{2t}{2!}\mathbb{E}(X^3) + \cdots,\end{aligned}$$

so plugging in $t = 0$ yields

$$\frac{\partial^2}{\partial t^2} m_X(0) = 0 + \mathbb{E}(X^2) + \frac{2\cdot 0}{2!}\mathbb{E}(X^3) + \cdots = \mathbb{E}(X^2).$$

The following example involves calculating the MGFs of several probability distributions.

Example C.11. *Compute the MGFs of the following probability distributions:*

(1) Bernoulli(p)
(2) Poisson(λ)
(3) $\mathcal{N}(0,1)$

We begin with (1). Note that if $X \sim$ Bernoulli(p), then

$$m_X(t) = \mathbb{E}(e^{tX}) = pe^{t\cdot 1} + (1-p)e^{t\cdot 0} = pe^t + 1 - p.$$

If $Y \sim$ Poisson(λ), then

$$\begin{aligned} m_Y(t) &= \mathbb{E}(e^{tY}) \\ &= \sum_{y=0}^{\infty} \mathbb{P}(Y=y)e^{ty} \\ &= \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!}e^{ty} \\ &= e^{-\lambda}\sum_{y=0}^{\infty}\frac{(\lambda e^t)^y}{y!} \\ &= e^{-\lambda}e^{\lambda e^t} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

If $Z \sim \mathcal{N}(0,1)$, note that

$$e^{-\frac{z^2}{2}}e^{tz} = e^{-\frac{z^2}{2}+tz} = e^{-\frac{z^2}{2}+zt-\frac{t^2}{2}+\frac{t^2}{2}} = e^{-\left(\frac{z^2-2zt+t^2}{2}\right)}e^{\frac{t^2}{2}} = e^{-\frac{(z-t)^2}{2}}e^{\frac{t^2}{2}},$$

so

$$\begin{aligned} m_Z(t) &= \mathbb{E}(e^{tZ}) \\ &= \int_{-\infty}^{\infty} e^{tz} f_Z(z)dz \\ &= \int_{-\infty}^{\infty} e^{tz}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-t)^2}{2}}e^{\frac{t^2}{2}}dz \\ &= e^{\frac{t^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-t)^2}{2}}dz \\ &= e^{\frac{t^2}{2}}\int_{-\infty}^{\infty} f_{Z_t}(z)dz \\ &= e^{\frac{t^2}{2}}, \end{aligned}$$

where we have used the fact that $f_{Z_t}(z) \doteq \frac{1}{\sqrt{2\pi}}e^{-\frac{(z-t)^2}{2}}$ is the p.d.f. of the $\mathcal{N}(t,1)$ distribution, and therefore that

$$\int_{-\infty}^{\infty} f_{Z_t}(z)dz = 1.$$

In Example C.11, all of the moment generating functions were defined for all $t \in \mathbb{R}$. Below we consider an example where a moment generating function exists only in some neighborhood of the origin.

Example C.12. *Let $X \sim$ Exp(λ). Then for $t < \lambda$,*

$$\mathbb{E}[e^{tX}] = \int_0^{\infty} e^{tx}\lambda e^{-\lambda x}dx = \lambda\int_0^{\infty} e^{(t-\lambda)x}dx = \frac{\lambda}{\lambda - t},$$

but if $t > \lambda$, then

$$\int_0^a e^{(t-\lambda)x}dx \overset{a\to\infty}{\to} \infty,$$

so the moment generating function is not well defined.

Below we summarize some important properties of MGFs.

Proposition C.13. *(1) If X and Y are independent random variables with MGFs m_X and m_Y, respectively, then the MGF of $X+Y$ is given by*

$$m_{X+Y}(t) = m_X(t) m_Y(t).$$

(2) For constants $a, b \in \mathbb{R}$, if we let $Y \doteq a + bX$, then the MGF of Y is given by

$$m_Y(t) = e^{at} m_X(bt).$$

(3) The MGF uniquely characterizes the probability distribution of a random variables. That is, if there is some $a > 0$ such that

$$m_X(t) = m_Y(t), \quad \text{for all } t \in (-a, a),$$

then X and Y have the same probability distribution, which we denote by $X \overset{d}{=} Y$.

We can use the previous proposition to find the probability distribution of linear combinations of random variables.

Example C.14. *Let $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then $Y \overset{d}{=} \mu + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$. Thus,*

$$m_Y(t) = \mathbb{E}(e^{(\mu+\sigma Z)t}) = e^{\mu t}\mathbb{E}(e^{\sigma t Z}) = e^{\mu t} m_Z(\sigma t) = e^{\mu t} e^{\frac{(\sigma t)^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

This allows us to compute the distribution of sums of independent normal random variables as well. For example, if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then the MGF of $S \doteq X_1 + X_2$ is given by

$$m_S(t) = m_{X_1+X_2}(t) = m_{X_1}(t) m_{X_2}(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}} e^{\mu_2 t + \frac{\sigma_2^2 t^2}{2}} = e^{(\mu_1+\mu_2)t + \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}$$

which is the MGF of a $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ distribution.

In the example below we derive the distribution of the sum of independent Poisson random variables.

Example C.15. *For $i = 1, 2, \ldots, n$, let $X_i \sim$ Poisson(λ_i) be independent random variables, where $\lambda_1, \lambda_2, \ldots, \lambda_n > 0$. What is the probability distribution of $S_n \doteq X_1 + X_2 + \cdots + X_n$?*

We have

$$m_{S_n}(t) = m_{X_1+X_2+\cdots+X_n}(t) = \prod_{i=1}^{n} m_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{\left(\sum_{i=1}^{n} \lambda_i\right)(e^t-1)},$$

which says that $S_n \sim$ Poisson$\left(\sum_{i=1}^{n} \lambda_i\right)$.

We recall the notion of convergence in distribution.

C.3. **Central Limit Theorem.** We begin by defining another form of convergence of random variables. The notion of convergence, called convergence in distribution, captures what it means for the probability distributions of a sequence of random variables to converge to another probability distribution. We begin by considering the following example.

Example C.16. *Let $\{X_n\}_{n=1}^{\infty}, X \sim Exp(\lambda)$. Then every X_n has the same probability distribution as X, so we would a notion of convergence that says that X_n is 'close' to X, at least in terms of their probability distributions. We can see that almost sure convergence and convergence in probability do not capture this. To see this note that for all $n \in \mathbb{N}$ and $\epsilon > 0$,*

$$\mathbb{P}[|X_n - X| > \epsilon] = 2\mathbb{P}(Y - X > \epsilon) \doteq c,$$

where $X, Y \overset{iid}{\sim} Exp(\lambda)$. Note that $c > 0$ and c does not depend on n, so this shows $\{X_n\}$ does not converge to X in probability, and therefore that it doesn't converge almost surely either.

This suggests another form of convergence is needed to study the convergence of random variables' probability distributions.

We now formally introduce convergence in distribution.

Definition C.17. *For each $n \in \mathbb{N}$, let X_n be a random variables with CDF F_n, and let X be another random variable with CDF F_X. Let $\mathscr{C}_X \doteq \{x : F_X \text{ is continuous at } x\}$. If*

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \textit{for all } x \in \mathscr{C}_X,$$

then we say that X_n ***converges in distribution*** *(or converges weakly, or converges in law) to X and we write $X_n \xrightarrow{d} X$.*

The intuition behind convergence in distribution is that for large values of $n \in \mathbb{N}$, if $X_n \xrightarrow{d} X$, then, roughly speaking,

$$\mathbb{P}(X_n \in A) \approx \mathbb{P}(X \in A), \quad A \subseteq \mathbb{R}.$$

Recall the central limit theorem, which says that if $\{X_n\}_{n=1}^{\infty}$ are i.i.d. random variables with mean μ and variance σ^2, then with

$$\bar{X}_n \doteq \frac{\sum_{i=1}^{n} X_i}{n}$$

we have that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1).$$

While we will come back to the central limit theorem, it really says that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0,1)$.

The following theorem says that MGFs can be used to characterized convergence in distribution.

Theorem C.18. *(Continuity theorem) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with MGFs $\{m_{X_n}\}_{n=1}^{\infty}$. Assume that X is a random variable with MGF m_X and that there is some $a > 0$ such that for all $t \in (-h, h)$, the MGFs of the X_n's and X are all defined on $(-a, a)$. Under this assumption, if*

$$\lim_{n\to\infty} m_{X_n}(t) = m_X(t), \quad \textit{for all } t \in (-a, a)$$

then $X_n \xrightarrow{d} X$.

Using Theorem C.18, one can use MGFs to prove the central limit theorem, which we state precisely below.

Theorem C.19. *(Central limit theorem) Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_n) = \mu$ and $Var(X_n) = \sigma^2$. Then*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0,1)$.

Proof. In light of Theorem C.18, we only consider the case when the MGF of the $X_i's$ and the MGF of X exist on some interval $(-a,a)$, where $a > 0$.

We begin by considering the case when $\mu = 0$; then we have

$$m_{X_1}^{(1)}(0) = \mathbb{E}(X) = 0, \qquad m_{X_1}^{(2)}(0) = \mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \sigma^2.$$

For each $n \in \mathbb{N}$, let

$$Z_n \doteq \frac{\bar{X}_n}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\bar{X}_n.$$

Using Proposition C.13, we can see that

$$m_{Z_n}(t) = m_{\bar{X}_n}\left(\frac{\sqrt{n}}{\sigma}t\right), \tag{75}$$

and

$$\begin{aligned} m_{\bar{X}_n}(s) &= m_{\frac{X_1}{n}+\frac{X_2}{n}+\cdots+\frac{X_n}{n}}(s) \\ &= \prod_{i=1}^{n} m_{\frac{X_i}{n}}(s) \\ &= \prod_{i=1}^{n} m_{\frac{X_i}{n}}\left(\frac{s}{n}\right) \\ &= \left[m_{X_1}\left(\frac{s}{n}\right)\right]^n. \end{aligned} \tag{76}$$

Combining (75) and (76), we obtain

$$m_{Z_n}(t) = m_{\bar{X}_n}\left(\frac{\sqrt{n}}{\sigma}t\right) = \left[m_{X_1}\left(\frac{\frac{\sqrt{n}}{\sigma}t}{n}\right)\right]^n = \left[m_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

To evaluate the limit of $m_{Z_n}(t)$ as $n \to \infty$, we begin by considering $\log m_{Z_n}(t)$:

$$\begin{aligned} \log m_{Z_n}(t) &= n \log m_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &\doteq \frac{1}{u_n^2}\log m_{X_1}(a_t u) \qquad \left(\text{with } a_t \doteq \frac{t}{\sigma} \text{ and } u_n \doteq \frac{1}{\sqrt{n}}\right). \end{aligned}$$

Since $u_n \to 0$ as $n \to \infty$, it follows that if

$$\lim_{n\to\infty} \log m_{Z_n}(t),$$

exists, then it will be equal to

$$\lim_{u\to 0} \frac{1}{u^2}\log m_{X_1}(a_t u).$$

We can apply L'Hospital's Rule to see that

$$\begin{aligned}\lim_{u\to 0}\frac{\log m_{X_1}(a_t u)}{u^2} &= \lim_{u\to 0}\frac{\frac{\partial}{\partial u}\log m_{X_1}(a_t u)}{\frac{\partial}{\partial u}u^2}\\ &= \lim_{u\to 0}\frac{a\frac{m^{(1)}_{X_1}(a_t u)}{m_{X_1}(a_t u)}}{2u}\\ &= \frac{a}{2}\lim_{u\to 0}\left(\frac{1}{m_{X_1}(a_t u)}\cdot\frac{m^{(1)}_{X_1}(a_t u)}{u}\right).\end{aligned}\tag{77}$$

Applying L'Hospital's Rule once more,

$$\lim_{u\to 0}\frac{m^{(1)}_{X_1}(a_t u)}{u} = \lim_{u\to 0}\frac{am^{(2)}_{X_1}(au)}{1} = am^{(2)}_{X_1}(0) = a\sigma^2,\tag{78}$$

and noting that

$$\lim_{u\to 0}\frac{1}{m_{X_1}(a_t u)} = \frac{1}{m_{X_1}(0)} = 1,$$

we can combine (77) and (78) to see that

$$\lim_{u\to 0}\frac{1}{u^2}\log m_{X_1}(a_t u) = \frac{a_t}{2}\cdot 1\cdot a_t\sigma^2 = \frac{a_t^2\sigma^2}{2} = \frac{t^2}{2}$$

As we noted before, it follows that

$$\lim_{n\to\infty}\log m_{Z_n}(t) = \lim_{u\to 0}\frac{1}{u^2}\log m_{X_1}(a_t u) = \frac{t^2}{2},$$

from which we obtain, using the continuity of $x\mapsto e^x$, that

$$\lim_{n\to\infty} m_{Z_n}(t) = e^{\frac{t^2}{2}},$$

which is the MGF of the $\mathcal{N}(0,1)$ distribution. Therefore, $Z_n \xrightarrow{d} Z$, where $Z\sim\mathcal{N}(0,1)$.

To complete the proof in the case when $\mu = \mathbb{E}(X_n)\neq 0$, define $Y_n \doteq X_n - \mu$, so that $\mathbb{E}(Y_n) = 0$. We have already shown that $\tilde{Z}_n \xrightarrow{d} Z$, where

$$\tilde{Z}_n \doteq \frac{\bar{Y}_n}{\sigma/\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)}{\sigma/\sqrt{n}} = \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}},$$

so the result follows. □

Note that we can formulate Theorem C.19 in terms of sums as well.

Remark C.20. *Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables with $\mathbb{E}(X_n)=\mu$ and $Var(X_n)=\sigma^2$. If we let*

$$S_n \doteq \sum_{i=1}^{n} X_i,$$

then $\mathbb{E}(S_n) = n\mu$ and $Var(S_n) = n\sigma^2$, so the central limit theorem says that

$$\frac{S_n-\mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{\frac{1}{n}(S_n-n\mu)}{\frac{1}{n}\sqrt{n\sigma^2}} = \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z,$$

where $Z \sim \mathcal{N}(0,1)$. Equivalently, for each $a, b \in \mathbb{R}$, we have that

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} \le b\right) = \Phi(b) - \Phi(a),$$

where

$$\Phi(x) \doteq \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx,$$

denotes the c.d.f. of the $\mathcal{N}(0,1)$ distribution.

C.4. **Convergence Theorems and Estimating π.** Consider the unit square $S = \{(x, y) : 0 \le x, y \le 1\}$, and draw a quarter circle inside of it, namely plot the function

$$f(x) = \sqrt{1-x^2}, \ \ x \in [0,1].$$

Let $C \doteq \{(x, f(x)) : x \in [0,1]\}$ denote the quarter circle. We know that the area of a circle with radius $r = 1$ is π, so the area of this quarter circle is $A(C) = \frac{\pi}{4}$. For each $n \in \mathbb{N}$, let $X_n, Y_n \sim$ Unif(0,1) be independent. Each random variable (X_n, Y_n) simply chooses a point uniformly at random on the unit square S, so it follows that for each $n \in \mathbb{N}$,

$$\mathbb{P}((X_n, Y_n) \in C) = \frac{\pi}{4}.$$

For each $n \in \mathbb{N}$, let $U_n \doteq 1_C(X_n, Y_n)$ and note that $\{U_n\}$ is a sequence of i.i.d. random variables such that

$$\mu \doteq \mathbb{E}(U_n) = \mathbb{E}(1_C(X_n, Y_n)) = \mathbb{P}((X_n, Y_n) \in C) = \frac{\pi}{4}.$$

Then the sample average

$$\bar{U}_n \doteq \frac{1}{n}\sum_{i=1}^{n} U_n,$$

simply measures the proportion of pairs (X_n, Y_n) that landed in the unit circle. To estimate $\tilde{\mu} \doteq \pi$, we let $V_n \doteq 4U_n$ and $\bar{V}_n \doteq 4\bar{U}_n$, so that $\mathbb{E}(V_n) = 4\mu = \pi$. Here, the V_n's are again i.i.d. random variables, so the strong law of large numbers ensures that $\bar{V}_n \overset{a.s.}{\to} \pi$. In view of this, to estimate π we just need to generate many i.i.d. Unif(0,1) random variables.

We can establish somewhat stronger results as well. To begin, we note that[5]

$$\sigma^2 \doteq \mathrm{Var}(V_n) = \mathrm{Var}(4U_n) = 4^2\left(\mathbb{E}(U_n^2) - (\mathbb{E}(U_n))^2\right) = 16\left(\mathbb{E}(U_n) - (\mathbb{E}(U_n))^2\right) = 16 \cdot \frac{\pi}{4}\left(1 - \frac{\pi}{4}\right) = \pi\,(4-\pi),$$

so the central limit theorem ensures that if n is sufficiently large, then

$$\begin{aligned}\mathbb{P}(|\bar{V}_n - \pi| \le \epsilon) &= \mathbb{P}(-\epsilon \le \bar{V}_n - \pi \le \epsilon)\\ &= \mathbb{P}\left(\frac{-\epsilon}{\sigma/\sqrt{n}} \le \frac{\bar{V}_n - \pi}{\sigma/\sqrt{n}} \le \frac{\epsilon}{\sigma/\sqrt{n}}\right)\\ &\approx \Phi\left(\frac{\epsilon}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{\epsilon}{\sigma/\sqrt{n}}\right).\end{aligned}$$

Using this, we can estimate the probability that our estimate for π is within 0.001 of the true value if we have $n = 1000$. We obtain

$$\mathbb{P}(|\bar{V}_{1000} - \pi| \le 0.001) \approx \Phi\left(\frac{0.001}{\pi(4-\pi)/\sqrt{1000}}\right) - \Phi\left(-\frac{0.001}{\pi(4-\pi)/\sqrt{1000}}\right) \approx 0.2892243.$$

If we had $n = 10000$, then we obtain

$$\mathbb{P}(|\bar{V}_{10000} - \pi|| \le 0.001) \approx 0.9997912.$$

[5] Here we use the fact that since $U_n \in \{0,1\}$, we have $U_n^2 = U_n$.

This is a simple example of a Monte Carlo algorithm. This simple example is extremely powerful, as the same approach works to estimate the area of much more complicated shapes, even those lying in high dimensional space such as $\mathbb{R}^d$, where $d \gg 10000$.

Below we present another Monte Carlo algorithm that can be used to estimate π. Recall that

$$\int_0^1 \sqrt{1-x^2}\,dx = \frac{\pi}{4},$$

and let $X \sim \text{Unif}(0,1)$. Then we can calculate

$$\mathbb{E}\left(\sqrt{1-X^2}\right) = \int_0^1 \sqrt{1-x^2} f_X(x)\,dx = \int_0^1 \sqrt{1-x^2}\,dx = \frac{\pi}{4}.$$

Additionally, the law of large numbers tells us that if $\{X_n\}_{n=1}^{\infty}$ is an i.i.d. collection of Unif(0,1) random variables, then if we let $Y_n \doteq 4\sqrt{1-X_n^2}$ for each $n \in \mathbb{N}$, then $\{Y_n\}_{n=1}^{\infty}$ are i.i.d. random variables with $\mathbb{E}(Y_n) = 4\mathbb{E}\left(\sqrt{1-X_n^2}\right) = \pi$. Then the law of large numbers tells us that $\bar{Y}_n \overset{a.s.}{\to} \pi$, so we can estimate π with $\bar{Y}_n$ for large values of n.

Example C.21. *Using the central limit theorem, determine (approximately) how large of a sample n we would need so that the following (approximately) holds:*

$$\mathbb{P}(|\bar{Y}_n - \pi| > 0.001) \le 0.0001.$$

We begin by noting that this is equivalent to find n such that

$$\mathbb{P}(|\bar{Y}_n - \pi| \le 0.001) \ge 0.9999.$$

In order to calculate $\sigma^2 \doteq Var(Y_n)$*, we first calculate* $\mathbb{E}(Y_n^2)$*;*

$$\mathbb{E}(Y_n^2) = 16\mathbb{E}(1-X_n^2) = 16(1-\mathbb{E}(X_n^2)) = 16\left(1 - \int_0^1 x^2\,dx\right) = 16\left(1-\frac{1}{3}\right) = \frac{32}{3},$$

which tells us that $\sigma^2 = \frac{32}{3} - \pi^2$*. Additionally, if* $Z \sim \mathcal{N}(0,1)$*, we have*

$$\mathbb{P}(|\bar{Y}_n - \pi| > 0.001) > \mathbb{P}\left(\left|\frac{\bar{Y}_n - \pi}{\sigma/\sqrt{n}}\right| > \frac{0.001}{\sigma/\sqrt{n}}\right) \approx \mathbb{P}\left(|Z| > \frac{0.001}{\sigma/\sqrt{n}}\right) = 2\Phi\left(-\frac{0.001}{\sigma/\sqrt{n}}\right),$$

so we want to find n such that

$$\Phi\left(-\frac{0.001}{\sigma/\sqrt{n}}\right) \le \frac{0.0001}{2} = 0.00005,$$

which is equivalent to

$$n \ge \left(\frac{-\Phi^{-1}(0.00005)\sigma}{0.001}\right)^2 = 9616474.$$

APPENDIX D. RANDOM WALKS AND STOCHASTIC PROCESSES

We begin by introducing the notion of a stochastic process.

Definition D.1. *A sequence of random variables $\{S_t\}_{t\in\mathscr{I}}$, where $\mathscr{I}$ is some index set is known as a stochastic process. We typically have $\mathscr{I} = \mathbb{N}_0$ or $\mathscr{I} = \mathbb{R}_+$, in which case we can interpret S_t as the value of the process at time t.*

In this course we will usually be dealing with *discrete-time stochastic processes*, that is processes for which $\mathscr{I} = \mathbb{N}_0$, in which case we usually write $\{S_n\} \doteq \{S_n\}_{n=0}^{\infty}$. While we are familiar with sequences of i.i.d. random variables, in general the values of stochastic processes at different time instants will not be independent or identically distributed. For example, if we let S_n the temperature on day n of 2022. Then, if S_n is large (i.e,. it is warm on day n), then we should also expect that S_{n+1} will be large as well. Additionally, for $n = 0,\dots,30$, it is winter, so we expect S_n to be much smaller than when $n = 240,\dots,270$, when it is summer.

We begin by considering a simpler example of a stochastic process. You start with initial wealth S_0 (for now assume that the amount is deterministic), then you make repeated bets of \$1 on fair coin tosses, namely you win \$1 if the coin lands on heads and you lose \$1 if the coin lands on tails. Let S_n denote your wealth after n bets; then S_n is a random variable for each $n \in \mathbb{N}$. Therefore, the sequence $\{S_n\}_{n=0}^{\infty}$ is a stochastic process.

If we let X_t denote the winnings from bet t, so that

$$X_t = \begin{cases} 1, & \text{coin flip } t \text{ lands on heads} \\ -1, & \text{coin flip } t \text{ lands on tails,} \end{cases}$$

then we have

$$S_n = S_0 + \sum_{s=1}^{t} X_t. \tag{79}$$

Unlike many of the settings that one sees in probability and statistics classes, here the random variables $\{S_n\}_{t=0}^{\infty}$ are neither identically distributed nor independent. To see why they are not identically distributed, note that

$$S_1 = \begin{cases} S_0 + 1, & \text{with probability } 1/2 \\ S_0 - 1, & \text{with probability } 1/2, \end{cases}$$

while

$$S_2 = \begin{cases} S_0 - 2, & \text{with probability } 1/4 \\ S_0, & \text{with probability } 1/2 \\ S_0 + 2, & \text{with probability } 1/4, \end{cases}$$

and so on. Clearly they are not independent as

$$\mathbb{P}(S_1 = S_0 + 1, S_2 = S_0 - 2) = 0 \neq \mathbb{P}(S_1 = S_0 + 1)\mathbb{P}(S_2 = S_0 - 2) = \frac{1}{8}.$$

More intuitively, if we know the value of S_s for some $s < t$, then it gives us information about S_n.

There are many natural questions that one might want to answer about $\{S_n\}_{t=1}^{\infty}$ such as:

1. What is the distribution of S_n for fixed $n \in \mathbb{N}$?
2. What is the distribution of S_n given $S_0, S_1, \dots, S_{n-1}$?

(3) What is the distribution of the time it takes to go bankrupt? Namely, what is the distribution of the random variable

$$\tau_0 \doteq \inf\{t \geq 0 : S_n = 0\}?$$

For instance, is there any chance that we never go bankrupt, i.e., what is $\mathbb{P}(\tau_0 = \infty)$?

The class of random variables to which $\{S_n\}$ belongs is known as a random walk.

D.1. **Introduction to Random Walks.**

Definition D.2. *Consider a sequence $\{X_n\}_{n\in\mathbb{N}}$ of i.i.d. random variables. Let S_0 be another random variable that is independent of $\{X_n\}$, and define*

$$S_n \doteq S_0 + \sum_{i=1}^{n} X_i, \;\; n \in \mathbb{N}.$$

The sequence $\{S_n\} \doteq \{S_n\}_{n\in\mathbb{N}_0}$ is known as a ***random walk*** *starting at S_0.*

If $X_n \in \{-1, 1\}$ for each $n \in \mathbb{N}$ and

$$\mathbb{P}(X_n = 1) = p, \qquad \mathbb{P}(X_n = -1) = q \doteq 1 - p, \;\; n \in \mathbb{N},$$

then $\{S_n\}$ is a ***simple random walk****. If $p = q = \frac{1}{2}$, then $\{S_n\}$ is* ***symmetric****.*

The following example illustrates some important points about simple random walks.

Example D.3. *Let $\{S_n\}$ be a simple random walk with $S_0 \doteq 0$. Compute the following:*

(1) $\mathbb{P}(S_2 = 2)$
(2) $\mathbb{P}(S_2 = 1)$
(3) $\mathbb{P}(S_3 = -1)$
(4) $\mathbb{P}(S_{100} - S_{98} = 2)$
(5) $\mathbb{P}(S_5 1 = -1)$

We complete the calculations below.

(1) $\mathbb{P}(S_2 = 2) = \mathbb{P}(X_1 = 1, X_2 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1) = p^2$.
(2) $\mathbb{P}(S_2 = 1) = 0$.
(3) Using the fact that $\{X_n\}$ are independent, we have

$$\begin{aligned}\mathbb{P}(S_3 = -1) = \mathbb{P}(X_1 = -1, X_2 = 1, X_3 = -1) + \mathbb{P}(X_1 = 1, X_2 = -1, X_3 = -1)\\ + \mathbb{P}(X_1 = -1, X_2 = -1, X_3 = 1)\\ = 3pq^2\end{aligned}$$

(4) Note that

$$\begin{aligned}\mathbb{P}(S_{100} - S_{98} = 2) &= \mathbb{P}\left(\left[S_0 + \sum_{i=1}^{100} X_i\right] - \left[S_0 + \sum_{i=1}^{98} X_i\right] = 2\right)\\ &= \mathbb{P}(X_{99} + X_{100} = 2)\\ &= p^2.\end{aligned}$$

(5) Begin by noting that $\mathbb{P}(S_{51} = -1) = \mathbb{P}\left(\sum_{i=1}^{53} X_i = -1\right)$. There are 51 terms in the summation, and each of those terms can either be 1 or -1. For the sum to add up to -1, that there must be exactly one more term that -1 than there are terms that are 1; this means that 26 of the X_n's must be -1 and 25 of the X_n's must be 1, this question is the same as determining the

probability that if you have 51 *experiments that are successful with probability* p*, what is the probability that exactly* 25 *of them are successes. From introductory statistics, we know that is the same as calculating* $\mathbb{P}(Y = 25)$*, where* $Y \sim$ *Binomial*$(51, p)$*. This is given by*

$$\mathbb{P}(Y = 25) = \binom{51}{25} p^2 5 q^{26},$$

so it follows that

$$\mathbb{P}(S_{53} = -1) = \binom{51}{25} p^2 5 q^{26}.$$

The following result gives the distribution of a simple random walk at each time instant.

Proposition D.4. *Let* $\{S_n\}$ *be a simple random walk. Then for each* $n \in \mathbb{N}$ *we have*

$$\mathbb{P}(S_n - S_0 = x) = \begin{cases} \binom{n}{\frac{1}{2}(n+x)} p^{\frac{1}{2}(n+x)} q^{\frac{1}{2}(n-x)}, & \textit{if } n + x \textit{ is even and } |x| \le n \\ 0, & \textit{otherwise} \end{cases}$$

Proof. Note that for each $n \in \mathbb{N}$, $S_n - S_{n-1} = X_n \in \{-1, 1\}$, so $\{S_n\}$ increases or decreases by 1 at each time instant.

We begin by considering the case when $n + x$ is even and $|x| \le n$. Suppose that $S_n - S_0 = x$, and let u denote the number of times the process increased, and d denotes the number of time it decreased. Then $u - d = x$ and $u + d = n$, so it follows that $u = \frac{1}{2}(n + x)$ and $d = \frac{1}{2}(n - x)$. Since $n + x$ is even, so is $n - x$, which ensures that u and d are both integers. Additionally, both u and d are nonnegative, since $|x| \le n$. The number of ways that we can choose the u increases from the total of n steps is given by $\binom{n}{u} = \binom{n}{\frac{1}{2}(n+x)}$. Note that by choosing the u increases, we also choose the d decreases, and that the probability of any particular configuration of u increases and d decreases is $p^u q^d$. Therefore,

$$\mathbb{P}(S_n - S_0 = x) = \binom{n}{\frac{1}{2}(n + x)} p^{\frac{1}{2}(n+x)} q^{\frac{1}{2}(n-x)}.$$

If $|x| > n$, then we would have $|S_n - S_0| = |x| = \left| \sum_{i=1}^{n} X_i \right| > n$, which is impossible since $X_i \in \{-1, 1\}$. Finally, if $|x| \le n$ and $n + x$ is odd, then we would have $u = \frac{1}{2}(n + x) \notin \mathbb{N}$, which is a contradiction, as the number of increases (and decreases) must be an integer.

□

The following corollary reformulates the conclusion of the previous proposition.

Corollary D.5. *Let* $\{S_n\}$ *be a simple random walk. Then*

$$\frac{1}{2}(S_n - S_0 + n) \sim Bin(n, p),$$

and therefore $\mathbb{E}(S_n - S_0) = n(2p - 1)$*, and* $Var(S_n - S_0 + n) = 4np(1 - p)$*.*

As noted earlier, the realizations of a random walk at different time instants are not independent from one another and do not necessarily have the same distribution. For example, knowing that $S_n = x$ gives you information about S_{n+1}, namely that it will be equal to $x+1$ with probability p and $x-1$ with probability q. The following result says that the *increments* of a random walk are themselves random walks.

Proposition D.6. *Let $\{S_n\}$ be a simple random walk and fix $k \in \mathbb{N}$. Then the process $\{\tilde{S}_n\}$ defined as*

$$\tilde{S}_n \doteq S_{k+n} - S_k, \;\; n \in \mathbb{N}_0,$$

is a simple random walk starting at $\tilde{S}_0 = 0$. Additionally, $\tilde{S}_n$ is independent of S_k for all $n \in \mathbb{N}$.

Proof. The proof follows immediately on noting that $\tilde{S}_0 = S_{k+0} - S_k = 0$ and that

$$\tilde{S}_n = \sum_{i=k+1}^{n} X_i.$$

Since the random variables $\{X_{k+i}\}_{i=1}^{\infty}$ are independent of $\{X_j\}_{j=1}^{k}$, the independence follows. □

As a consequence, we see that the simple random walk $\{S_n\}$ has stationary and independent increments.

Definition D.7. *A stochastic process $\{S_n\}$ has* ***independent increments*** *if for all $k \in \mathbb{N}_0$ and $0 \le n_1 < n_2 < \cdots < n_k$, the random variables*

$$\left\{S_{n_i} - S_{n_{i-1}}\right\}_{i=1}^{k} = \left\{S_{n_1} - S_0, S_{n_2} - S_{n_1}, \ldots, S_{n_k} - S_{n_{k-1}}\right\},$$

are independent of each other.

The process $\{S_n\}$ has ***stationary increments*** *if for all $0 \le n_1 < n_2 < n_3 < n_4$ such that $n_2 - n_1 = n_4 - n_3$, the random variables $S_{n_2} - S_{n_1}$ and $S_{n_4} - S_{n_3}$ have the same distribution.*

Note that if $\{S_n\}$ has stationary and independent increments, then for all $0 \le n_1 < n_2$, the random variables S_{n_1} and S_{n_2} may not be independent, but the random variables $S_{n_2} - S_{n_1}$ and $S_{n_1} - S_0$ are independent. Additionally, for each $k \in \mathbb{N}$, $S_{n+k} - S_n \overset{d}{=} S_k - S_0$.

Remark D.8. *Going forward we will write*

$$\mathbb{P}_x(\cdot) \doteq \mathbb{P}(\cdot | S_0 = x), \qquad \mathbb{E}_x(\cdot) \doteq \mathbb{E}(\cdot | S_0 = x),$$

to denote the conditional probability/expectation given that the stochastic process starts at state x.

Note that if $\{S_n\}$ has stationary and independent increments, if $S_0 = 0$, then

$$\mathbb{P}_0(S_n = x) = \mathbb{P}(S_n = x | S_0 = 0) = \mathbb{P}(S_n - S_0 = x).$$

Sometimes it will be interesting to consider random walks where the initial value S_0 is random.

Remark D.9. *Let μ be a probability measure on $\mathbb{Z}$. We write $\mathbb{P}_\mu$ to denote the probability measure under which $S_0 \sim \mu$; this yields*

$$\mathbb{P}_\mu(\cdot) \doteq \sum_{x \in \mathbb{Z}} \mathbb{P}(\cdot | S_0 = x)\mu(x), \qquad \mathbb{E}_\mu(\cdot) \doteq \sum_{x \in \mathbb{Z}} \mathbb{E}(\cdot | S_0 = x)\mu(x),$$

Example D.10. *Suppose that $\mu \sim Unif\{-1,0,1\}$, and that $\mathbb{P}(S_{n+1}-S_n=1)=\frac{4}{5}$. Then*

$$\mathbb{P}_\mu(S_1=1)=\mathbb{P}_{-1}(S_1=1)\cdot\frac{1}{3}+\mathbb{P}_0(S_1=1)\cdot\frac{1}{3}+\mathbb{P}_1(S_1=1)\cdot\frac{1}{3}=\frac{4}{5}\cdot\frac{1}{3}=\frac{4}{15}.$$

The following proposition tells us that random walks satisfy the Markov property. It says that if we are interested in the probability distribution of a random walk at the n-th time instant, knowing the random walk's value at the $n-1$-th time instant gives us just as much information as knowing its values at all prior time instants.

Proposition D.11. *Let $\{S_n\}$ be a random walk starting at $S_0=s_0\in\mathbb{R}$. Denote the i.i.d. increments of the walk by*

$$X_n \doteq S_n-S_{n-1},\ \ n\in\mathbb{N},$$

so that

$$S_n=S_0+\sum_{i=1}^{n}X_i.$$

Then, for each $n\in\mathbb{N}$ and $s,s_1,\dots,s_{n-1}\in\mathbb{R}$, we have

$$\mathbb{P}_{s_0}\left[S_n=s|S_0=s_0,S_1=s_1,\dots,S_{n-1}=s_{n-1}\right]=\mathbb{P}_{s_0}\left[S_n=s|S_{n-1}=s_{n-1}\right]=\mathbb{P}_{s_{n-1}}\left[S_1=s\right].$$

Proof. By definition of the probability measure $\mathbb{P}_s$, the result holds when $n=1$, so we suppose that it holds for some $n\in\mathbb{N}$. Then

$$\begin{aligned}\mathbb{P}_{s_0}\left[S_n=s|S_0=s_0,S_1=s_1,\dots,S_n=s_n\right]&=\frac{\mathbb{P}_{s_0}\left[S_n=s,S_0=s_0,S_1=s_1,\dots,S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_0=s_0,S_1=s_1,\dots,S_n=s_n\right]}\\&=\frac{\mathbb{P}_{s_0}\left[X_n=s-s_{n-1},S_0=s_0,S_1=s_1,\dots,S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_0=s_0,S_1=s_1,\dots,S_{n-1}=s_{n-1}\right]}\\&=\frac{\mathbb{P}_{s_0}\left[X_n=s-s_{n-1}\right]\mathbb{P}_{s_0}\left[S_0=s_0,S_1=s_1,\dots,S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_0=s_0,S_1=s_1,\dots,S_{n-1}=s_{n-1}\right]}\\&=\mathbb{P}_{s_0}\left[X_n=s-s_{n-1}\right].\end{aligned}$$

Additionally,

$$\begin{aligned}\mathbb{P}_{s_0}\left[S_n=s|S_{n-1}=s_{n-1}\right]&=\frac{\mathbb{P}_{s_0}\left[S_n=s,S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_{n-1}=s_{n-1}\right]}\\&=\frac{\mathbb{P}_{s_0}\left[S_n-S_{n-1}=s-s_{n-1},S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_{n-1}=s_{n-1}\right]}\\&=\frac{\mathbb{P}_{s_0}\left[X_n=s-s_{n-1},S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_{n-1}=s_{n-1}\right]}\\&=\frac{\mathbb{P}_{s_0}\left[X_n=s-s_{n-1}\right]\mathbb{P}_{s_0}\left[S_{n-1}=s_{n-1}\right]}{\mathbb{P}_{s_0}\left[S_{n-1}=s_{n-1}\right]}\\&=\mathbb{P}_{s_0}\left[X_n=s-s_{n-1}\right].\end{aligned}$$

Finally, note that since $\{X_n\}$ are i.i.d. and independent of S_0, we have

$$\begin{aligned}\mathbb{P}_{s_0}\left[X_n=s-s_{n-1}\right]&=\mathbb{P}_{s_{n-1}}\left[X_n=s-s_{n-1}\right]\\&=\mathbb{P}_{s_{n-1}}\left[X_1=s-s_{n-1}\right]\\&=\mathbb{P}\left[X_1=s-s_{n-1}|S_0=s_{n-1}\right]\\&=\mathbb{P}\left[S_0+X_1=s|S_0=s_{n-1}\right]\\&=\mathbb{P}_{s_{n-1}}\left[S_1=s\right].\end{aligned}$$

□

Using the previous proposition, we can show the following corollary.

Corollary D.12. *Let $\{S_n\}$ be a random walk starting at $S_0 = s_0 \in \mathbb{R}$. Then, for each $n \in \mathbb{N}$ and $k \in \mathbb{N}$, for all $s, s_1, \dots, s_{n-1} \in \mathbb{R}$, we have*

$$\mathbb{P}_{s_0}[S_{n-1+k} = s | S_0 = s_0, S_1 = s_1, \dots, S_{n-1} = s_{n-1}] = \mathbb{P}_{s_0}[S_{n-1+k} = s | S_{n-1} = s_{n-1}] = \mathbb{P}_{s_{n-1}}[S_k = s].$$

The following example illustrates how we can calculate the probability of observing a random walk with particular properties.

Example D.13. *Let $\{S_n\}_{n=0}^{\infty}$ be a simple random walk with*

$$p \doteq \mathbb{P}(S_n - S_{n-1} = 1), \ \ n \in \mathbb{N}.$$

Calculate

$$\mathbb{P}_0[S_3 \geq 1, S_6 = 2, S_{11} = -1]$$

We have

$$\begin{aligned}\mathbb{P}_0[S_3 \geq 1, S_6 = 2, S_{11} = -1] &= \mathbb{P}_0[S_{11} = -1 | S_3 \geq 1, S_6 = 2]\mathbb{P}[S_3 \geq 1, S_6 = 2] \\ &= \mathbb{P}_0[S_{11} = -1 | S_3 \geq 1, S_6 = 2]\left(\mathbb{P}_0[S_3 = 1, S_6 = 2] + \mathbb{P}_0[S_3 = 3, S_6 = 2]\right) \\ &= \mathbb{P}_0[S_{11} = -1 | S_3 \geq 1, S_6 = 2]\left(\mathbb{P}_0[S_6 = 2 | S_3 = 1]\mathbb{P}_0[S_3 = 1] + \mathbb{P}_0[S_6 = 2 | S_3 = 3]\mathbb{P}_0[S_3 = 3]\right)\end{aligned}$$

Proposition D.11 *ensures that*

$$\begin{aligned}\mathbb{P}_0[S_{11} = -1 | S_3 \geq 1, S_6 = 2] &= \mathbb{P}_0[S_{11} = -1 | S_6 = 2] \\ &= \mathbb{P}_0[S_{11} - S_6 = -3 | S_6 = 2] \\ &= \mathbb{P}_0[S_5 = -3] \\ &= \binom{5}{\frac{1}{2}(5 + (-3))} p^{\frac{1}{2}(5+(-3))}(1-p)^{\frac{1}{2}(5-(-3))} \\ &= 5p(1-p)^4.\end{aligned}$$

Similarly,

$$\mathbb{P}_0[S_6 = 2 | S_3 = 1] = \mathbb{P}_0[S_3 = 1] = \binom{3}{\frac{1}{2}(3+1)} p^{\frac{1}{2}(3+1)}(1-p)^{\frac{1}{2}(3-1)} = 3p^2(1-p),$$

and

$$\mathbb{P}_0[S_6 = 2 | S_3 = 3] = \binom{3}{\frac{1}{2}(3 + (-1))} p^{\frac{1}{2}(3+(-1))}(1-p)^{\frac{1}{2}(3-(-1))} = 3p(1-p)^2.$$

Finally,

$$\mathbb{P}_0[S_3 = 1] = \binom{3}{\frac{1}{2}(3+1)} p^{\frac{1}{2}(3+1)}(1-p)^{\frac{1}{2}(3-1)} = 3p^2(1-p)$$

and

$$\mathbb{P}_0[S_3 = 3] = p^3,$$

which yields

$$\mathbb{P}_0[S_3 \geq 1, S_6 = 2, S_{11} = -1] = 5p(1-p)^4\left(3p^2(1-p) \cdot 3p^2(1-p) + 3p(1-p)^2 \cdot p^3\right) = 60p^5(1-p)^6.$$

We now consider the problem of understanding the probability distribution of the amount of time that it takes for a random walk to end up in a particular state.

D.2. **Hitting Times of Random Walks.** Suppose that $\{S_n\}$ models someone's wealth each day. Then we might be interested in the probability that their wealth eventually exceeds a certain threshold, or the probability that they eventually run out of wealth completely (and how long we should expect each of things to take).

Definition D.14. *For $x \in \mathbb{Z} \doteq \{0, 1, -1, 2, -2, \dots\}$, the **first passage time/hitting time** of level x is the random variable τ_x defined as*

$$\tau_x \doteq \inf\{n > 0 : S_n = x\}.$$

Note that τ_x is a discrete random variable with values in $\mathbb{N}$.

We would like to determine the probability distribution of τ_x for a simple random walk. In order to do so, we begin by proving the reflection principle.

Theorem D.15. *(Reflection principle) Let $\{S_n\}$ be a simple random walk. Then for $x, y \in \mathbb{N}$, define the following quantities:*

- *Let $N_{x,y}(n)$ denote the number of paths from x to y in n steps*
- *Let $N^0_{x,y}(n)$ denote the number of paths from x to y in n steps that hit 0*
- *Let $N^{\neq 0}_{x,y}(n)$ denote the number of paths from x to y in n steps that do not hit 0 at time $i = 1, \dots, n$*

A path that is n steps long is a collection of points $\{a_0, a_1, \dots, a_n\}$ such that $a_i - a_{i-1} \in \{-1, 1\}$ for each $i \in \{1, \dots, n\}$.

*The **reflection principle** says that*

$$N^0_{x,y}(n) = N_{-x,y}(n),$$

and therefore that

$$N^{\neq 0}(x, y) = N_{x,y}(n) - N^0_{x,y}(n) = N_{x,y}(n) - N_{-x,y}(n).$$

Proof. To convince ourselves that $N^0_{x,y}(n) = N_{-x,y}(n)$, it suffices to draw a picture. The second claim is clearly true. □

The following lemma will be used in the proof of Theorem D.17.

Lemma D.16. *For each $x > 0$,*

$$N^{\neq 0}_{0,x}(n) = \frac{x}{n} N_{0,x}(n).$$

Proof. Let $u_{y,x}(n)$ denote the number of steps up that are required to go up to go from y to x in n steps, and let $d_{y,x}(n)$ denote the number of steps down that are required to go up to go from y to x in n steps. Note that that $u_{y,x}(n) - d_{y,x}(n) = x - y$ and $u_{y,x}(n) + d_{y,x}(n) = n$, so $u_{y,x}(n) = \frac{1}{2}(n + x - y)$ and $d_{y,x}(n) =$

$\frac{1}{2}(n+y-x)$, and that $N_{y,x}(n) = \binom{n}{u_{y,x}(n)}$, so it follows from Theorem D.15 that

$$\begin{aligned}
N_{0,x}^{\neq 0}(n) &= N_{1,x}^{\neq 0}(n-1) \\
&= N_{1,x}(n-1) - N_{-1,x}(n-1) \\
&= \binom{n-1}{u_{1,x}(n-1)} - \binom{n-1}{u_{-1,x}(n-1)} \\
&= \binom{n-1}{\frac{1}{2}(n-x-2)} - \binom{n-1}{\frac{1}{2}(n+x)} \\
&= \binom{n-1}{\frac{1}{2}(n-x)-1} - \binom{n-1}{\frac{1}{2}(n+x)} \\
&= \binom{n-1}{u_{0,x}(n)-1} - \binom{n-1}{u_{0,x}(n)}.
\end{aligned} \tag{80}$$

Finally, observe that

$$\begin{aligned}
\binom{n-1}{u_{0,x}(n)-1} - \binom{n-1}{u_{0,x}(n)} &= \frac{(n-1)!}{(n-1-(u_{0,x}(n)-1))!(u_{0,x}(n)-1)!} - \frac{(n-1)!}{(n-1-u_{0,x}(n))!u_{0,x}(n)!} \\
&= \frac{(n-1)!}{(n-u_{0,x}(n))!(u_{0,x}(n)-1)!} - \frac{(n-1)!}{(n-1-u_{0,x}(n))!u_{0,x}(n)!} \\
&= \frac{(n-1)!u_{0,x}(n)}{(n-u_{0,x}(n))!u_{0,x}(n)!} - \frac{(n-1)!(n-u_{0,x}(n))}{(n-u_{0,x}(n))!u_{0,x}(n)!} \\
&= (u_{0,x}(n)-(n-u_{0,x}(n)))\left(\frac{(n-1)!}{(n-u_{0,x}(n))!u_{0,x}(n)!}\right) \\
&= \frac{u_{0,x}(n)-d_{0,x}(n)}{n}\left(\frac{n!}{(n-u_{0,x}(n))!u_{0,x}(n)!}\right) \\
&= \frac{u_{0,x}(n)-d_{0,x}(n)}{n}\binom{n}{u_{0,x}(n)} \\
&= \frac{x}{n}\binom{n}{u_{0,x}(n)}.
\end{aligned} \tag{81}$$

Recalling that $N_{0,x}(n) = \binom{n}{u_{0,x}(n)}$ and combining (80) and (81), we obtain

$$N_{0,x}^{\neq 0}(n) = \frac{x}{n}N_{0,x}(n).$$

□

The following theorem characterizes the distribution of the hitting times of level x for a simple random walk starting at 0.

Theorem D.17. *Let $\{S_n\}$ be a simple random walk. Then for each $n \in \mathbb{N}$, if $|x| \le n$ and $n+x$ is even, we have*

$$\mathbb{P}_0(\tau_x = n) = \frac{|x|}{n}\mathbb{P}_0(S_n = x) = \frac{|x|}{n}\binom{n}{\frac{1}{2}(n+x)} p^{\frac{1}{2}(n+x)} q^{\frac{1}{2}(n-x)}.$$

Proof. If $S_0 = 0$, then $S_n = \sum_{i=1}^{n} X_i$, where the $\{X_i\}$ are i.i.d. Since the $\{X_i\}$ are i.i.d., it follows that for each $n \in \mathbb{N}$,

$$(X_1, X_2, \dots, X_n) \overset{d}{=} (X_n, X_{n-1}, \dots, X_1)$$[6],

which tells us that

$$(S_1, S_2, \dots, S_n) \overset{d}{=} (X_n, X_n + X_{n-1}, \dots, X_n + X_{n-1} + \cdots + X_1).$$

Additionally,

$$(X_n, X_n + X_{n-1}, \dots, X_n + X_{n-1} + \cdots + X_1) = (S_n - S_{n-1}, S_n - S_{n-2}, \dots, S_n),$$

so

$$(S_1, S_2, \dots, S_n) \overset{d}{=} (S_n - S_{n-1}, S_n - S_{n-2}, \dots, S_n). \tag{82}$$

Now, without loss of generality, suppose that $x \geq 0$ and that $x \leq n$[7]. Then

$$\begin{aligned}
\mathbb{P}_0(\tau_x = n) &= \mathbb{P}_0(S_1 < x, S_2 < x, \dots, S_{n-1} < x, S_n = x) \\
&= \mathbb{P}_0(S_n > S_1, S_n > S_2, \dots, S_n > S_{n-1}, S_n = x) \\
&= \mathbb{P}(S_n - S_1 > 0, S_n - S_2 > 0, \dots, S_n - S_{n-1} > 0, S_n = x).
\end{aligned}$$

Now apply (82) to see that

$$\begin{aligned}
\mathbb{P}_0(\tau_x = n) &= \mathbb{P}(S_n - S_1 > 0, S_n - S_2 > 0, \dots, S_n - S_{n-1} > 0, S_n = x) \\
&= \mathbb{P}(S_{n-1} > 0, S_{n-2} > 0, \dots, S_1 > 0, S_n = x).
\end{aligned}$$

Using Lemma D.16, we see that

$$\mathbb{P}(S_{n-1} > 0, S_{n-2} > 0, \dots, S_1 > 0, S_n = x) = N_{0,x}^{\neq 0}(n) p^{u_{0,x}(n)} q^{d_{0,x}(n)} = \frac{x}{n} N_{0,x}(n) p^{u_{0,x}(n)} q^{d_{0,x}(n)},$$

so if we recall that $u_{0,x}(n) = \frac{1}{2}(n+x)$ and $d_{0,x}(n) = \frac{1}{2}(n-x)$, then it follows that

$$\mathbb{P}_0(\tau_x = n) = \frac{x}{n}\binom{n}{u_{0,x}(n)} p^{u_{0,x}(n)} q^{d_{0,x}(n)} = \frac{x}{n}\binom{n}{\frac{1}{2}(n+x)} p^{\frac{1}{2}(n+x)} q^{\frac{1}{2}(n-x)},$$

as claimed. The proof in the case when $x < 0$ is similar and is omitted. □

We apply Theorem D.17 in the example below.

Example D.18. *A group of people is waiting in line at a store. Each minute either someone exists the line (either because they were served or because they were tired of waiting) or someone joins the line. Someone joins with a probability of*0.4 *and someone leaves with a probability of*0.6.

Suppose the line starts out with 10 *people in it. Calculate the probability that it takes at most an hour for the line to have no one in it.*

Let S_n *denote the number of people in the line at time* n. *Then* $\{S_n\}$ *is a simple random walk with*

$$\mathbb{P}(S_n - S_{n-1} = 1) = 0.4, \quad \mathbb{P}(S_n - S_{n-1} = -1) = 0.6.$$

We want to calculate $\mathbb{P}_{10}(\tau_0 \leq 60)$. *Observe that for each* $k \in \mathbb{N}$,

$$\mathbb{P}_{10}(\tau_0 = k) = \mathbb{P}_0(\tau_{-10} = k),$$

[6] I.e., for all $x_1, x_2, \dots, x_n \in \{-1, 1\}$, $\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \mathbb{P}(X_n = x_1, X_{n-1} = x_2, \dots, X_1 = x_n)$.

[7] If $x > n$, then the probability is clearly 0.

so we have

$$\begin{aligned}\mathbb{P}_{10}(\tau_0 \leq 60) &= \sum_{k=1}^{60} \mathbb{P}_{10}(\tau_0 = k)\\ &= \sum_{k=1}^{60} \mathbb{P}_0(\tau_{-10} = k)\\ &= \sum_{k=10}^{60} \mathbb{P}_0(\tau_{-10} = k),\end{aligned}$$

since the probability is 0 *whenever* $k < |-10|$*. Additionally, the probability of each term above is* 0 *when* $-10+k$ *is odd, namely whenever* k *is odd, which, when combined with Theorem* D.17*, yields*

$$\mathbb{P}_{10}(\tau_0 \leq 60) = \sum_{k=5}^{30} \mathbb{P}_0(\tau_{-10} = 2k) = \sum_{k=5}^{30} \frac{|-10|}{2k}\binom{2k}{\frac{1}{2}(2k+(-10))}(0.4)^{\frac{1}{2}(2k+(-10))}(0.6)^{\frac{1}{2}(2k-(-10))} \approx 0.7356.$$

D.3. **The Gambler's Ruin Problem.** The Gambler's ruin problem is a classical problem in probability theory - it is concerned with the probability that one can earn a particular amount of money (through betting on rounds of some game of change) before going bankrupt (or vice versa).

To make the problem more concrete, suppose that we have \$10 in our pockets, and we need \$80. We plan to try to convert our \$10 to \$80 by betting on roulette; we use the 'safest' roulette strategy, which is betting on 'red' each round. This strategy is relatively 'safe', as the probability that we win each round is $p \doteq \frac{18}{37}$, which is close to 50%. We stop playing whenever our fortune reaches \$80, or when it drops to \$0, whichever happens first. Additionally, we pick a constant stake of \$$c$ to bet each round. If the ball lands on 'red', then we get \$$2c$, and if the ball lands on 'black' or 'green', we don't get anything.

Let us begin by analyzing our expected winnings in each round. Let W denote our net winnings from betting \$c on a single round, and let P denote the payout of that round. Then $W = P - c$, so

$$\mathbb{E}[W] = \mathbb{E}[P-c] = \mathbb{E}[P] - c = 2c \cdot \mathbb{P}(P = 2c) + 0 \cdot \mathbb{P}(P = 0) = 2c \cdot \frac{18}{37} + 0 \cdot \frac{19}{37} - c = -\frac{c}{37} < 0.$$

This says that if we bet \$$c$ on each round, then we expect to lose \$$\frac{c}{37}$, on average.

Note that each round is independent of all of the other rounds, so we can represent our total wealth as a random walk. In order to do so, $\{X_n\}_{n=1}^{\infty}$ be i.i.d. random variables such that

$$\mathbb{P}(X_n = c) = p, \qquad \mathbb{P}(X_n = -c) = 1 - p.$$

Then, our total wealth after n rounds is given by

$$\begin{cases} S_0 = 10 \\ S_n \doteq S_0 + \sum_{i=1}^{n} X_i, \ \ n \geq 1. \end{cases}$$

Then $\{S_n\}$ is a discrete-time stochastic process that describes our wealth after n rounds. If we introduce the hitting times

$$\tau_0 \doteq \inf\{n \geq 1 : S_n = 0\}, \qquad \tau_{80} \doteq \inf\{n \geq 1 : S_n = 80\},$$

then we are interested in calculating $\mathbb{P}(\tau_0 < \tau_{80})$, namely the probability that we go bankrupt before accumulating a total wealth of \$80. The following proposition illustrates how one can calculate this probability.

Proposition D.19. *Consider the gambler's ruin problem where our initial wealth is d dollars, our target wealth is D dollars, where we stake \$1 on each round, and where the probability of winning each round is p and the probability of losing is $q \doteq 1-p$. Denote our wealth at time instant n by $\{S_n\}$. If we let*

$$\tau_0 \doteq \inf\{n \geq 1 : S_n = 0\}, \qquad \tau_D \doteq \inf\{n \geq 1 : S_n = K\},$$

then the probability that we go bankrupt before accumulating D dollars is

$$\mathbb{P}_d(\tau_0 < \tau_D) = \begin{cases} \frac{\left(\frac{q}{p}\right)^d - \left(\frac{q}{p}\right)^D}{1-\left(\frac{q}{p}\right)^D}, & \text{if } p \neq q \\ 1 - \frac{d}{D}, & \text{if } p = q = \frac{1}{2}. \end{cases}$$

We will see the proof of this result later in the course. However, we can apply it to see that in the setting described above where $d = 10$, $D = 80$, $p = \frac{18}{37}$, and $q = \frac{19}{37}$, we have

$$\mathbb{P}(\tau_0 < \tau_D) = \frac{\left(\frac{19}{18}\right)^{10} - \left(\frac{19}{18}\right)^{80}}{1 - \left(\frac{19}{18}\right)^{80}} = 0.9903858,$$

so it is overwhelmingly likely that we would go bankrupt before reaching \$80.

D.4. **Maxima of Random Walks.** Below we give another formulation of the reflection principle (see D.15).

Theorem D.20. *(Reflection principle 2) For $x, y \in \mathbb{N}$, let*

$$\mathscr{P}^-_{x,y}(n) \doteq \text{ the number of paths from 0 to } x-y \text{ in } n \text{ steps that visit } x \text{ on the way,}$$

and

$$\mathscr{P}^+_{x,y}(n) \doteq \text{ the number of paths from } 0 \text{ to } x+y \text{ in } n \text{ steps.}$$

Then

$$\mathscr{P}^-_{x,y}(n) = \mathscr{P}^+_{x,y}(n) = \mathscr{P}_{x,y}(n) \doteq \binom{n}{\frac{1}{2}(n+x+y)}$$

Below we study the maximum of a random walk. This measures the largest value that a random walk hits in a given before a given time. For a random walk $\{S_n\}$, its **maximum** is the stochastic process $\{M_n\}$ defined as

$$M_n \doteq \max_{1 \leq i \leq n} S_i, \quad n \in \mathbb{N}_0.$$

Theorem D.21. *Let $\{S_n\}$ be a symmetric random walk, so that*

$$\mathbb{P}_0[S_n - S_{n-1} = 1] = \mathbb{P}_0[S_n - S_{n-1} = -1] = \frac{1}{2}.$$

For each x, denote the hitting time of level x by

$$\tau_x \doteq \min\{n \in \mathbb{N} : S_n = x\}.$$

Then, for each $x, y > 0$,

$$\mathbb{P}_0[M_n \geq x, S_n = x-y] = \mathbb{P}_0[\tau_x \leq n, S_n = x-y] = \mathbb{P}_0[S_n = x+y].$$

In particular,

$$\mathbb{P}_0[M_n \geq x] = \mathbb{P}_0[S_n \geq x] + \mathbb{P}_0[S_n > x].$$

Proof. We begin by observing that $M_n \geq x$ if and only if $\tau_x \leq n$. Therefore, for $x, y > 0$,

$$\mathbb{P}_0[M_n \geq x, S_n = x - y] = \mathbb{P}_0[\tau_x \leq n, S_n = x - y],$$

so it follows from this, Theorem D.20, and the fact that $p = \frac{1}{2}$ that

$$\mathbb{P}_0[M_n \geq x, S_n = x - y] = \mathbb{P}_0[\tau_x \leq n, S_n = x - y] = \mathbb{P}_0[S_n = x + y]. \tag{83}$$

Additionally, if $x > 0$, and $y \leq 0$, then $x - y \leq x$, so

$$\mathbb{P}_0[M_n \geq x, S_n = x - y] = \mathbb{P}_0[S_n = x - y] \tag{84}$$

Together, the law of total probability, (83), and (84) tell us that

$$\begin{aligned}\mathbb{P}_0[M_n \geq x] &= \sum_{y=-\infty}^{\infty} \mathbb{P}_0[M_n \geq x, S_n = x - y] \\ &= \sum_{y=-\infty}^{0} \mathbb{P}_0[M_n \geq x, S_n = x - y] + \sum_{y=1}^{\infty} \mathbb{P}_0[M_n \geq x, S_n = x - y] \\ &= \sum_{y=-\infty}^{0} \mathbb{P}_0[S_n = x - y] + \sum_{y=1}^{\infty} \mathbb{P}_0[S_n = x + y] \\ &= \mathbb{P}_0[S_n \geq x] + \mathbb{P}_0[S_n > x].\end{aligned}$$

□

The previous result allows us to calculate probabilities involving the maximum of a simple symmetric random walk.

Example D.22. *Let $\{S_n\}$ be a simple symmetric random walk and let $M_n \doteq \max_{0 \leq i \leq n} S_i$. Calculate $\mathbb{P}_0[M_8 = 4]$.*

We have

$$\begin{aligned}\mathbb{P}_0[M_8 = 4] &= \mathbb{P}_0[M_8 \geq 4] - \mathbb{P}_0[M_8 \geq 3] \\ &= \mathbb{P}_0[S_8 \geq 4] + \mathbb{P}_0[S_8 > 4] + \mathbb{P}_0[S_8 \geq 3] + \mathbb{P}_0[S_8 > 3] \\ &= 2\mathbb{P}_0[S_8 \geq 4] + \mathbb{P}_0[S_8 > 4] + \mathbb{P}_0[S_8 \geq 3].\end{aligned}$$

Using Proposition D.4*, we see that for each x,*

$$\mathbb{P}_0[S_n = x] = \begin{cases} \binom{n}{\frac{1}{2}(n+x)} p^{\frac{1}{2}(n+x)} q^{\frac{1}{2}(n-x)}, & \textit{if } n + x \textit{ is even and } |x| \leq n \\ 0, & \textit{otherwise} \end{cases}$$

so we can (tediously) calculate each of the terms above.

D.5. **Further Questions.** Suppose that $\{S_n\}$ is a simple random walk. If $\{S_n\}$ models a quantity such as an investor's wealth, then we might be interested in understanding, in the long term, how $\{S_n\}$ behaves. While we will revisit this question later, the following points provide a great deal of insight.

First, recall that $\mathbb{E}_0[S_n] = n(2p - 1)$, so $\frac{S_n}{n}$ is the empirical average of n i.i.d. random variables with mean $2p - 1$, which ensures that, as $n \to \infty$,

$$\frac{S_n}{n} \xrightarrow{a.s.} 2p - 1.$$

If $p > \frac{1}{2}$, then $2p - 1 > 0$, so this tells us that S_n diverges to ∞ as $n \to \infty$. Similarly, if $p < \frac{1}{2}$, then $2p - 1 < 0$, which ensures that S_n diverges to $-\infty$ as $n \to \infty$. This means that if $p \neq \frac{1}{2}$, then, with probability 1, there is some point after which the random walk will never return to 0. However, if $p = \frac{1}{2}$, then $2p - 1 = 0$, so it is unclear what will happen.

Appendix E. Markov Chains

In statistics we often consider situations in which our data $\{X_n\}_{n=0}^{\infty}$ are i.i.d. samples from a population. However, in many situations it is not realistic to assume that the X_i are independent or that they have the same distribution. For instance, suppose that X_i measures the value of a stock on day i. Then, it is not reasonable to assume that X_{i-1} and X_i are independent; for example, if a stock was highly valuable yesterday, then it will probably also be highly valuable today. Additionally, in any realistic model, the distribution of the stock price should change over time as well. For instance, it might make sense to assume that the volatility (variability) of a stock will increase with its value.

However, the assumption that our data is i.i.d. is extremely useful, as it allows us to derive limit theorems such as the law of large numbers and the central limit theorem. If we make too few assumptions about the structure of our data, then analogous results might fail to hold, and so we might not be able to say anything useful about $\{X_n\}_{n=0}^{\infty}$, for example. Accordingly, it is natural to ask what a reasonable alternative to the i.i.d. assumption might be. It turns out that in many settings, a useful balance is given by the assumption that 'the future, given the present, is independent of the past'. In other words, we assume that as long as we know what happens today, then knowing what happened on past days doesn't give us any additional information about what will happen in the future.

Mathematically, for any event $A \subseteq \mathbb{R}$,

$$\mathbb{P}(X_{n+1} \in A | X_1, X_2, \ldots, X_n) = \mathbb{P}(X_{n+1} \in A | X_n).$$

E.1. **Introduction to Markov Chains.**

In this course, we will focus primarily on processes that have a discrete state space $\mathscr{S}$. In this setting, this property is formulated as follows:

Definition E.1. *Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of random variables taking values in a discrete set $\mathscr{S}$. We say that $\{X_n\}_{n=0}^{\infty}$ is a Markov chain if for all $n \in \mathbb{N}_0$, and all $x_0, x_1, \ldots, x_{n-1}, x, y \in \mathscr{S}$, we have*

$$\mathbb{P}[X_{n+1} = y | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x] = \mathbb{P}[X_{n+1} = y | X_n = x] \tag{85}$$

*A process satisfying (85) for all $n \in \mathbb{N}_0$, and all $x_0, x_1, \ldots, x_{n-1}, x_n, x, y \in \mathscr{S}$ is said to be a **time-homogeneous Markov chain**. If the stronger condition that for all $n \in \mathbb{N}_0$ and all $x_0, \ldots, x_{n-1}, x, y \in \mathscr{S}$ we have*

$$\mathbb{P}[X_{n+1} = y | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_n, X_n = x] = \mathbb{P}[X_1 = y | X_0 = x],$$

*holds, then the process $\{X_n\}_{n=0}^{\infty}$ is known as a **time-homogeneous Markov chain**. In this case, the matrix $P \doteq [P_{x,y}]_{x,y \in \mathscr{S}}$ with entries*

$$P_{x,y} \doteq \mathbb{P}[X_1 = y | X_0 = x], \quad x, y \in \mathscr{S},$$

*is known as the **transition matrix** (or **transition kernel**) of $\{X_n\}$.*

It is convenient to consider time-homogeneous Markov chains for many reasons. For example, if we hope to extract some statistical information from a system, it doesn't matter when we start observing it. More practically, their analysis is much more straightforward.

The following example illustrates how one might use Markov chains to model physical systems, such as weather.

Example E.2. *Each day if it is sunny, there is a 95% chance that it will be sunny the next day, a 3% chance that it will be cloudy the next day, and a 2% chance that it will be rainy the next day. Similarly, if it is cloudy, there is a 80% chance the next day is sunny, a 15% chance the next day will be cloudy, and a 5% chance the next day is rainy. Finally, if it is rainy, there is a 75% chance the next day is sunny, a 15% chance the next day will be cloudy, and a 10% chance the next day is rainy.*

We can model this system as a Markov chain. Its state space is $\{S,C,R\}$, and its transition matrix is given by

$$P = \begin{pmatrix} 0.95 & 0.03 & 0.02 \\ 0.8 & 0.15 & 0.05 \\ 0.75 & 0.15 & 0.10. \end{pmatrix}$$

If it is sunny today, what is the probability it will be sunny in two days?

Let $\{X_n\}$ denote the weather today. Then Markov's property ensures that

$$\begin{aligned}\mathbb{P}(X_2 = S|X_0 = S) &= \mathbb{P}(X_2 = S|X_0 = S, X_1 = S)\mathbb{P}(X_1 = S|X_0 = S) + \mathbb{P}(X_2 = S|X_0 = S, X_1 = C)\mathbb{P}(X_1 = C|X_0 = S) \\ &\quad + \mathbb{P}(X_2 = S|X_0 = S, X_1 = R)\mathbb{P}(X_1 = R|X_0 = S) \\ &= \mathbb{P}(X_2 = S|X_1 = S)\mathbb{P}(X_1 = S|X_0 = S) + \mathbb{P}(X_2 = S|X_1 = C)\mathbb{P}(X_1 = C|X_0 = S) \\ &\quad + \mathbb{P}(X_2 = S|X_1 = R)\mathbb{P}(X_1 = R|X_0 = S) \\ &= P_{S,S}P_{S,S} + P_{C,S}P_{S,C} + P_{R,S}P_{S,R} \\ &= 0.95 \cdot 0.95 + 0.8 \cdot 0.03 + 0.02 \cdot 0.75 = 0.9415.\end{aligned}$$

There are other interesting questions that we may want to answer too. For example, in the long run, what proportion of the time will it be sunny? Namely, what is

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} 1_S(X_i)?$$

Note that the law of large numbers is not applicable here, since $\{X_i\}$ are not i.i.d. A related question is, does this proportion depend on the weather on day 1?

For a more interesting example, we recall a homework problem from earlier in the course.

Example E.3. *Flip a fair coin repeatedly. How many flips do you expect it to take before you get 2 heads in a row?*

We consider a state space $\mathscr{S} = \{HT, TH, TT, HH\}$. How could we model this question using a Markov chain on $\mathscr{S}$?

Consider the transition matrix on $\mathscr{S}$ given by

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5, \end{pmatrix}$$

and let $\{X_n\}$ describe our two most recent flips. With $\tau_{HH} \doteq \inf\{n \geq 0 : X_n = HH\}$, we would like to calculate $\mathbb{E}_{TT}(\tau_{HH})$.

This will require first step analysis, which we will see soon.

We could also model this question using $\mathscr{S} = \{0,1,2\}$ *and*

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 1 \end{pmatrix},$$

and in this case X_n *would track the number of heads in a row, and would stop once it reached 2.*

Note that if the state space is infinite, then the transition matrix P is infinite.

Example E.4. *Let* $\{S_n\}$ *be a simple random walk with* $p = \mathbb{P}(S_{n+1} - S_n = 1)$ *and* $q \doteq 1 - p$. *Then* $\mathscr{S} = \{0,1,-1,2,-2,\dots\}$, *and*

$$P_{x,y} = \mathbb{P}(S_{n+1} = y | S_n = x) = \begin{cases} p, & y = x+1 \\ q, & y = x-1 \\ 0, & \textit{otherwise} \end{cases}$$

Definition E.5. *A square matrix P is said to be a* ***stochastic matrix*** *if*

(1) $P_{x,y} \in [0,1]$ *for all* $x, y \in \mathscr{S}$.
(2) *For each* $x \in \mathscr{S}$, $\sum_{y \in \mathscr{S}} P_{x,y} = 1$.

The rows of P are ***probability vectors****, namely row x satisfies*

$$P_{x,\cdot} = \mathbb{P}[X_1 \in \cdot | X_0 = x].$$

It is often helpful to consider a Markov chain with a random initial distribution μ. *In this case, for a probability measure* $\mu = (\mu_1, \mu_2, \dots) \in \mathbb{R}^{|\mathscr{S}|}$, *we write*

$$\mathbb{P}_\mu(\cdot) = \mathbb{P}(\cdot | X_0\ \mu),$$

to denote the probability measure under which $X_0 \sim \mu$. *Observe that*

$$\mathbb{P}_\mu(X_0 = x) = \mu(x).$$

Markov chains are easy to simulate, and have found widespread use in statistics, optimization, physics, and other fields. To simulate a Markov chain, one implements the following algorithm using a random number generator.

To simulate a Markov chain with transition matrix P and initial distribution μ for n steps, do the following.

Algorithm E.6. *Input:* P, μ, n.

```
(1) Generate $X_0$ according to $\mu$.
(2) For $i \in \{1,\dots,n\}$:
      • Given $X_{i-1} = x$, let $p \doteq P_{x,\cdot}$ (the x-th row of P).
      • Generate $X_i$ according to $P_{x,\cdot}$ (i.e., sample a discrete random variable with state space
        $\mathscr{S}$ according to the probability distribution $P_{x,\cdot}$).
Output: $X_0, X_1, \dots, X_n$.
```

In general, if we want to understand the long-term behavior of the chain, it will be helpful to identify and calculate the marginal distribution of X_n. More precisely, we want to study $\mathbb{P}_\mu(X_n = x)$ for initial distribution μ, time $n \in \mathbb{N}$, and $x \in \mathscr{S}$.

The following theorem illustrates how to calculate the marginal distribution of a Markov chain started in a particular state.

Proposition E.7. *For each $n \in \mathbb{N}$, the matrix $P^n = P \times P \times \cdots \times P$ is stochastic. Furthermore, for each $n \in \mathbb{N}$,*

$$\mathbb{P}_x(X_n = y) = \mathbb{P}[X_n = y | X_0 = x] = (P^n)_{x,y}.$$

Proof. Let P and Q be stochastic matrices on $\mathscr{S}$. Then, for each $x \in \mathscr{S}$,

$$\begin{aligned}\sum_{y\in\mathscr{S}} (PQ)_{x,y} &= \sum_{y\in\mathscr{S}} \sum_{z\in\mathscr{S}} P_{x,z} Q_{z,y} \\ &= \sum_{z\in\mathscr{S}} \sum_{y\in\mathscr{S}} P_{x,z} Q_{z,y} \\ &= \sum_{z\in\mathscr{S}} P_{x,z} \sum_{y\in\mathscr{S}} Q_{z,y} \\ &= \sum_{z\in\mathscr{S}} P_{x,z} \cdot 1 \\ &= 1,\end{aligned}$$

so PQ is stochastic. It follows that P^2 is stochastic, and the general result follows on inducting on n. Finally, note that, using the time-homogeneity of the process and then Markov's property,

$$\begin{aligned}(P^2)_{x,y} &= \sum_{z\in\mathscr{S}} P_{x,z} P_{z,y} \\ &= \sum_{z\in\mathscr{S}} \mathbb{P}[X_1 = z | X_0 = x] \mathbb{P}[X_1 = y | X_0 = z] \\ &= \sum_{z\in\mathscr{S}} \mathbb{P}[X_1 = z | X_0 = x] \mathbb{P}[X_2 = y | X_1 = z] \\ &= \sum_{z\in\mathscr{S}} \mathbb{P}[X_1 = z | X_0 = x] \mathbb{P}[X_2 = y | X_0 = x, X_1 = z] \\ &= \sum_{z\in\mathscr{S}} \mathbb{P}[X_2 = y, X_1 = z | X_0 = x] \\ &= \mathbb{P}[X_2 = y | X_0 = x].\end{aligned}$$

The proof of the result for general $n \in \mathbb{N}$ is similar. □

The following corollary is an immediate consequence of Proposition E.7.

Corollary E.8. *For any initial distribution μ,*

$$\mathbb{P}_\mu(X_n = x) = (\mu^T P^n)_x.$$

Proof. Let μ be a probability distribution on $\mathscr{S}$, so that μ is a $|\mathscr{S}| \times 1$ vector, and P^n is an $|\mathscr{S}| \times |\mathscr{S}|$ matrix. Using Proposition E.7, the x entry of $\mu^T P^n$ is given by

$$(\mu^T P^n)_x = \sum_{z\in\mathscr{S}} \mu_y (P^n)_{y,x} = \sum_{z\in\mathscr{S}} \mu_y \mathbb{P}[X_n = x | X_0 = y] = \mathbb{P}_\mu[X_n = x].$$

□

We now consider several important examples of Markov chains.

Example E.9. *(2-allele Wright Fisher model) Consider a population with a constant size of N individuals. Each gene has two alleles (types); a and A, and each person's gene has two alleles (either aa, aA, or AA). For simplicity, we call the alleles* $1, 2,$ *and* 3.

Let $X_n = (X_n^1, X_n^2, X_n^3)$ *denote the number of individuals with each type of alleles in the system in generation* n.

The dynamics of the system are defined as follows; given that the state of the population at generation $n-1$ *was* $X_n = (\alpha_1, \alpha_2, \alpha_3)$, *where* $\alpha_1 + \alpha_2 + \alpha_3 = N$, *the probability that each of the N individuals in generation has an allele of type* $i \in \{1,2,3\}$ *is* $\frac{\alpha_i}{N}$.

Interpretation: *each of the N individuals in generation* $n+1$ *chooses a parent from generation* n *uniformly at random, then inherits the allele from that parent.*

This means that given that the state of the system in generation n *was* $\boldsymbol{\alpha} \doteq (\alpha_1, \alpha_2, \alpha_3)$, *we have, when* $\boldsymbol{\beta} \doteq (\beta_1, \beta_2, \beta_3)$ *and* $\beta_1 + \beta_2 + \beta_3 = N$,

$$\begin{aligned}\mathbb{P}[X_{n+1} = \boldsymbol{\beta} | X_n = \boldsymbol{\alpha}] &= \mathbb{P}[X_{n+1} = (\beta_1, \beta_2, \beta_3) | X_n = (\alpha_1, \alpha_2, \alpha_3)] \\ &\doteq \binom{N}{\beta_1\ \beta_2\ \beta_3} \left(\frac{\alpha_1}{N}\right)^{\beta_1} \left(\frac{\alpha_2}{N}\right)^{\beta_2} \left(\frac{\alpha_3}{N}\right)^{\beta_3} \\ &= \frac{N!}{\beta_1! \beta_2! \beta_3!} \left(\frac{\alpha_1}{N}\right)^{\beta_1} \left(\frac{\alpha_2}{N}\right)^{\beta_2} \left(\frac{\alpha_3}{N}\right)^{\beta_3}\end{aligned}$$

What do you expect will happen as $n \to \infty$? *Will the population ever reach a stable state or will it repeatedly change?*

The following Python code simulates a Wright-Fisher process:

```
def WrightFisherSim(N, x0, n):
    Xi = x0
    WrightFisherPath = numpy.array(x0)
    for i in range(1,n+1):
        prob = [Xi[0]/N, Xi[1]/N, Xi[2]/N]
        Xi = numpy.random.multinomial(N, prob, size=None)
        WrightFisherPath = numpy.append(WrightFisherPath, Xi)
    return(WrightFisherPath)
```

Below is another example of a Markov chain.

Example E.10. *(Reliability) Each day eduroam is either up, down, or being repaired. If eduroam is up at the start of the day, then it will be up at the start of the next day with probability 0.9, and it will be down at the start of the next day with probability 0.1. When eduroam breaks, it takes 3 days for the IT department to fix it.*

Let E_n denote the status of eduroam on day n. What is its state space, $\mathscr{S}$? And what is its transition matrix, P?

The chain has 5 *possible states; D (down at the start of day n), U (up at the start of day n, 1 (on the first day of repair after going down), 2 (on the second day of repair after going down), and 3 (on the third day of repair after going down).*

To model this system using a Markov chain, let $\mathscr{S} \doteq \{D, U, 1, 2, 3\}$, *and consider the transition matrix*

$$P \doteq \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0.1 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

E.2. **Chapman-Kolmogorov Equations.** We have seen that for a (time-homogeneous) Markov chain $\{X_n\}$, the marginal distribution of the chain at time n, given that it started at state x, is given by

$$\mathbb{P}[X_n = y | X_0 = x] = (P^n)_{x,y}.$$

Similarly, if the chain starts according to a probability distribution μ on $\mathscr{S}$, then

$$\mathbb{P}_\mu(X_n = y) = \mathbb{P}[X_n = y | X_0 \sim \mu] = (\mu^T P^n)_y.$$

The following proposition is known as the Chapman-Kolmogorov equations. The proof follows from time-homogeneity and Markov's property.

Proposition E.11. *For each $n, m \in \mathbb{N}_0$, for $x, y \in \mathscr{S}$, we have*

$$\begin{aligned} \mathbb{P}[X_{n+m} = y | X_0 = x] &= (P^{n+m})_{x,y} \\ &= \sum_{z \in \mathscr{S}} \mathbb{P}[X_m = z | X_0 = x] \mathbb{P}[X_{m+n} = y | X_m = z] \\ &= \sum_{z \in \mathscr{S}} \mathbb{P}[X_m = z | X_0 = x] \mathbb{P}[X_n = y | X_m = z] \\ &= \sum_{z \in \mathscr{S}} (P^m)_{x,z} (P^n)_{z,y}. \end{aligned}$$

Proof. This follows immediately on noting that $P^{n+m} = P^n P^m$ and applying Proposition E.7. □

Now, consider a collection of time instants $0 \le n_1 < n_2 < \cdots < n_k$, where $k \in \mathbb{N}$. Then for any collection of states $x_1, x_2, \ldots, x_k$, we have, for any probability measure μ on $\mathscr{S}$,

$$\mathbb{P}_\mu(X_{n_1} = x_1, X_{n_2} = x_2, \ldots, X_{n_k} = x_k) = (\mu^T P^{n_1})_{x_1} (P^{n_2 - n_1})_{x_1, x_2} \cdots (P^{n_k - n_{k-1}})_{x_{k-1}, x_k},$$

which shows that μ and P fully determine the joint distribution of the chain at different time instants. Consequently, for any $m < n$,

$$\mathbb{P}[X_{n+m} = y | X_0 = x_0, \ldots, X_{m-1} = x_{m-1}, X_m = x] = \mathbb{P}[X_{n+m} = y | X_m = x] = (P^n)_{x,y}.$$

We now consider an example of an Urn Process.

Example E.12. *An urn contains* 5 *marbles; some are red and some are blue. At each time instant, a marble is chosen uniformly at random and replaced by a new marble. The new marble is red with probability* p *and blue with probability* $1-p$. *Suppose that the urn starts with* 5 *red marbles. Find the probability that the* 3*-rd marble selected is red.*

Let X_n *be the number of red marbles in the urn after* n *draws. If we let* A *denote the event that the 3rd draw is red, then*

$$\mathbb{P}(A) = \mathbb{P}(A|X_2 = 5)\mathbb{P}(X_2 = 5) + \mathbb{P}(A|X_2 = 4)\mathbb{P}(X_2 = 4) + \mathbb{P}(A|X_2 = 3)\mathbb{P}(A|X_2 = 3)$$

We conclude this section with a final example.

Example E.13. *Let* $\{X_n\}$ *be the Markov chain with state space* $\mathscr{S} = \{1,2,3,4\}$ *and transition matrix*

$$P = \begin{pmatrix} 0 & \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{8} & \frac{7}{8} & 0 & 0 \\ 0 & 0 & \frac{4}{5} & \frac{1}{5} \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

with initial distribution $\mu^T = \left(\frac{8}{47}, \frac{21}{47}, 0, \frac{18}{47}\right)$. *Calculate the following:*

(1) $\mathbb{P}_\mu[X_5 = 3|X_3 = 2]$.
(2) $\mathbb{P}_\mu[X_3 = 2]$.
(3) $\mathbb{P}_\mu[X_8 = 1, X_6 = 2, X_3 = 2]$.

We have

$$\mathbb{P}_\mu[X_5 = 3|X_3 = 2] = (P^2)_{2,3} = 0.09375,$$

and

$$\mathbb{P}_\mu[X_3 = 2] = (\mu^T P^3)_2 = 0.35767121,$$

and

$$\mathbb{P}_\mu[X_8 = 1, X_6 = 2, X_3 = 2] = (\mu^T P^3)_2 (P^3)_{2,2} (P^2)_{2,1} = 0.01235210712919844.$$

The Python code below may be helpful:

```
P = numpy.array([[0,1/4,3/4,0],[1/8,7/8,0,0], [0,0,4/5, 1/5], [0,0,0,1]])
muT = numpy.array([8/47,21/47,0,18/47])
P2 = numpy.linalg.matrix_power(P, 2)
P3 = numpy.linalg.matrix_power(P, 3)
muTP3 = numpy.matmul(muT,P3)
```

E.3. **First Step Analysis and Hitting Times.** The following example illustrates the principle of first step analysis.

Example E.14. *A queue at a bank starts with no one in it. Each minute, there is a 40% chance someone joins the line and a 60% chance someone leaves it. Furthermore, if the line already has 4 people in it, then any arrivals simply walk away rather than waiting in line.*

If there is currently one person in the queue, how long, on average, do you expect it to take for the queue to be empty again?

Let $\{X_n\}$ *denote the Markov chain on* $\mathscr{S} = \{0,1,2,3,4\}$ *with transition matrix*

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 & 0 \\ 0 & 0.6 & 0 & 0.4 & 0 \\ 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0.6 & 0.4 \end{pmatrix}$$

and initial distribution of $X_0 = 1$. *For each* $x \in \mathscr{S}$, *let*

$$\tau_0 \doteq \inf\{t \geq 0 : X_0 = 0\},$$

denote the hitting time of state 0 *when the chains starts at state* x. *Then* $\mathbb{E}[\tau_0|X_0 = 1]$, *for example, is the expect number of minutes for the line to empty when it starts with 1 person in it.*

To calculate $\mathbb{E}[\tau_0]$, *we use the principle of first step analysis, which is based on the law of total expectation: if the chain starts at* 1, *then at time 1 it must be at either state* 0 *or state* 2, *so we have*

$$\begin{aligned} \mathbb{E}[\tau_0|X_0 = 1] &= \mathbb{E}[\tau_0|X_1 = 0, X_0 = 1]\mathbb{P}[X_1 = 0|X_0 = 1] + \mathbb{E}[\tau_0|X_1 = 2, X_0 = 1]\mathbb{P}[X_1 = 2|X_0 = 1] \\ &= [1 + \mathbb{E}[\tau_0|X_1 = 0]]\,\mathbb{P}[X_1 = 0|X_0 = 1] + [1 + \mathbb{E}[\tau_0|X_1 = 2]]\,\mathbb{P}[X_1 = 2|X_0 = 1] \\ &= (1 + \mathbb{E}[\tau_0|X_1 = 0])(0.6) + (1 + \mathbb{E}[\tau_0|X_1 = 2])(0.4). \end{aligned}$$

Using this same approach, we can see that

$$\begin{aligned} \mathbb{E}[\tau_0|X_0 = 1] &= (1 + m_0)(0.6) + (1 + m_2)(0.4) \\ \mathbb{E}[\tau_0|X_0 = 2] &= (1 + m_1)(0.6) + (1 + m_3)(0.4) \\ \mathbb{E}[\tau_0|X_0 = 3] &= (1 + m_2)(0.6) + (1 + m_4)(0.4) \\ \mathbb{E}[\tau_0|X_0 = 4] &= (1 + m_3)(0.6) + (1 + m_4)(0.4), \end{aligned}$$

where we let $m_x = \mathbb{E}[\tau_0|X_0 = x]$ *for* $x \in \mathscr{S}$, *and let* $m \doteq [m_0, m_1, \ldots, m_4]^T$, *then, noting that* $m_0 = 0$, *we can rewrite the system of equations as*

$$\begin{aligned} m_1 &= 1 + 0.6m_0 + 0.4m_2 = 1 + 0.4m_2 \\ m_2 &= 1 + 0.6m_1 + 0.4m_3 \\ m_3 &= 1 + 0.6m_2 + 0.4m_4 \\ m_4 &= 1 + 0.6m_3 + 0.4m_4, \end{aligned}$$

or, equivalently,

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.4 & 0 & 0 \\ -0.6 & 1 & -0.4 & 0 \\ 0 & -0.6 & 1 & -0.4 \\ 0 & 0 & -0.6 & 0.6 \end{pmatrix} m,$$

from which we obtain

$$m = \begin{pmatrix} 1 & -0.4 & 0 & 0 \\ -0.6 & 1 & -0.4 & 0 \\ 0 & -0.6 & 1 & -0.4 \\ 0 & 0 & -0.6 & 0.6 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 4.0123 \\ 7.5308 \\ 10.30 \\ 11.97 \end{pmatrix}.$$

Since $m_1 = 4.0123$, *this says that if the queue starts with a single person in it, then on average it will take 4.0123 minutes for the queue to be empty again.*

The following theorem illustrates this principle more generally.

Theorem E.15. *Consider a Markov chain $\{X_n\}$ on $\mathscr{S}$ with transition matrix P. For some $x_* \in \mathscr{S}$, for each $x \in \mathscr{S}$, let*

$$\tau_{x_*} = \inf\{n \geq 0 : X_n = x_*\}$$

Further, assume that

$$\mathbb{P}(\tau_x < \infty) = 1,$$

for all $x \in \mathscr{S}$ (otherwise we will have $\mathbb{E}[\tau_x] = \infty$). Let B denote the transition matrix P with column x_* and row x_* removed, namely, B is the $|\mathscr{S}| - 1 \times |\mathscr{S}| - 1$ matrix with

$$B = [B_{x,y}]_{x,y \in \mathscr{S} \setminus \{x_*\}}.$$

Let e be a column of ones, and define the vector m by $m_x = \mathbb{E}[\tau_{x_*} | X_0 = x]$. Then m is the smallest non-negative solution to

$$m = e + Bm.$$

Proof. We prove the result using first step analysis. Note that

$$\begin{aligned} m_x &= \mathbb{E}[\tau_{x_*} | X_0 = x] \\ &= \sum_{y \in \mathscr{S}} \mathbb{E}[\tau_{x_*} | X_1 = y, X_0 = x] \mathbb{P}[X_1 = y | X_0 = x] \\ &= \sum_{y \in \mathscr{S}} p_{x,y} \mathbb{E}[\tau_{x_*} | X_0 = x, X_1 = y], \end{aligned}$$

and for each $x \in \mathscr{S}$,

$$\mathbb{E}[\tau_{x_*} | X_0 = x, X_1 = x_*] = 1,$$

and for each $x \in \mathscr{S} \setminus \{x_*\}$, we have

$$\mathbb{E}[\tau_{x_*} | X_0 = x, X_1 = y] = 1 + \mathbb{E}[\tau_{x_*} | X_0 = y] = 1 + m_y.$$

Therefore, for each $x \in \mathscr{S} \setminus \{x_*\}$, we have

$$\begin{aligned} m_x &= \sum_{y \in \mathscr{S}} p_{x,y} \mathbb{E}[\tau_{x_*} | X_0 = x, X_1 = y] \\ &= p_{x,x_*} \cdot 1 + \sum_{y \in \mathscr{S} \setminus \{x_*\}} p_{x,y}(1 + m_y) \\ &= p_{x,x_*} + \sum_{y \in \mathscr{S} \setminus \{x_*\}} p_{x,y} 1 + \sum_{y \in \mathscr{S} \setminus \{x_*\}} p_{x,y} m_y \\ &= \sum_{y \in \mathscr{S}} p_{x,y} + \sum_{y \in \mathscr{S} \setminus \{x_*\}} p_{x,y} m_y \\ &= 1 + \sum_{y \in \mathscr{S} \setminus \{x_*\}} p_{x,y} m_y, \end{aligned}$$

as claimed. We omit the proof that m is the smallest non-negative solution. □

Let's consider another example.

Example E.16. *Consider a Markov chain on $\mathscr{S} = \{a, b, c, d\}$ with transition matrix*

$$P = \begin{pmatrix} 0 & 0.7 & 0.3 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 1. \end{pmatrix}$$

Define $\tau_d \doteq \inf\{n \geq 0 : X_n = d\}$ *and compute* $m_x \doteq \mathbb{E}[\tau_d | X_0 = x]$ *for each* $x \in \mathscr{S}$.

We have, with e denoting a vector of ones, and $m^T \doteq \begin{pmatrix} m_a & m_b & m_c \end{pmatrix}$ *that* $m = e + Bm$, *so*

$$m = (I-B)^{-1}e = \begin{pmatrix} 5.263158 \\ 3.631579 \\ 5.736842 \end{pmatrix}.$$

We consider another example on a finite state space.

Example E.17. *We roll a fair 6-sided die repeatedly. How many flips do we expect it to take until we observe the sequence* $1,1,1$*?*
Consider the Markov chain $\{X_n\}$ *on* $\mathscr{S} \doteq \{0,1,11,111\}$ *and transition matrix*

$$P = \begin{pmatrix} 5/6 & 1/6 & 0 & 0 \\ 5/6 & 0 & 1/6 & 0 \\ 5/6 & 0 & 0 & 1/6 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Define $\tau \doteq \inf\{n \geq 0 : X_n = 123\}$ *and for each* $x \in \mathscr{S}$, *let* $m_x = \mathbb{E}[\tau | X_0 = x]$. *Then if we let* $e^T = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$, *and* $m = \begin{pmatrix} m_0 & m_1 & m_{11} \end{pmatrix}$, *and*

$$B = \begin{pmatrix} 5/6 & 1/6 & 0 \\ 5/6 & 0 & 1/6 \\ 5/6 & 0 & 0 \end{pmatrix},$$

then

$$m = e + Bm,$$

so

$$m = (I-B)^{-1}e = \begin{pmatrix} 258 \\ 252 \\ 216 \end{pmatrix},$$

which tells us that we expect it to take 258 rolls, on average.

If we instead were interested in calculating the expected number of rolls to observe the sequence $1,2,3$, *then we could consider the Markov chain* $\{Y_n\}$ *on* $\mathscr{S}_Y \doteq \{0,1,12,123\}$ *with transition matrix*

$$P_Y = \begin{pmatrix} 5/6 & 1/6 & 0 & 0 \\ 4/6 & 1/6 & 1/6 & 0 \\ 4/6 & 1/6 & 0 & 1/6 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then we would have, with $m_{Y,x} \doteq \mathbb{E}[\tau_{Y,123} | Y_0 = x]$ *and*

$$\tau_{Y,123} \doteq \inf\{n \geq 0 : Y_n = 123\},$$

then

$$m_Y = (I - B_Y)^{-1}e = \begin{pmatrix} 216 \\ 210 \\ 180 \end{pmatrix},$$

where

$$B_Y = \begin{pmatrix} 5/6 & 1/6 & 0 \\ 4/6 & 1/6 & 1/6 \\ 4/6 & 1/6 & 0 \end{pmatrix}.$$

This tells us that we would expect it to take 216 *rolls, on average, to observe the sequence* $1,2,3$.

The next example involves calculating the expected first passage time of a simple random walk. It relies on some basic results from the theory of linear second-order difference equations. Accordingly, it may be helpful to refer to `https://en.wikipedia.org/wiki/Linear_difference_equation` and references therein.

Example E.18. *First step analysis reveals that, with* $m_i \doteq \mathbb{E}[\tau_0 | X_0 = i]$,

$$m_i = 1 + q m_{i-1} + p m_{i+1}, \quad i \in \mathbb{N}, \tag{86}$$

and $m_0 = 0$. *One can show that the solution to* (86) *must be of the form*

$$m_i = C_1 + C_2 \left(\frac{q}{p}\right)^i + \frac{i}{q-p}, \quad i \in \mathbb{N},$$

where C_1 *and* C_2 *are suitable constants. Thus, we need only to find the value of* C_1 *and* C_2. *To do so, note that for each* $i \in \mathbb{N}_0$,

$$\begin{aligned} m_{i+1} &= \mathbb{E}[\tau_0 | X_0 = i+1] \\ &= \mathbb{E}[\tau_0 - \tau_1 + \tau_1 | X_0 = i+1] \\ &= \mathbb{E}[\tau_0 - \tau_1 | X_0 = i+1] + \mathbb{E}[\tau_1 | X_0 = i+1] \\ &= \mathbb{E}[\tau_i | X_0 = i+1] + \mathbb{E}[\tau_0 | X_0 = i] \\ &= \mathbb{E}[\tau_0 | X_0 = 1] + \mathbb{E}[\tau_0 | X_0 = i] \\ &= m_1 + m_i. \end{aligned}$$

Similarly,

$$m_{i+1} = m_1 + (m_1 + m_{i-1}) = 2m_1 + m_{i-1}.$$

Inductively, this yields, for $i \in \mathbb{N}_0$,

$$m_{i+1} = m_0 + (i+1)m_1 = (i+1)m_1.$$

Thus

$$m_i = i m_1 = C_1 + C_2 \left(\frac{q}{p}\right)^i + \frac{i}{q-p}$$

so $C_1 = C_2 = 0$, *and therefore*

$$m_i = \frac{i}{q-p}.$$

The next example involves calculating the expected first passage time of a more complicated random walk, namely one whose transition probabilities depend on the current state. It is somewhat beyond the scope of this class, but may be interesting to read anyway.

Example E.19. *Let $\{S_n\}$ be a random walk with transition matrix*

$$\begin{cases} p_{i,i+1} = p_i, & i \in \mathbb{N} \\ p_{i,i-1} = q_i = 1 - p_i, & i \in \mathbb{N} \\ p_{0,0} = 1. \end{cases}$$

Note that the transition probabilities here depend on the current state of the chain, so it is different from the other simple random walks we have seen so far. For $i \in \mathbb{N}$, compute $\mathbb{E}[\tau_0 | X_0 = i]$, where

$$\tau_0 \doteq \inf\{n \geq 0 : X_n = 0\}.$$

Begin by defining the quantity

$$\begin{cases} \alpha_i \doteq \prod_{j=1}^{i} \frac{q_j}{p_j} & i \geq 1 \\ \alpha_0 = 1. \end{cases}$$

Additionally, assume that as $n \to \infty$,

$$\sum_{i=1}^{n} \alpha_i \to \infty. \tag{87}$$

First step analysis reveals that, with $m_i \doteq \mathbb{E}[\tau_0 | X_0 = i]$,

$$m_i = 1 + q_i m_{i-1} + p_i m_{i+1}, \quad i \in \mathbb{N}, \tag{88}$$

and $m_0 = 0$. If we define $\Delta_i \doteq m_i - m_{i-1}$, then (88) becomes

$$q_i \Delta_i = 1 + p_i \Delta_{i+1}, \quad i \in \mathbb{N}. \tag{89}$$

Observe, by taking $i = 1$ and combining (88) and (89), that

$$\Delta_1 = m_1 - m_0 = m_1, \tag{90}$$

and

$$\Delta_2 = \frac{q_1}{p_1}\Delta_1 - \frac{1}{p_1} = \frac{q_1}{p_1}(m_1 - m_0) - \frac{1}{p_1} = \alpha_1 m_1 - \frac{1}{p_1} \tag{91}$$

If we define

$$\begin{cases} b_i \doteq \sum_{j=1}^{i} \frac{1}{p_j \alpha_j} & i \in \mathbb{N} \\ b_0 \doteq 0, \end{cases}$$

then $b_1 = \frac{1}{p_1 \alpha_1}$, so the expressions in (90) and (91) can be rewritten as

$$\Delta_1 = m_1 - m_0 = \alpha_0 m_1 - \alpha_0 \beta_0$$

and

$$\Delta_2 = \alpha_1 m_1 - \alpha_1 b_1,$$

respectively. Iterating this procedure and observing that

$$\alpha_2 b_1 + \frac{1}{p_2} = \alpha_2 \frac{1}{p_1 \alpha_1} + \frac{1}{p_2} = \alpha_2 \left(\frac{1}{p_1 \alpha_1} + \frac{1}{\alpha_2 p_2} \right) = \alpha_2 b_2,$$

we can calculate

$$\begin{aligned}\Delta_3 &= \frac{q_2}{p_2}\Delta_2 - \frac{1}{p_2}\\ &= \frac{q_2}{p_2}(\alpha_1 m_1 - \alpha_1 b_1) - \frac{1}{p_2}\\ &= \alpha_2 m_1 - \left(\alpha_2 b_1 + \frac{1}{p_2}\right)\\ &= \alpha_2 m_1 - \alpha_2 b_2.\end{aligned}$$

In general, we can repeat this argument recursively to show that for each $i \in \mathbb{N}$,

$$\Delta_{i+1} = \alpha_i m_1 - \alpha_i b_i. \tag{92}$$

Note that

$$\sum_{i=0}^{n} \Delta_{i+1} = \sum_{i=0}^{n}(m_{i+1} - m_i) = \sum_{i=0}^{n} m_{i+1} - \sum_{i=0}^{n} m_i = \sum_{i=0}^{n} m_{i+1} - \sum_{i=0}^{n-1} m_{i+1} = m_{n+1}$$

*so we can apply (*92*) to see that*

$$m_{n+1} = \sum_{i=1}^{n}(\alpha_i m_1 - \alpha_i b_i) = m_1 \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i b_i. \tag{93}$$

Recall from Theorem E.15 *that m is the smallest non-negative solution to (*93*), so, using the fact that we have written m_{n+1} in terms of m_1 for all $n \geq 0$, it suffices to find the smallest non-negative m_1 such that for all $n \in \mathbb{N}$,*

$$m_1 \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i b_i \geq 0.$$

Rewriting this expression, we will require, for all $n \geq 0$, that

$$m_1 \geq \frac{\sum\limits_{i=1}^{n} \alpha_i}{\sum\limits_{i=1}^{n} \alpha_i b_i}.$$

If we define, for each $n \geq 1$,

$$f(n) \doteq \frac{\sum\limits_{i=1}^{n} \alpha_i}{\sum\limits_{i=1}^{n} \alpha_i b_i},$$

*then, under assumption (*87*), one can show that $\{f(n)\}$ is non-increasing and*

$$f(n) \downarrow \sum_{i=1}^{\infty} \frac{1}{p_i \alpha_i} \geq 0.$$

It follows that

$$m_1 = \sum_{i=1}^{\infty} \frac{1}{p_i \alpha_i}.$$

*Combining this with (*93*), one sees that for $i \geq 2$,*

$$m_i = \sum_{j=0}^{i-1} \alpha_j \left(\sum_{k=j+1}^{\infty} \frac{1}{p_k \alpha_k} \right).$$

Note that depending on the values of p_i, q_i, we may have that $m_1 = \infty$ or that $m_1 < \infty$.

E.4. **Limiting Distributions.** Suppose that we are interested in understanding the long-term behavior of a Markov chain $\{X_n\}$ on $\mathscr{S}$ with transition matrix P. That is to say, suppose we would like to analyze

$$\lim_{n\to\infty} \mathbb{P}_\mu[X_n = x], \quad x \in \mathscr{S},$$

where μ is some distribution on $\mathscr{S}$. In the previous section we saw that this is equivalent to analyzing

$$\lim_{n\to\infty} (\mu^T P^n)_x, \quad x \in \mathscr{S}.$$

There are many natural questions one might have about such a limit, for example,

1. When does the limit exist (i.e., when does the sequence $p_n \doteq (\mu^T P^n)_x$ converge)?
2. Does it depend on the initial distribution μ?
3. How can we compute this limit? What if we don't have a computer or calculator?
4. If the limit of the sequence of probabilities exists, how quickly do they converge to the limit?

There are many other questions related to the long-term behavior of Markov chains as well. For example,

1. If the chain runs for a very long time, what proportion of the time will it spend in each state $x \in \mathscr{S}$, on average.
2. Can we compute the long term expected value of a function of the chain? For example, if there is a cost associated with the chain being in each state, can we estimate the average cost the chain will incur over the long term?
3. And more.

Definition E.20. *Let $\{X_n\}$ be a Markov chain on $\mathscr{S}$ with transition matrix P. We say that a probability distribution α on $\mathscr{S}$ is a **limiting distribution** for $\{X_n\}$ if for any initial distribution μ on $\mathscr{S}$ and all $x \in \mathscr{S}$,*

$$\lim_{n\to\infty} \mathbb{P}_\mu[X_n = x] = \alpha_x \doteq \alpha(\{x\}).$$

Equivalently, α is a limiting distribution if for each initial distribution μ on $\mathscr{S}$ we have

$$\lim_{n\to\infty} \mu^T P^n = \alpha,$$

meaning that for each $x \in \mathscr{S}$,

$$\lim_{n\to\infty} (\mu^T P^n)_x = \alpha_x.$$

Addiitonally, if α is a limiting distribution for $\{X_n\}$, then

$$\lim_{n\to\infty} P^n = \mathscr{A},$$

where $\mathscr{A}$ is a matrix in which every row is the vector α.

Note that any limiting distribution α for $\{X_n\}$ is independent of the initial distribution μ, which tells us that the limiting distribution of a Markov chain is unique. The following proposition tells us that the limiting distribution of a Markov chain describes the average proportion of time that the Markov chain spends in each state. Before we state the proposition, we recall a result regarding the convergence of Cesàro means.

Proposition E.21. *(Cesàro means) Let $\{a_n\}$ be a convergent sequence with limit a. For each $n \in \mathbb{N}$, let*

$$x_n \doteq \frac{1}{n}\sum_{i=1}^{n} a_i,$$

so that x_n is the mean of $a_1, a_2, \dots, a_n$. Then $\{x_n\}$ converges to a.

Proof. Fix $\epsilon > 0$. Since $a_n \to a$, there is some $n_1 \in \mathbb{N}$ such that if $n \geq n_1$, then $|a_n - a| < \frac{\epsilon}{2}$. Note that, for all $n \geq n_1$,

$$\begin{aligned} |x_n - a| &= \left| \frac{1}{n} \sum_{i=1}^{n} a_i - a \right| \\ &= \frac{1}{n} \left| \sum_{i=1}^{n} (a_i - a) \right| \\ &\leq \frac{1}{n} \left| \sum_{i=1}^{n_1 - 1} (a_i - a) \right| + \frac{1}{n} \left| \sum_{i=n_1}^{n} (a_i - a) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n_1-1} |a_i - a| + \frac{1}{n} \sum_{i=n_1}^{n} |a_i - a| \\ &\leq \frac{n_1}{n} \max_{1 \leq i \leq n_1 - 1} |a_i - a| + \max_{n_1 \leq i \leq n} |a_i - a| \end{aligned} \tag{94}$$

Let $n_2 > \max\left\{ \frac{2n_1 \max_{1 \leq i \leq n_1 - 1} |a_i - a|}{\epsilon}, n_1 \right\}$ and observe that for all $n \geq n_2$, we have

$$\frac{n_1}{n} \max_{1 \leq i \leq n_1 - 1} |a_i - a| < \frac{\epsilon}{2}, \quad \max_{n_1 \leq i \leq n} |a_i - a| < \frac{\epsilon}{2}. \tag{95}$$

Combining (94) and (95), we see that for all $n \geq n_2$, $|x_n - a| < \epsilon$, and therefore that $x_n \to a$ as $n \to \infty$. □

Recall that we write, for an initial distribution μ, $\mathbb{E}_\mu(\cdot) \doteq \mathbb{E}[\cdot | X_0 \sim \mu]$.

Proposition E.22. *Let $\{X_n\}$ be a Markov chain on $\mathscr{S}$. Suppose that $\{X_n\}$ has a limiting distribution α. Then for each initial distribution μ and $x \in \mathscr{S}$,*

$$\alpha_x = \lim_{n \to \infty} \mathbb{E}_\mu \left[\frac{1}{n+1} \sum_{k=0}^{n} 1_{\{X_k = x\}} \right].$$

Proof. Let μ be an initial distribution for $\{X_n\}$ and denote the transition matrix of $\{X_n\}$ by P. Then for each $x \in \mathscr{S}$, recalling that for each $k \in \mathbb{N}$,

$$(\mu^T P^k)_x = \mathbb{P}_\mu[X_k = x],$$

we have

$$\begin{aligned} \mathbb{E}_\mu \left[\frac{1}{n+1} \sum_{k=0}^{n} 1_{\{X_k = x\}} \right] &= \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{E}_\mu \left[1_{\{X_k = x\}} \right] \\ &= \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}_\mu \left[X_k = x \right] \\ &= \frac{1}{n+1} \sum_{k=0}^{n} (\mu^T P^k)_x. \end{aligned}$$

Since α is a limiting distribution, we know that for each $x \in \mathscr{S}$

$$\lim_{n \to \infty} (\mu^T P^n)_x = \alpha_x,$$

so Proposition E.21 ensures that as $n \to \infty$,

$$\frac{1}{n+1} \sum_{k=0}^{n} (\mu^T P^k)_x \to \alpha_x,$$

as well. The result follows. □

The following example allows us to explicitly calculate the limiting distribution of a Markov chain.

Example E.23. *Consider the Markov chain $\{X_n\}$ on $\mathscr{S} = \{1,2\}$ with transition matrix*

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix},$$

where $a, b \in (0,1)$. On the homework we showed that

$$P^n = \frac{1}{a+b}\begin{pmatrix} b + a(1-a-b)^n & a(1-(1-a-b)^n) \\ b(1-(1-a-b)^n) & a + b(1-a-b)^n \end{pmatrix}.$$

Note that, since $a, b \in (0,1)$, we have $|1-a-b| < 1$, and therefore as $n \to \infty$, $(1-a-b)^n \to 0$. Thus, as $n \to \infty$, $P^n \to \mathscr{A}$, where

$$\mathscr{A} = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}.$$

Therefore, the limiting distribution of $\{X_n\}$ is

$$\alpha^T = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}.$$

Using Proposition E.22*, we see that this means that on average, in the long term, the Markov chain will spend $\frac{b}{a+b}$ of its time in state* 1 *and $\frac{a}{a+b}$ in state 2.*

We have seen that limiting distributions allow us to understand, in some sense, the long-term behavior of Markov chains. But many questions remain, such as :

(1) How do I know if a Markov chain has a limiting distribution?
(2) If a Markov chain has a limiting distribution, then it can be expressed by calculating the limit as $n \to \infty$ of $\mu^T P^n$, for an initial distribution μ. Is there a simpler way to calculate the limiting distribution?

E.5. **Stationary Distributions.** We now introduce the notion of a stationary distribution, which will be useful for answering these questions and more. Recall that we are dealing time time homogeneous Markov chains, namely Markov chains where for all $n \in \mathbb{N}$, $x, y \in \mathscr{S}$,

$$\mathbb{P}[X_{n+1} = y | X_n = x] = P_{x,y},$$

for some transition matrix P.

Definition E.24. *Let $\{X_n\}$ be a Markov chain on $\mathscr{S}$. We say that a probability distribution π on $\mathscr{S}$ is a* ***stationary distribution*** *for $\{X_n\}$ if $X_0 \sim \pi$ implies that $X_1 \sim \pi$.*

The following proposition provides several equivalent definitions of a stationary distribution of a Markov chain.

Proposition E.25. *For a Markov chain $\{X_n\}$ on $\mathscr{S}$ with transition matrix P, the following are equivalent:*

(1) π is a stationary distribution for $\{X_n\}$.
(2) If $X_0 \sim \pi$, then $X_1 \sim \pi$.
(3) If $X_0 \sim \pi$, then $X_n \sim \pi$ for all $n \geq 0$.
(4) For each $n \geq 0$, if $X_n \sim \pi$, then $X_{n+1} \sim \pi$.
(5) $\pi^T P = \pi$
(6) For each $n \geq 0$, $\pi^T P^n = \pi$.

(7) *For all* $x \in \mathscr{S}$, $\pi_x = \sum_{y\in\mathscr{S}} \pi_x P_{x,y}$.

The following observation explains the connection between stationary distributions and eigenvalues.

Observation E.26. *Consider a state space* $\mathscr{S}$ *and let* P *be a* $|\mathscr{S}| \times |\mathscr{S}|$ *transition matrix. Then a vector* $v \in \mathbb{R}^{|\mathscr{S}|}$ *is a stationary distribution for* P *if and only if it is a probability distribution* [a] *and* v *is a left-eigenvector for* P *with corresponding eigenvalue* 1.

[a] Recall that v is a probability distribution if $v_x \geq 0$ for all $x \in \mathscr{S}$ and $\|v\|_1 \doteq \sum_{x\in\mathscr{S}} |v_x| = 1$.

The following proposition tells us that limiting distributions are in fact stationary distributions.

Proposition E.27. *Let* P *be a transition matrix on* $\mathscr{S}$. *Suppose that there is some probability distribution* μ *on* $\mathscr{S}$ *such that*

$$\lim_{n\to\infty} \mu^T P^n = \pi^T,$$

for some probability distribution π. *Then* π *is a stationary distribution.*

Proof. Observe that

$$\pi^T P = \left[\lim_{n\to\infty} \mu^T P^n\right] P = \lim_{n\to\infty} \left[\mu^T P^n P\right] = \lim_{n\to\infty} \mu^T P^{n+1} = \pi,$$

so π is a stationary distribution. □

We can verify that Proposition E.27 holds directly in the context of Example E.23.

Example E.28. *Consider the Markov chain* $\{X_n\}$ *from Example* E.23. *Let*

$$\pi^T \doteq \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}.$$

We have seen that π *is the limiting distribution of* $\{X_n\}$, *but it is also the stationary distribution, as*

$$\pi^T P = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix} \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} = \begin{pmatrix} \frac{b(1-a)+ba}{a+b} & \frac{ab+a(1-b)}{a+b} \end{pmatrix} = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix} = \pi^T.$$

It is natural to ask whether the converse of Proposition E.27 is true as well. The following example illustrates this, and also shows that stationary distributions need not be unique.

Example E.29. *(Stationary distributions are not necessarily limiting distributions, and stationary distributions aren't always unique) Let* $\{X_n\}$ *be the Markov chain on* $\mathscr{S} = \{1,2\}$ *with transition matrix*

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

For each $\lambda \in [0,1]$, *let*

$$\pi_\lambda^T \doteq \begin{pmatrix} \lambda & 1-\lambda \end{pmatrix}.$$

Then

$$\pi_\lambda^T P = \pi_\lambda^T,$$

so π_λ is a stationary distribution for each $\lambda \in [0,1]$. However, we can easily see that $\{X_n\}$ does not have a limiting distribution, as limiting distributions are unique, and with

$$e_1^T \doteq \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad e_2^T \doteq \begin{pmatrix} 0 & 1 \end{pmatrix},$$

we have

$$\lim_{n\to\infty} (e_1^T P^n) = \lim_{n\to\infty} e_1 = e_1,$$

while

$$\lim_{n\to\infty} (e_2^T P^n) = \lim_{n\to\infty} e_2 = e_2,$$

Since different limiting behavior emerges for different initial distributions, we conclude that $\{X_n\}$ does not have a limiting distribution.

In light of Example E.5, we will now seek to answer the following questions:

(1) When does a Markov chain have a stationary distribution?
(2) When is the stationary distribution unique?
(3) When is the stationary distribution the limiting distribution?

Somehow the problem with the Markov chain in Example E.5 was that the state space consisted of two separate regions that the Markov chain could not move between. We now introduce some terminology so that we can more carefully describe the quantitative behavior of chains like that one.

E.6. **Recurrent and Transient States.** We begin by introducing the notions of *accessibility and communication* in a Markov chain. Throughout this section, we assume that $\{X_n\}$ is a Markov chain on some state space $\mathscr{S}$ with transition matrix P.

Definition E.30. *We say that a state y is accessible from a state x if there is some $n \in \mathbb{N}_0$ such that*

$$\mathbb{P}_x(X_n = y) = \mathbb{P}[X_n = y | X_0 = x] = (P^n)_{x,y} > 0,$$

and we write $x \to y$. Similarly, we say that x is accessible from y if there is some $m \in \mathbb{N}_0$ such that

$$\mathbb{P}_y(X_m = x) = \mathbb{P}[X_m = x | X_0 = y] = (P^m)_{y,x} > 0,$$

and we write $y \to x$. If $x \to y$ and $y \to x$, then we say that x and y communicate if, then we write $x \leftrightarrow y$.

Note that two states communicate if the Markov chain can travel between them. The following observation says that ↔ is an equivalence relation (see `https://en.wikipedia.org/wiki/Equivalence_relation`).

Observation E.31. *Communication is an equivalence relation, namely it satisfies:*

(1) (Reflexivity) $x \leftrightarrow x$.
(2) (Symmetry) if $x \leftrightarrow y$, then $y \leftrightarrow x$.
(3) (Transitivity) if $x \leftrightarrow y$ and $y \leftrightarrow z$, then $x \leftrightarrow z$.

If multiple states communicate with one another, we refer to them as belonging to the same communication class.

Definition E.32. *If $x \leftrightarrow y$, then x and y are said to belong to the same* ***communication class****. We write*

$$\mathscr{C}_x \doteq \{y : x \leftrightarrow y\}.$$

The following observation says that if x and y communicate, then they have the same communication class.

Observation E.33. *Suppose that $x \leftrightarrow y$. Then $\mathscr{C}_x = \mathscr{C}_y$.*

We are often interested in chains with a single communication class.

Definition E.34. *A Markov chain on $\mathscr{S}$ is said to be* ***irreducible*** *if it has a single communication class. That is, if for all states $x, y \in \mathscr{S}$, there is some $n \geq 0$ such that $(P^n)_{x,y} > 0$.*

The following example allows us to visualize transience and communication.

Example E.35. *Let $\{X_n\}$ be a Markov chain on $\mathscr{S} = \{a, b, c, d, e\}$ with transition matrix*

$$P = \begin{pmatrix} 0.5 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0.7 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.3 & 0.7 \end{pmatrix}$$

Then state a is transient, as $\mathbb{P}_a[R_a < \infty] = [R_a < \infty | X_0 = a] = 1/2$. However, states b, c, d, and e are not transient. This can be verified using first step analysis. Additionally, we have that $\mathscr{C}_a = \{a\}$, $\mathscr{C}_b = \mathscr{C}_C = \{b, c\}$, and $\mathscr{C}_d = \mathscr{C}_e = \{d, e\}$.

Since this chain has multiple communication classes, it is not irreducible.

We consider another example.

Example E.36. *Consider the Markov chain $\{X_n\}$ on $\mathscr{S} = \{0, 1, 2\}$ with transition matrix*

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

Note that $1 \leftrightarrow 2$, $2 \leftrightarrow 3$, and $1 \leftrightarrow 3$, so the chain is irreducible.

We have seen several examples of Markov chains where the chain can get stuck in one position forever. Such chain show up in many models, for example in ecology, biology, and chemical kinetics.

Definition E.37. *A state $x \in \mathscr{S}$ is* ***absorbing*** *if $P_{x,x} = 1$. A Markov chain with at least one absorbing state is called an* ***absorbing chain****.*

An absorbing state is an example of a closed communication class.

Definition E.38. *A set of states $\mathscr{C}$ is* ***closed*** *if for all $x \in \mathscr{S}$, if $x \to y$, then $y \in \mathscr{C}$. A set of states $\mathscr{C}$ is* ***open*** *if there is some $y \in \mathscr{C}^c$ such that $x \to y$.*

This means that a set of states is closed if no state outside of it is accessible from within it.

Example E.39. *Consider the Markov chain on $\mathscr{S} = \{1, 2, \dots, 7\}$ with transition matrix*

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

The communication classes are $\mathscr{C}_1 = \{1, 5\}$, $\mathscr{C}_2 = \{2\}$, $\mathscr{C}_3 = \{3, 4, 7\}$, and $\mathscr{C}_6 = \{6\}$. Note that $\mathscr{C}_1$ and $\mathscr{C}_3$ are closed, and $\mathscr{C}_2$ and $\mathscr{C}_6$ are open.

Closed communication classes are helpful because they allow us to break a Markov chain's state space down into smaller, irreducible pieces.

Proposition E.40. *Let P be a transition matrix on $\mathscr{S}$, and let $\mathscr{C} \subseteq \mathscr{S}$ be a communication class. Then $\mathscr{C}$ is closed if and only if either $\mathscr{C} = \mathscr{S}$ (i.e., the corresponding Markov chain is irreducible) or if $[P^m]_{x,y} = 0$ for all $x \in \mathscr{C}$ and $y \in \mathscr{S} \setminus \mathscr{C}$ for all $m \in \mathbb{N} = \{1, 2, \dots\}$.*

The following example illustrates how one should interpret Proposition E.40.

Example E.41. *Consider a Markov chain on $\mathscr{S} = \{0, 1, 2, 3, 4\}$ with transition matrix*

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

In this case there are two communication classes, $\mathscr{C}_0 = \{0, 1, 2\}$ and $\mathscr{C}_3 = \{3, 4\}$. Observe that $\mathscr{C}_0$ is open, as $2 \to 3$, and $\mathscr{C}_2$ is closed. The sub-matrices corresponding to $\mathscr{C}_0$ and $\mathscr{C}_3$ are given by

$$P_{0,1,2} = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{pmatrix}, \quad P_{3,4} = \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \end{pmatrix}$$

Observe that $P_{0,1,2}$ is not stochastic (i.e., is not the transition matrix of some Markov chain), while $P_{3,4}$ is stochastic (i.e., it is the transition matrix of some Markov chain).

We now introduce the notion of a **return time**.

Definition E.42. *For each $x \in \mathscr{S}$, let*

$$R_x \doteq \inf\{n > 0 : X_n = x\}$$

be the ***first passage time*** *of $\{X_n\}$ to state x. Note that $R_x \geq 1$ for each $x \in \mathscr{S}$. For each $x \in \mathscr{S}$, let*

$$f_x \doteq \mathbb{P}_x[R_x < \infty] = \mathbb{P}[R_x < \infty | X_0 = x],$$

denote the ***return probability*** *.*

- *We say that a state $x \in \mathscr{S}$ is* ***transient*** *if $f_x < 1$.*
- *We say that a state $x \in \mathscr{S}$ is* ***recurrent*** *if $f_x = 1$.*
- *A recurrent state x is said to be* ***positive recurrent*** *if $\mathbb{E}_x[R_x] < \infty$.*
- *A state that is* ***recurrent*** *but not* ***positive recurrent*** *is said to be* ***null recurrent****.*

Observe that a state is transient if there is some chance that the Markov chain never returns to it. Similarly, a state is recurrent if it the Markov chain always returns to the state (in a finite amount of time). Finally, a state is positive recurrent if the Markov chain, on average, returns to the state in a finite amount of time.

Example E.43. *Consider a Markov chain $\{X_n\}$ on $\mathscr{S} = \{1, 2, 3\}$ with transition matrix*

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

Note that states 1 and 2 are recurrent, but state 3 is transient.

It is not immediately clear that it is possible for a state to be null recurrent. It turns out that a finite state Markov chain has no null recurrent states.

Proposition E.44. *Let $\mathscr{S}$ be a finite state space, and let P be a transition matrix on $\mathscr{S}$. Then, for each $x \in \mathscr{S}$, either x is positive recurrent or x is transient. In particular, each $x \in \mathscr{S}$ cannot be null recurrent.*

Proof. We omit the proof for now but may return to it later. □

Below we present an example of a Markov chain for which each state is null-recurrent.

Example E.45. *Let $\{X_n\}$ be a simple symmetric random walk on $\mathscr{S} = \{0, 1, -1, 2, -2, \dots\}$. For each $x \in \mathscr{S}$, x is null recurrent.*

The following proposition gives us a helpful criteria to determine whether a state is transient or recurrent.

Proposition E.46. *Consider a Markov chain on $\mathscr{S}$ with transition matrix P. Then, a state $x \in \mathscr{S}$ is recurrent if and only if*

$$\sum_{n=1}^{\infty} (P^n)_{x,x} = \infty.$$

And $x \in \mathscr{S}$ is transient if and only if

$$\sum_{n=1}^{\infty}(P^n)_{x,x} = \frac{f_x}{1-f_x} < \infty,$$

where $f_x \doteq \mathbb{P}_x(R_x < \infty)$.

Proof. Observe that $(P^n)_{x,y} = \mathbb{P}_x[X_n = y] = \mathbb{E}_x\left[1_{\{X_n=y\}}\right]$, so we let

$$N_x \doteq \sum_{n=0}^{\infty} 1_{\{X_n=x\}},$$

denote the number of times that the chain visits state x. We begin by assuming that x is transient, and observing that $N_x \sim \text{Geometric}(1-f_x)$. To see this, think of each excursion of the Markov chain away from x. The chain returns to x with probability f_x, and each excursion is independent of the others, so the probability that an excursion is finite is f_x. And the probability that a given excursion is infinite (i.e., that the chain does not return to x) is $1-f_x$. Noting that

$$\begin{aligned}\{N_x = k\} &= \{\text{the chain visits } x \text{ exactly } k \text{ times then never returns}\}\\ &= \{\text{the first } k-1 \text{ excursions away from } x \text{ are finite, and the } k\text{-th is infinite}\},\end{aligned}$$

we see that

$$\mathbb{P}_x(N_x = k) = f_x^{k-1}(1-f_x),$$

so $N_x \sim \text{Geometric}(1-f_x)$. Therefore, $\mathbb{E}[N_x] = \frac{1}{1-f_x}$, and, using the fact that

$$(P^n)_{x,x} = \mathbb{P}_x[X_n = x] = \mathbb{E}_x[1_{\{X_n=x\}}],$$

we have

$$\begin{aligned}\mathbb{E}_x[N_x] &= \mathbb{E}_x\left[\sum_{n=0}^{\infty} 1_{\{X_n=x\}}\right]\\ &= \mathbb{E}_x\left[1_{\{X_0=x\}}\right] + \mathbb{E}_x\left[\sum_{n=1}^{\infty} 1_{\{X_n=x\}}\right]\\ &= 1 + \sum_{n=1}^{\infty}\mathbb{E}_x\left[1_{\{X_n=x\}}\right]\\ &= 1 + \sum_{n=1}^{\infty}(P^n)_{x,x}.\end{aligned}$$

Rearranging the previous expression and using the fact that $\mathbb{E}[N_x] = \frac{1}{1-f_x}$, we see that

$$\sum_{n=1}^{\infty}(P^n)_{x,x} = \frac{1}{1-f_x} - 1 = \frac{1}{1-f_x} - \frac{1-f_x}{1-f_x} = \frac{f_x}{1-f_x}.$$

We now consider the case when x is recurrent. A similar argument shows that $\mathbb{E}[N_x] = \infty$, and therefore that

$$\sum_{n=1}^{\infty}(P^n)_{x,x} = \infty.$$

□

From Proposition E.46 and its proof, we obtain the following corollary, which gives several equivalent criteria for transience and recurrence.

Corollary E.47. *For a Markov chain* $\{X_n\}$ *on* $\mathscr{S}$ *with transition matrix* P*, the following are equivalent for each* $x \in \mathscr{S}$*:*

(1) x is transient.

(2) $\mathbb{P}_x[N_x = \infty] = 0$.

(3) $\sum_{n=1}^{\infty} (P^n)_{x,x} = \frac{f_x}{1-f_x} < \infty$.

Additionally, the following are equivalent:

(1) x *is recurrent.*

(2) $\mathbb{P}_x[N_x = \infty] = 1$.

(3) $\sum_{n=1}^{\infty} (P^n)_{x,x} = \infty$.

The following result says that recurrent and transience are *class properties*, meaning that for a communication class $\mathscr{C}$, the transience or recurrence of a state $x \in \mathscr{S}$ determines the transience or recurrence of all of $\mathscr{C}$.

Theorem E.48. *Let $\mathscr{C}$ be a communication class. The states of $\mathscr{C}$ are either all recurrent or all transient.*

The following corollary follows from Proposition E.44 and Theorem E.48.

Corollary E.49. *Let $\{X_n\}$ be a finite irreducible Markov chain on $\mathscr{S}$. Then every state in $\mathscr{S}$ is positive recurrent.*

The following result can be used to determine whether the states within a communication class are transient or recurrent.

Theorem E.50. *Let $\mathscr{C} \subseteq \mathscr{S}$ be a communication class for a Markov chain. Then:*

(1) If $\mathscr{C}$ is open then every element in $\mathscr{C}$ is transient.

(2) If $\mathscr{C}$ is closed and contains only finitely many elements, then every element in $\mathscr{C}$ is recurrent.

We apply Theorem E.50 in the example below.

Example E.51. *Consider a Markov chain $\{X_n\}$ on $\mathscr{S} = \{1,2,3,4,5\}$ with transition matrix*

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

The communication classes are $\mathscr{C}_1 = \{1,2,3\}$, which is open, and $\mathscr{C}_4 = \{4\}$ and $\mathscr{C}_5 = \{5\}$, which are both closed. It follows that states 1,2, and 3 are transient, while states 4 and 5 are recurrent.

Note also that $\{X_n\}$ does not have a limiting distribution, as $\alpha_1^T = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \end{pmatrix}$ and $\alpha_2^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \end{pmatrix}$ lead to different limiting behavior for the sequence $\mathbb{P}_{\alpha_i}(X_n \in \cdot) = \alpha_i^T P^n$.

Furthermore, $\{X_n\}$ has uncountably many stationary distributions, as the vector $\pi_\lambda^T = \begin{pmatrix} 0 & 0 & 0 & \lambda & 1-\lambda \end{pmatrix}$ is a stationary distribution for each $\lambda \in [0,1]$.

The following proposition shows that we can partition a finite Markov chain's state space in terms of closed communication classes of recurrent states and transient states.

Proposition E.52. *The state space $\mathscr{S}$ of a finite Markov chain can be partitioned into transient and recurrent states as*

$$\mathscr{S} = T \cup R_1 \cup \cdots \cup R_m,$$

where T is the set of all transient states, and each R_i is a closed communication class of recurrent states.

The following example shows how we can use Proposition E.52 to decompose the state space of the Markov chain from Example E.51.

Example E.53. *Consider a Markov chain $\{X_n\}$ on $\mathscr{S} = \{1,2,3,4,5\}$ with transition matrix*

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

Recall that the communication classes are $\mathscr{C}_1 = \{1,2,3\}$, which is open, and $\mathscr{C}_4 = \{4\}$ and $\mathscr{C}_5 = \{5\}$, which are both closed. Recall that states 1,2, and 3 are transient, while states 4 and 5 are recurrent. The canonical decomposition of $\mathscr{S}$ is then given by $S = T \cup R_1 \cup R_2$, where $T = \{1,2,3\}$, $R_1 = \{4\}, R_2 = \{5\}$.

The following result answers some of our questions about the existence of stationary distributions, but not all of them. The proof is a consequence of our earlier observation that a stationary distribution is an eigenvector corresponding to an eigenvalue of 1, and the Perron-Frobenius Theorem. Recall, from Example E.29, however, stationary distributions of finite-state Markov chains need not be unique.

Proposition E.54. *If $\{X_n\}$ is a Markov chain on a finite state space, then it has* at least *one stationary distribution.*

The following example shows that Markov chains on infinite state spaces need not have any stationary distributions.

Example E.55. *Let $\{X_n\}$ be a simple symmetric random walk on $\mathscr{S} = \{0,1,-1,2,-2,\dots\}$. Suppose towards a contraction that a stationary distribution exists, namely that there is some probability distribution π on $\mathscr{S}$ such that*

$$\pi^T P = \pi^T. \tag{96}$$

Note that

$$P_{i,j} = \begin{cases} \frac{1}{2}, & j = i \pm 1 \\ 0, & \textit{otherwise.} \end{cases}$$

Therefore, (96) *is equivalent to saying that, for each $i \in \mathscr{S}$,*

$$\pi_i = \frac{1}{2}\pi_{i-1} + \frac{1}{2}\pi_{i+1},$$

which yields

$$\pi_{i+1} = 2\pi_i - \pi_{i-1},$$

or, equivalently,

$$\pi_{i+2} = 2\pi_{i+1} - \pi_i.$$

We now consider two cases: either $\pi_1 = \pi_0$ *or* $\pi_1 \neq \pi_0$.

Begin by assuming that $\pi_1 = \pi_0$. *Then*

$$\pi_2 = 2\pi_1 - \pi_0 = 2\pi_0 - \pi_0 = \pi_0.$$

Similarly,

$$\pi_3 = 2\pi_2 - \pi_1 = 2\pi_0 - \pi_0 = \pi_0,$$

so we can inductively prove that $\pi i = \pi_0$ *for all* $i \in \mathscr{S}$. *If* $\pi_0 = 0$, *this says that*

$$\sum_{i \in \mathscr{S}} \pi_i = 0,$$

which means that π *is not a probability measure, which is a contradiction. On the other hand, if* $\pi_0 > 0$, *this says*

$$\sum_{i \in \mathscr{S}} \pi_i = \infty,$$

which also means that π *is not a probability measure. Therefore, we conclude that we cannot have* $\pi_1 = \mu_0$.
We now consider the case when $\pi_1 \neq \pi_0$. *Then*

$$\pi_2 = 2\pi_1 - \pi_0 = \pi_1 + (\pi_1 - \pi_0).$$

which tells us that

$$\pi_2 - \pi_1 = \pi_1 - \pi_0.$$

Thus

$$\pi_3 = 2\pi_2 - \pi_1 = \pi_2 + (\pi_2 - \pi_1) = \pi_2 + (\pi_1 - \pi_0),$$

so we can show inductively that

$$\pi_{i+1} = \pi_i + (\pi_1 - \pi_0).$$

Once more arguing by induction, we can show that we must have

$$\pi_{i+1} = \pi_0 + i(\pi_1 - \pi_0).$$

If $\pi_1 > \pi_0$, *then it follows that for all* $i > \frac{2}{\pi_1 - \pi_0}$,

$$\pi_{i+1} = \pi_0 + i(\pi_1 - \pi_0) > \pi_0 + \frac{2}{\pi_1 - \pi_0}(\pi_1 - \pi_0) > \pi_0 + 2 > 1,$$

which again tells us that π *is not a probability measure. Similarly, if* $\pi_1 < \pi_0$, *then for all* $i > \frac{2}{\pi_0 - \pi_1}$, *we have, since* $\pi_1 - \pi_0 < 0$,

$$\pi_{i+1} = \pi_0 + i(\pi_1 - \pi_0) < \pi_0 + \frac{2}{\pi_0 - \pi_1}(\pi_1 - \pi_0) = \pi_0 - 2 < 0,$$

which again tells us that π *is not a probability. measure.*

We conclude that $\{X_n\}$ *does not have a stationary distribution.*

The next result shows that an irreducible Markov chain has at most one stationary distribution. However, it does not guarantee the existence of a stationary distribution, as irreducible Markov chains on infinite state spaces may not have stationary distributions, as, for example, we saw in Example E.55.

Proposition E.56. *Let $\{X_n\}$ be an irreducible Markov chain. Then $\{X_n\}$ has at most one stationary distribution.*

This leads to the following important result regarding finite Markov chains.

Corollary E.57. *A finite irreducible Markov chain has a unique positive stationary distribution. Here a distribution π on $\mathscr{S}$ is said to be positive $\pi_x > 0$ for all $x \in \mathscr{S}$.*

E.7. **The Ergodic Theorem.** In some settings it is not practical to explicitly calculate the stationary distribution of a Markov chain. This may cause difficulties, but if the stationary distribution is in fact the limiting distribution of the chain, then the stationary distribution can easily be estimated via simulation. Accordingly, it is natural to try to determine when a stationary distribution is a limiting distribution.

Before we state our main results, recall from Proposition E.54 that a finite-state Markov chain always has at least one stationary distribution. Additionally, we know that if a Markov chain has a limiting distribution, then the limiting distribution is unique. And from Corollary E.57 we know that the stationary distribution of an irreducible Markov chain, if it exists, is unique.

This leads to the first situation (for finite Markov chains) in which a stationary distribution need not be a limiting distribution, namely, when the stationary distribution is not unique (i.e., when the chain is not irreducible).

Example E.58. *Consider the transition matrix on $\mathscr{S} = \{a, b, c, d\}$ given by*

$$P = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

*Then $\pi_1^T = \begin{pmatrix} 0 & 0 & 1/2 & 1/2 \end{pmatrix}$ and $\pi_2^T = \begin{pmatrix} 5/11 & 6/11 & 0 & 0 \end{pmatrix}$ are stationary distributions for the chain. Note that the chain is **not** irreducible.*

The other setting in which stationary distributions need not be limiting distributions is that of periodic Markov chain. Before we introduce the notion of periodicity, we consider an example.

Example E.59. *Consider a Markov chain on $\mathscr{S} = \{1, 2, 3, 4\}$ with transition matrix*

$$P = \begin{pmatrix} 0 & 0.4 & 0 & 0.6 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix}.$$

Observe that the Markov chain repeatedly jumps between $\{1, 3\}$ and $\{2, 4\}$, so, for example, if we let $\mu^T \doteq \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$, then the sequence of probabilities $\{\mu^T P^n\}$ does not converge.

Recall that the greatest common divisor, or gcd, of a set $\mathscr{I} \subseteq \mathbb{N}$ of positive integers is the quantity as

$$\gcd(\mathscr{I}) \doteq \sup\{n \in N : i/n \in \mathbb{N} \text{ for all } i \in \mathscr{I}\}.$$

The gcd of $\mathscr{I}$ is the largest positive integer n such that i/n is a whole number for all $i \in \mathscr{I}$.

Definition E.60. *Let P be a transition matrix on $\mathscr{S}$. The **period** of a state $x \in \mathscr{S}$ is defined as*

$$d(x) \doteq gcd\{n \in \mathbb{N} : (P^n)_{x,x} > 0\}.$$

*If $d(x) = 1$, state x is said to be **aperiodic**. If $(P^n)_{x,x} = 0$ for all $n \in \mathbb{N}$, we let $d(x) \doteq \infty$.*

*If all states are aperiodic, the chain is said to be **aperiodic**.*

We explore the notion of periodicity in the example below.

Example E.61. *Consider the Markov chain on $\mathscr{S} = \{1,2,3,4\}$ with transition matrix*

$$P = \begin{pmatrix} 0 & 0.4 & 0 & 0.6 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix},$$

from Example E.59. *Then*

$$\{n \in \mathbb{N} : (P^n)_{1,1} > 0\} = \{n \in \mathbb{N} : (P^n)_{2,2} > 0\} = \{2,4,6,\ldots\},$$

and

$$\{n \in \mathbb{N} : (P^n)_{3,3} > 0\} = \{n \in \mathbb{N} : (P^n)_{4,4} > 0\} = \{2,4,6,\ldots\},$$

so $d(1) = d(2) = d(3) = d(4) = 2$, so every state has a period of 2.

Now consider the transition matrix $\bar{P}$ on $\bar{\mathscr{S}} = \{a, b\}$ given by

$$\bar{P} = \begin{pmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix},$$

and note that $d(a) = d(b) = 1$, so every state is aperiodic.

Just as we saw with transience and recurrence, periodicity is a class property.

Proposition E.62. *The states of a communication class all have the same period.*

One consequence of Proposition E.62 is the following:

Corollary E.63. *If $\{X_n\}$ is an irreducible Markov chain on $\mathscr{S}$, and $P_{x,x} > 0$ for some $x \in \mathscr{S}$, then $\{X_n\}$ is aperiodic.*

Note that all of the Markov chains we saw for which the stationary distribution was not a limiting distribution did not satisfy the conditions in the definition below. In particular, they were not ergodic; ergodic chains constitute the most important class of Markov chains because their stationary and limiting behavior agree.

Definition E.64. *A Markov chain is called* **ergodic** *if it is irreducible (i.e., has a single communication class) and aperiodic.*

We begin by stating the Ergodic Theorem for finite-state Markov chains. Recall from Corollary E.57 that finite irreducible (and therefore ergodic) Markov chains have a unique positive stationary distribution.

Theorem E.65. *Let $\{X_n\}$ be an ergodic Markov chain on a finite state space with (unique and positive) stationary distribution π. Then for each $x \in \mathscr{S}$, with*

$$R_x \doteq \inf\{n > 0 : X_n = x\},$$

we have

$$\pi_x = \frac{1}{\mathbb{E}_x[R_x]}.$$

Furthermore, π is the unique limiting distribution of $\{X_n\}$, meaning that for all initial distributions μ on $\mathscr{S}$ and all $x \in \mathscr{S}$,

$$\lim_{n\to\infty} \mathbb{P}_\mu[X_n = x] = \lim_{n\to\infty} (\mu^T P^n)_x = \pi_x.$$

The following example illustrates how to compute the stationary distribution of an ergodic Markov chain.

Example E.66. *Consider a Markov chain on $\mathscr{S} = \{a, b, c, d\}$ with transition matrix*

$$P = \begin{pmatrix} 0 & 0.4 & 0 & 0.6 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix}.$$

Note that the chain is irreducible and $P_{c,c} > 0$, so it is ergodic. Therefore, it has a unique stationary distribution which solves

$$\pi^T P = \begin{pmatrix} \pi_a & \pi_b & \pi_c & \pi_d \end{pmatrix} \begin{pmatrix} 0 & 0.4 & 0 & 0.6 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0.9 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix} = \begin{pmatrix} \pi_a & \pi_b & \pi_c & \pi_d \end{pmatrix} = \pi^T.$$

However, the solution to the above equation is not unique; if v solves the equation above, so too does cv for each $c \in \mathbb{R}$ (as we saw on the homework). In order to remedy this, we need to use the additional constraint that $\pi_a + \pi_b + \pi_c + \pi_d = 1$. The system of equations becomes

$$\begin{pmatrix} \pi_a & \pi_b & \pi_c & \pi_d \end{pmatrix} \begin{pmatrix} 0 & 0.4 & 0 & 0.6 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0.1 & 0.9 & 1 \\ 0.5 & 0 & 0.5 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \pi_a & \pi_b & \pi_c & \pi_d & 1 \end{pmatrix},$$

which has the unique solution

$$\pi^T \approx \begin{pmatrix} 0.306 & 0.123 & 0.204 & 0.367 \end{pmatrix}.$$

Note that this tells us that, starting from state a, it will, on average, take

$$\mathbb{E}_a[R_a] = \frac{1}{\pi_a} \approx \frac{1}{0.306} = 3.268,$$

steps to return a.

Now, suppose that Markov chain starts according to initial distribution $\mu^T = \begin{pmatrix} 0.1 & 0.1 & 0.5 & 0.3 \end{pmatrix}$. *Since the Markov chain is ergodic, we know that*

$$(\mu^T P^n) \to \pi^T,$$

since the stationary distribution is the limiting distribution for ergodic Markov chain. This tells us that, in the long term, the average proportion of time it will spend in state a is about 0.306, the average proportion of time it will spend in state b is 0.123, and so on.

We now consider the queueing example from Example E.14.

Example E.67. *A queue at a bank starts with no one in it. Each minute, there is a 40% chance someone joins the line and a 60% chance someone leaves it. Furthermore, if the line already has 4 people in it, then any arrivals simply walk away rather than waiting in line.*

In the long term, on average, what proportion of the time is the queue full? Does this depend on the initial distribution?
The queue is modeled by a Markov chain on $\mathscr{S} = \{0,1,2,3,4\}$ *with transition matrix*

$$P = \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 \\ 0.6 & 0 & 0.4 & 0 & 0 \\ 0 & 0.6 & 0 & 0.4 & 0 \\ 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0.6 & 0.4 \end{pmatrix}.$$

The chain is ergodic, since it is irreducible and aperiodic, as $P_{0,0} > 0$. *Thus, its limiting distribution is the unique stationary distribution, so it is enough to solve the system of equations given by*

$$\begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 & \pi_4 \end{pmatrix} \begin{pmatrix} 0.6 & 0.4 & 0 & 0 & 0 & 1 \\ 0.6 & 0 & 0.4 & 0 & 0 & 1 \\ 0 & 0.6 & 0 & 0.4 & 0 & 1 \\ 0 & 0 & 0.6 & 0 & 0.4 & 1 \\ 0 & 0 & 0 & 0.6 & 0.4 & 1 \end{pmatrix} = \begin{pmatrix} \pi_0 & \pi_1 & \pi_2 & \pi_3 & \pi_4 & 1 \end{pmatrix},$$

or, equivalently,

$$\begin{cases} -0.4\pi_0 + 0.6\pi_1 = 0 \\ 0.4\pi_0 - \pi_1 + 0.6\pi_2 = 0 \\ 0.4\pi_1 - \pi_2 + 0.6\pi_3 = 0 \\ 0.4\pi_2 - \pi_3 + 0.6\pi_4 = 0 \\ 0.4\pi_3 - 0.6\pi_4 = 0 \\ \pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1. \end{cases}$$

Since the stationary distribution is the (appropriately normalized) left eigenvector of P corresponding to an eigenvalue of 1, *we can calculate it in python with the following:*

```
P = np.array([[0.6, 0.4,0,0,0],[0.6,0,0.4,0,0],
[0,0.6,0,0.4,0],[0,0,0.6,0,0.4],[0,0,0,0.6,0.4]])
evals, evecs = np.linalg.eig(P.T)
evec1 = evecs[:,np.isclose(evals, 1)]
#Since np.isclose will return an array, we've indexed with an array
```

```
#Convert to a vector
evec1 = evec1[:,0]
#Normalize the eigenvector to sum to 1
stationary = evec1 / evec1.sum()
stationary = stationary.real
stationary
```

We obtain

$$\pi^T = \begin{pmatrix} 0.384 & 0.256 & 0.171 & 0.114 & 0.075 \end{pmatrix},$$

which tells us that the queue, on average, is full about 7.5% of the time.

We now consider an example where we can analytically derive the stationary distribution.

Example E.68. *Consider a simple symmetric random walk with partially reflecting boundaries at* 0 *and* N. *The transition matrix* P *is given by, for* $x \in \{1,\dots,N-1\}$,

$$P_{x,x+1} = P_{x,x-1} = \frac{1}{2},$$

and

$$P_{0,0} = P_{0,1} = P_{N,N-1} = P_{N,N} = \frac{1}{2}.$$

The stationary distribution π *satisfies*

$$\begin{cases} \pi_0 = \frac{1}{2}(\pi_0 + \pi_1) \\ \pi_i = \frac{1}{2}(\pi_{i-1} + \pi_{i+1}), \quad 1 \le i \le N-1 \\ \pi_N = \frac{1}{2}(\pi_{N-1} + \pi_N). \end{cases}$$

We obtain

$$\pi_0 = \pi_1, \quad \pi_1 = \pi_2, \dots,$$

which means that

$$\pi_i = \frac{1}{N+1},$$

for all $0 \le i \le N$.

E.8. **Ergodic Theorem for Unichains.** Consider a Markov chain $\{X_n\}$ on $\mathscr{S} = \{1,2,3\}$ with transition matrix

$$P = \begin{pmatrix} 0.3 & 0.5 & 0 \\ 0.6 & 0.4 & 0 \\ 0.2 & 0.2 & 0.6 \end{pmatrix},$$

and note that $\{X_n\}$ is not irreducible, as

$$\mathbb{P}_3[\tau_3 = \infty] \ge 0.4 > 0,$$

which means that

$$\mathbb{P}_3[\tau_3 < \infty] < 1.$$

Since state 3 is transient, we cannot apply the ergodic theorem as stated in Theorem E.65. However, if we consider the stochastic matrix

$$\tilde{P} \doteq \begin{pmatrix} 0.3 & 0.5 \\ 0.6 & 0.4 \end{pmatrix},$$

then we can see that $\tilde{P}$ is the transition matrix of an ergodic Markov chain on $\mathscr{C}_1 = \{1,2\}$.

$$P = \begin{pmatrix} \tilde{P}_{1,1} & \tilde{P}_{1,2} & 0 \\ \tilde{P}_{2,1} & \tilde{P}_{2,2} & 0 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}.$$

Observe that if the chain starts in $\mathscr{C}_1$, then it will never enter state 3 and will then evolve according to the transition matrix $\tilde{P}$. On the other hand, if it starts in state 3, then, since state 3 is transient, the chain will eventually jump to $\mathscr{C}_1$, after which point it will continue evolving according to $\tilde{P}$. Thus, in the long term, we should expect that the chain will behave like an ergodic Markov chain with transition matrix $\tilde{P}$.

Thus, if we found the unique stationary distribution $\tilde{\pi}^T = \begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2\end{pmatrix}$ for $\tilde{P}$, then we would expect that the unique stationary distribution for $\{X_n\}$ would be given by $\pi^T = \begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2 & 0\end{pmatrix}$, and that π would be the limiting distribution of $\{X_n\}$ as well. This is true, as we see later in this section. A Markov chain that consists of a single recurrent communication class and a collection of transient states is known as a unichain.

Definition E.69. *Suppose that $\{X_n\}$ is a Markov chain with finite state space $\mathscr{S}$, and that we can write*

$$\mathscr{S} = \mathscr{C} \cup T,$$

where $\mathscr{C}$ is a communication class of recurrent states, and T is a collection of transient states. Then $\{X_n\}$ is called a ***unichain****. If we write $\tilde{P}$ to denote the transition matrix of $\{X_n\}$ restricted to $\mathscr{C}$, then $\tilde{P}$ is a stochastic matrix. If $\tilde{P}$ is the transition matrix of an ergodic Markov chain, then we refer to $\{X_n\}$ as an ergodic unichain.*

The idea behind finite-state ergodic unichains is that they may move around in the set of transient states for a finite time period, after which they will jump into the set of recurrent states and behave like an ergodic Markov chain. This is captured in the following theorem.

Theorem E.70. *Let $\{X_n\}$ be an ergodic unichain on finite state space $\mathscr{S} = \{1,2,\dots,r-1,r,r+1,\dots,t\}$, and let $T = \{r+1,r+2,\dots,t\}$ denote the set of transient states, and $\mathscr{C} = \{1,2,\dots,r-1,r\}$ denote the communication class of recurrent states. Then the unique stationary distribution of $\{X_n\}$ is given by*

$$\pi^T = \begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2 & \dots & \tilde{\pi}_r & 0 & 0 & \cdots & 0\end{pmatrix},$$

where the last $t-r$ entries are all 0. Furthermore, π is the limiting distribution of the chain.

The following example illustrates how we can apply this theorem.

Example E.71. *Consider the Markov chain from on $\mathscr{S} = \{1,2,3\}$ from the beginning of this section with transition matrix*

$$P = \begin{pmatrix} 0.3 & 0.5 & 0 \\ 0.6 & 0.4 & 0 \\ 0.2 & 0.2 & 0.6 \end{pmatrix}.$$

If we let $T = \{3\}$ and $\mathscr{C} = \{1,2\}$, then T is the collection of transient states, and $\mathscr{C}$ is the set of recurrent states. Furthermore,

$$\tilde{P} \doteq \begin{pmatrix} 0.3 & 0.5 \\ 0.6 & 0.4 \end{pmatrix}$$

is the transition matrix of an ergodic Markov chain on $\mathscr{C}$*, so we can apply Theorem* E.70 *to find the limiting distribution and the stationary distribution of* $\{X_n\}$*. The stationary distribution* $\tilde{\pi}$ *corresponding to* $\tilde{P}$ *is the unique solution to*

$$\begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2\end{pmatrix}\begin{pmatrix}0.3 & 0.5 & 1\\ 0.6 & 0.4 & 1\end{pmatrix} = \begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2 & 1\end{pmatrix},$$

and is given by

$$\tilde{\pi}^T = \begin{pmatrix}0.462 & 0.538\end{pmatrix}.$$

Therefore, the limiting and stationary distribution of $\{X_n\}$ *is given by*

$$\pi = \begin{pmatrix}0.462 & 0.538 & 0\end{pmatrix}.$$

We apply Theorem E.70 again in the example below.

Example E.72. *Consider the Markov chain* $\{X_n\}$ *on* $\mathscr{S} = \{1,2,3,4,5\}$ *with transition matrix*

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

Here the transient states are $T = \{4,5\}$ *and the closed communication class of recurrent states is* $\mathscr{C} = \{1,2,3\}$*. Note that the unique solution to*

$$\begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2 & \tilde{\pi}_3\end{pmatrix}\begin{pmatrix}1/2 & 1/2 & 0 & 1\\ 1/4 & 1/4 & 1/2 & 1\\ 1/4 & 1/2 & 1/4 & 1\end{pmatrix} = \begin{pmatrix}\tilde{\pi}_1 & \tilde{\pi}_2 & \tilde{\pi}_3 & 1\end{pmatrix}$$

is given by

$$\tilde{\pi}^T = \begin{pmatrix}0.333 & 0.4 & 0.267\end{pmatrix},$$

so the ergodic theorem for unichains tells us that the limiting and stationary distribution of $\{X_n\}$ *is given by*

$$\pi^T = \begin{pmatrix}0.333 & 0.4 & 0.267 & 0 & 0\end{pmatrix}.$$