Markov Chains

Kshitij Tomar

July 14, 2024

Abstract

Studying complicated probability models requires a much deeper understanding of randomness in systems than that viewed in an advanced high school statistics class. This paper is meant to act as an introduction to motivated high school students and undergraduates who are aiming to learn about Markov Chains and its underlying theory. We will introduce the necessity of Markov Chain theory, discuss its many applications, and build up to the Ergodic Theorem and the Convergence Theorem. From there onwards, this paper discusses the Metropolis-Hastings Algorithm and Monte Carlo Integration.

1 Introduction

Independence amongst multiple random variables is a foundational concept in probability theory. For example, we know that rolling pairs of standard dice and flipping coins in a sequential pattern are considered independent events. In addition to that, we know that each subsequent roll of a die or coin flip follows the same probability distribution as any other die or coin flip. That is, these sequences of random events are identically distributed. On one hand, we find that it is quite a common assumption to make in basic systems. The theory regarding these types of systems has been very well studied, thus making analysis of these systems quite easy. On the other hand, it is quite difficult, with basic probability knowledge alone, to study more complicated systems. This is a problem because there are many naturally occurring systems we would like to study who have some sort of systematic dependence.

This means that we need to introduce a new tool, Markov Chains, that looks upon random variables in a more general way so that we can study systems that have some structure of dependency. An example of a system we would study with Markov Chains is weather. If we wanted to know the probability of today being Sunny, we would need to make major assumptions about how weather changes over time. The main problem in predictive weather modeling is the fact that weather is always dependent on the weather that immediately precedes it. In other words, a rainy day today would impact the chances that there is a sunny day tomorrow. That is, there is a dependence between the weathers in these days. What is important to note here is that it's not just weather. The study of particles randomly hopping on a one-dimensional lattice has previously been defined as a Markov Chain to study protein synthesis and traffic flow [10]. Markov Chains have also appeared in conjunction with other techniques to develop an analytical model for evaluating the performance of security protocols [2]. In addition, modern algorithms such as Google's Pagerank fundamentally use Markov Chain theory to predict the best search queries for an individual based on their previous search history [9]. In physics, the Ising model and Glauber dynamics are also systems that have previously used Markov Chain methods for a precise understanding of the properties in these systems. Historically, the Ehrenfest Urn model was also evident as part of the system that had an underlying Markov Chain structure. Recently, subsurface flows and complicated fluid dynamics have been simulated using statistical methods such as Markov Chain Monte Carlo. For more information on these applications, check out the research papers from [1] and [3]. Similarly, there are artificial intelligence algorithms that implement Markov Chains. Some interesting ones include [4] and [6]. For a more detailed discussion on Markov Chains in AI, check out [8].

This paper assumes that the reader will have a basic understanding of probability theory and a level of understanding of linear algebra as a first-year or second-year course in college. There are many times throughout this paper where the knowledge of matrices will be applied for important deductions or steps of proof. If the reader does not feel confident throughout any part of this paper, it is advised to brush up on those topics before continuing through the rest of the paper.

2 Looking at Markov Chains

Definition 2.1. Let χ be a countable state space. A sequence of random variables (X_t) is called a **Markov Chain**, where $X_t \in \chi$ for all t in an indexing set A,

$$\mathbb{P}(X_t = y \mid X_{t-1} = x, X_{t-2}, \ldots, X_0) = \mathbb{P}(X_t = y \mid X_{t-1} = x). \tag{2.1}$$

Equation 2.1 is known as the Markov Property, and it considers systems where the probability of traveling from one state to another is independent of all the previous states except the state immediately preceding it. In this paper, the indexing set A will mostly represent time and the finite state space will be a subset of $\mathbb{N}$. There are some interesting situations where this is not the case though. Do note that there are continuous-time Markov Chains and continuous-space Markov Chains, where time t is indexed on an uncountably infinite set or space χ is an uncountably infinite set. With regard to these Markov Chains, the ones covered in this paper are more commonly known as Discrete Time Markov Chains.

Compared to **i.i.d (independent and identically distributed)** sequences of random variables, Markov Chains, and their resulting consequences allow deeper analysis of a general class of systems. In other words, they can describe more complicated systems than i.i.d systems alone. This is what makes Markov Chains useful. The following is a classical example of a Markov Chain.

Example 2.1 (Gambler's Ruin). Let $k > 0$ be the number of coins a gambler starts with. He makes a bet at his local casino in which every time he flips a coin and gets heads, he gains one coin. Every time he flips a coin and gets a tails, he loses one coin. Assume that the gambler continues betting until he has $n > k$ coins or he has 0 coins. What is the probability that the gambler walks away with n coins?

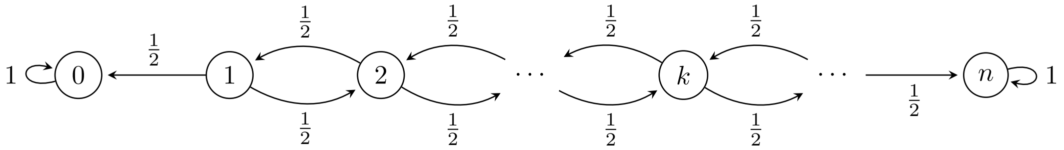

Figure 1: Gambler's Ruin Illustration

Proof. Let $\mathbb{P}_k$ be the probability of reaching the state n coins when starting from k coins where $1 \leq k \leq n-1$ coins. Realize that $\mathbb{P}_0 = 0$ and $\mathbb{P}_n = 1$. Notice that $\mathbb{P}_k = \mathbb{P}(\text{tails})\mathbb{P}_{k-1} + \mathbb{P}(\text{heads})\mathbb{P}_{k+1}$. Therefore,

$$\mathbb{P}_k = \frac{1}{2}\mathbb{P}_{k-1} + \frac{1}{2}\mathbb{P}_{k+1}.$$

We now proceed to prove that $\mathbb{P}_k = \frac{k}{k+1}\mathbb{P}_{k+1}$ via induction. For $k = 0$, $\mathbb{P}_1 = \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_2 = \frac{1}{2}\mathbb{P}_2$. Thus, the base case meets this forms.

Next, assume that for k, $\mathbb{P}_k = \frac{k}{k+1}\mathbb{P}_{k+1}$. Then,

$$\begin{aligned}
\mathbb{P}_{k+1} &= \frac{1}{2}\mathbb{P}_k + \frac{1}{2}\mathbb{P}_{k+2}. \\
\mathbb{P}_{k+1} &= \frac{k}{2(k+1)}\mathbb{P}_{k+1} + \frac{1}{2}P_{k+2} \\
\mathbb{P}_{k+1} - \frac{k}{2(k+1)}\mathbb{P}_{k+1} &= \frac{1}{2}\mathbb{P}_{k+2} \\
\frac{k+2}{2(k+1)}\mathbb{P}_{k+1} &= \frac{1}{2}\mathbb{P}_{k+2} \\
\mathbb{P}_{k+1} &= \frac{k+1}{k+2}\mathbb{P}_{k+2}.
\end{aligned}$$

Therefore, $\mathbb{P}_k = \frac{k}{k+1}\mathbb{P}_{k+1}$ for all $0 \leq k < n$. However, this implies that

$$\begin{aligned}\mathbb{P}_k &= \frac{k}{k+1}\frac{k+1}{k+2}\mathbb{P}_{k+2}\\ \mathbb{P}_k &= \frac{k}{k+1}\frac{k+1}{k+2}\frac{k+2}{k+3}\cdots\frac{n-1}{n}\mathbb{P}_n\\ \mathbb{P}_k &= \frac{k}{n}.\end{aligned}$$

■

It is quite common to represent systems for Markov chains as directed graphs. This makes it easier to visualize the chain. If there is a non-zero probability of moving from one state to another, we draw an arrow between those states and label it with that probability. A natural question arises from looking at the Gambler's Ruin example: how long does this game take? We know that eventually one of the states n or 0 must be the final ending state of this game. In other words, what is the expected number of coin flips for the gambler to receive n coins or the gambler to go home broke? To look at the solution of a problem like this one, it becomes imperative to define the "number of coin flips" in a more rigorous way.

Definition 2.2. The **hitting time** of a Markov Chain (X_t) on a state space χ for a state $x \in \chi$ is

$$\tau_x = \min\{t \geq 0 : X_t = x\},$$

and

$$\tau_x^+ = \min\{t \geq 1 : X_t = x\}.$$

When $X_0 = x$, τ_x^+ is known as the **first return time**.

We can rephrase our question to match a more precise definition. This will help us understand what and how we should aim to solve this problem.

Example 2.2. For the same gambler in Exercise 2.1, what is the expected hitting time for state n or state 0, given that we begin on state k?

Proof. For the expected value, a similar recursive approach is possible. Let the expected length of the game be f_k. Notice that $f_n = 0$. On the other hand, the game ends when we start with 0 coins. Thus, $f_0 = 0$. Now, we can write the movement along this chain for $1 \leq k \leq n-1$ as the recursive statement in which we can move to an adjacent state and then move from that respective state onwards to state n.

$$f_k = \frac{1}{2}(1 + f_{k+1}) + \frac{1}{2}(1 + f_{k-1}).$$

By induction, we aim to prove that

$$f_k = k + \frac{k}{k+1}f_{k+1}.$$

When $k = 1$, $f_1 = \frac{1}{2}(1 + f_2) + \frac{1}{2}(1 + f_0) = 1 + \frac{1}{2}f_2$ and the base case holds.

Next, assume that for $1 \leq k \leq n-1$ that $f_k = k + \frac{k}{k+1}f_{k+1}$. Then,

$$\begin{aligned}f_{k+1} &= \frac{1}{2}(1 + f_k) + \frac{1}{2}(1 + f_{k+2})\\ f_{k+1} &= \frac{1}{2}(1 + k + \frac{k}{k+1}f_{k+1}) + \frac{1}{2}(1 + f_{k+2})\\ f_{k+1} - \frac{k}{2(k+1)}f_{k+1} &= \frac{1}{2}(2 + k + f_{k+2})\\ \frac{k+2}{2(k+1)}f_{k+1} &= \frac{1}{2}(2 + k + f_{k+2})\\ f_{k+1} &= \frac{k+1}{k+2}(2 + k + f_{k+2})\\ f_{k+1} &= k + 1 + \frac{k+1}{k+2}f_{k+2}.\end{aligned}$$

Therefore, $f_k = k + \frac{k}{k+1} f_{k+1}$ for all $0 \leq k < n$. However, this implies that for $0 \leq k \leq n$,

$$f_k = k + \frac{k}{k+1}(k+1+\frac{k+1}{k+2}f_{k+2})$$
$$f_k = k + \frac{k}{k+1}(k+1+\frac{k+1}{k+2}(k+2+\frac{k+2}{k+3}\cdots(n-1)\cdots))$$
$$f_k = k + \cdots + k$$
$$f_k = k(n-k).$$

■

When looking at random variables, it is of interest to understand how likely it is for one to travel from one state to another. For example, if today was rainy on Saturday then would it be rainy again on Sunday? What are the chances that it will be sunny on Sunday? What if there were more than 1000 states of interest? To completely define the probabilities in these complicated systems, we can store these probabilities in a matrix.

Definition 2.3. Let χ be a finite state space. A **transition matrix** P of the Markov Chain (X_t) on χ is a $|\chi| \times |\chi|$ matrix such that for all t and $x, y \in \chi$

$$\mathbb{P}(X_t = y \mid X_{t-1} = x) = P(x,y) \tag{2.2}$$

$P(x,y)$ is known as the **transition probability** of moving from a state x to a state y. Denote $P(x,\cdot)$ as the x-th row of the transition matrix.

This means the x-th row of the matrix is going to be the probability distribution $P(x,\cdot)$. Therefore, the total sum of all components of any row in this matrix must sum to one and all entries in this matrix are non-negative, as this is true of all probability distributions. This property makes the transition matrix **stochastic** and is also why Markov Chains are considered **stochastic processes**.

The benefit of writing these probabilities in transition matrices is that they allow for some neat manipulations. For all $x, y \in \chi$, if we wanted to know $\mathbb{P}(X_t = y \mid X_{t-2} = x)$, on a Markov Chain (X_t) then,

$$\mathbb{P}(X_t = y \mid X_{t-2} = x) = \sum_{z\in\chi} \mathbb{P}(X_t = y \mid X_{t-1} = z)\, \mathbb{P}(X_{t-1} = z \mid X_{t-2} = x) \tag{2.3}$$
$$= \sum_{z\in\chi} P(y,z)P(z,x) \tag{2.4}$$
$$= P^2(x,y) \tag{2.5}$$

It is relatively simple to show (via induction) that traveling t-steps in a Markov Chain and reaching some state y from a starting state x is the same as the matrix index $P^t(x,y)$. This is sometimes referred to as the **t-step transition probability**. What is important about this is that it allows the same matrix to help explain an arbitrary number of steps on a Markov Chain, a useful property that this paper will aim to exploit.

From example 2.1, the probability of eventually traveling to state zero is $1 - \frac{k}{n}$. This is because as $t \to \infty$ the chain will end up at either 0 or n, from which the chain will be stuck in that state. Such states are known as **absorbing states**. However, some questions are brought up by an example like this one. Are all Markov Chains convergent onto some distribution? If not, then what types of chains will eventually follow a predictable pattern?

In other words, we aim to find the limiting distribution of this Markov Chain. Let's suppose that we begin with a distribution μ_0, which defines which states we could potentially start from, and transition matrix P. Then, we know that for $t \geq 0$, $\mu_{t+1} = \mu_t P$. In general for some integer $k \geq 0$,

$$\mu_{k+t} = \mu_k P^t. \tag{2.6}$$

Therefore, we also know that $\mu_t = \mu_0 P^t$. This shows that multiplying by the transition matrix allows us to study the Markov Chain when it has progressed one more step than before. Naturally, we want to understand what happens when $t \to \infty$. The resultant distribution is known as a limiting distribution. All limiting distributions are considered stationary, and rightfully named so.

Definition 2.4. A row vector π is a **stationary distribution** for a Markov Chain (X_t) on χ with a transition matrix P when

$$\pi = \pi P. \tag{2.7}$$

Another equivalent formulation is that a row vector π is a stationary distribution if for all $x \in \chi$,

$$\pi(x) = \sum_{y \in \chi} \pi(y) P(y, x). \tag{2.8}$$

Note that $\pi(x)$ represents the entry in the row vector that corresponds to the state x. It does not necessarily have to be true that the state x is an integer.

Markov Chains and their subsequent techniques are employed on complicated systems to typically approximate the limiting distributions for these systems. Limiting distributions consist of long-term probabilities on any particular state, regardless of the starting state or time t of the chain. Even if the starting state is completely not random, we could still fully understand the dependent structure by understanding the limiting distribution. Because of this useful property, it becomes important to check if every Markov Chain has a unique limiting distribution. If not, what constraints do we need to put on our Chain so that we can guarantee that it does have a limiting distribution? What is really cool about stationary distributions is they help to understand the randomness in a system. Do note that stationary distributions are not necessarily limiting distributions. Consider the figure below for a start at this distinction.

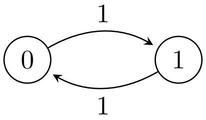

Figure 2: Bit Flip System

As seen in Figure 2, we have a system that switches bits after every instance. In this system, it is not hard to see that this system lacks randomness. Without loss of generality, assume that this chain begins at state 0. Then, we know with certainty that the next state is 1 and the subsequent chain after that is 0. In other words, for all $t \geq 0$, $\mathbb{P}(X_{2t} = 0 \mid X_0 = 0) = 1$ and $\mathbb{P}(X_{2t+1} = 1 \mid X_0 = 0) = 1$. This distribution fails to have a limiting distribution. We can try the same technique as before to see why this is the case. First, we assumed the starting distribution is $\mu_0 = \begin{bmatrix} 0 & 1 \end{bmatrix}$ and the transition matrix

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

As we progress in this chain,

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

and then progress again

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

As we keep increasing the time t, μ_t does not approach any particular distribution but rather switches between μ_0 and μ_1. To look at this more concretely, the reader can check that P^2 is the identity matrix and that $P^3 = P$ for this particular system. Since an original distribution μ_0 doesn't reach a distribution as time progresses, we find that this system fails to have a limiting distribution. However, $\mu = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$ is a stationary distribution on this system because

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Therefore, the existence of a stationary distribution does not necessarily mean that a limiting distribution exists. However, both of these types of distributions are worth studying for a little more because of there interesting properties. So, let's look at what happens when we add another state.

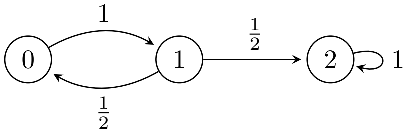

Figure 3: Added State to BFS

When we added the absorbing state 2, notice that the distribution would eventually end at 2 as $t \to \infty$. This is because as we continue the walk along the chain, regardless of the starting position, the walk will eventually end up at state 2. If it doesn't, we could continue changing states along the chain until it does land on state 2. Therefore, a further distinction of states within any given system will be required.

Definition 2.5. For a Markov Chain (X_t) on state space χ, let P be the transition matrix for the Markov Chain. A state $x \in \chi$ is called **essential** when for every $y \in \chi$ which has a $r_1 > 0$ such that $P^{r_1}(x, y) > 0$ implies that there exists a $r_2 > 0$ such that $P^{r_2}(y, x) > 0$. A state is considered **inessential** when the state is not essential.

Here states have been divided into two categories. Essential states help to distinguish states that are absorbing from other states. We are essentially looking at the states that always have a path back from any other state. In other words, it is always possible to return to that state. These states are considered essential because they seem to always be included in the stationary distribution.

Lemma 2.6. *Let (X_t) be a Markov Chain on state space χ and P be the transition matrix for the Markov Chain. If $x \in \chi$ is an essential state such that for $y \in \chi$ there is a $r > 0$ where $P^r(x, y) > 0$, then y is an essential state.*

Proof. For any $z \in \chi$ in which there exists a $r_1 > 0$ such that $P^{r_1}(y, z) > 0$ then, $P^{r+r_1}(x, z) > 0$. Since x is essential, there must exist a $r_2 > 0$ such that $P^{r_2}(z, x) > 0$. However, this means that $P^{r_2+r_1}(z, y) > 0$. Therefore, y is also an essential state. ■

However, this lemma motivates a further distinction of chains into collections of states. This is because if a state is essential and it connects to another state, that state is also essential. Therefore, we find that there will be a collection of essential states if there is at least one state. Consequently, there may also be a collection of inessential states.

Definition 2.7. For a state $x \in \chi$, a **communication class** $[x]$ is a collection of all states $y \in \chi$ such that all states in $[x]$ are essential or all states in $[x]$ are inessential such that for every $a, b \in [x]$ there is a $r > 0$ such that $P^r(a, b) > 0$ or $P^r(b, a) > 0$.

Notice that every state $y \in [x]$ for an essential class $[x]$ is absorbing if the arrival of the chain, from a state in another communication class, onto any of these states implies that the chain never leaves the essential communication class. Is there always at least an essential collection of states?

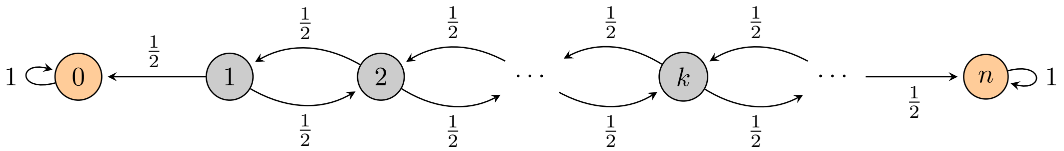

Figure 4: The figure above is the graph for the Gambler's Ruin example from Example 2.1. The states that are colored in orange are essential states. The rest of the states are colored in grey and represent inessential states. Both $[0]$ and $[n]$ form essential communication classes that only include one state each.

Lemma 2.8. *Every finite Markov Chain has at least one essential communication class.*

Proof. Consider a sequence of random variables $(X_0, X_1, \dots)$. Inductively create a sequence of random variables (X_t) such that x is a random variable in our sequence if $X_{t-1} = y$ can go to x in the chain. In other words, $P(X_t = x \mid X_{t-1} = y) > 0$ but $P(X_{t+r} = X_{t-1} \mid X_{t-1} = y) = 0$ for any $r \geq 0$.

Note that $X_t \neq X_{t+k}$ for any $t, k \geq 0$. This is because if there are repeating elements then, there is at least one essential state. However, by Lemma 2.6 this means that all states in the chain are essential, implying the existence of an essential communication class. Then, we are done. Similarly, if we find that any state is essential in the chain then we stop because we have found an essential class.

Assuming that one of the previous situations does not occur, there must be a final state (since we are dealing with finite spaces), which must be essential. Therefore, there is always at least one essential communication class. ■

Now that we have created the distinction between the states $[0] = [1] = \{0, 1\}$ and $[2] = \{2\}$ from Figure 2, and noticed this overlapping property for essential states, we should try to generalize the results from our example to all inessential states. We previously noticed that the inessential states have a zero probability in the limiting distribution. Meaning, that if we were to take samples from the system for a long long time, then we would not expect to be in any inessential state. It's time that we prove this observation.

Proposition 2.9. *Let χ be a finite state space. If π is the stationary distribution for the transition matrix P, then $\pi(y) = 0$ for all inessential states y.*

Proof. Assume that all states are in the finite state space χ. Let C be an essential communication class. From Lemma 2.8, we know that this must always exist. Let $P(C)$ represent the rows of the matrix such that for all $x \in C$, the row $P(x, \cdot)$ is part of $P(C)$. Then,

$$\begin{aligned}
\pi P(C) &= \sum_{z \in C} \pi(z) P(z, \cdot) = \sum_{z \in C} \sum_{y \in \chi} \pi(y) P(y, z) \\
&= \sum_{z \in C} \Bigg[\sum_{y \in C} \pi(y) P(y, z) + \sum_{y \notin C} \pi(y) P(y, z) \Bigg] \\
&= \sum_{y \in C} \pi(y) \sum_{z \in C} P(y, z) + \sum_{z \in C} \sum_{y \notin C} \pi(y) P(y, z) \\
&= \pi(C) + \sum_{z \in C} \sum_{y \notin C} \pi(y) P(y, z). \qquad (2.9)
\end{aligned}$$

Notice that for any essential class, the sum of all possibilities in the transition matrix is stochastic. Therefore, $\sum_{z \in C} P(y, z) = 1$. In Equation 2.9, $\pi(C)$ represents all entries in the stationary distribution $\pi(x)$ such that $x \in C$. Thus, $\pi(C) = \pi P(C)$. In Equation 2.9, this implies that the $\pi(y)P(y, z) = 0$ for all $z \in C$ and $y \notin C$.

Suppose y_0 is an inessential state. Then, there exists a sequence of events $y_0, y_1, y_2 \dots$ such that for all $k \geq 0$, y_k is an inessential state. Such a creation of a sequence of events is detailed in Lemma 2.8. Eventually, there exists $r > 0$ such that y_r is an essential state. Since we know that $\pi(y_{r-1})P(y_{r-1}, y_r) = 0$ and $P(y_{r-1}, y_r) > 0$, it is implied that $\pi(y_{r-1}) = 0$. However,

$$0 = \pi(y_{r-1}) = \sum_{z \in \chi} \pi(z) P(z, y_{r-1})$$

Therefore, $\pi(z)P(z, y_{r-1}) = 0$ for all $z \in \chi$. Thus, $\pi(y_{r-2}) = 0$ since $P(y_{r-2}, y_{r-1}) > 0$. Using the same reasoning, we can inductively travel backward along the chain until we receive that $\pi(y_0) = 0$ for any inessential state y_0. ■

Distinguishing between these types of states allows for easier manipulation of systems, especially when simulations are run on computers. Notice that the proposition does not make any claims about the existence or the uniqueness of the stationary distribution. Another extension of Figure 2 is the following figure.

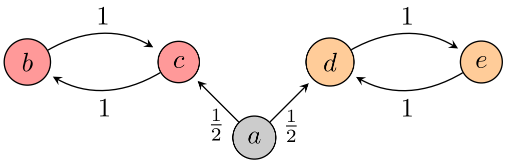

Figure 5: Process with Two Stationary Distributions

The above figure contains more than one essential class. If we were to imagine starting at state a, we would be forced into one of the absorbing states, which, in general, could have limiting different distributions. In the context of studying limiting distributions, it becomes important to realize that inessential states are very bad candidates for limiting distribution analysis. In cases where there are multiple absorbing classes, it is much better to divide the entire system into its separate classes and analyzing each one independently for the existence of limiting distributions. Therefore, this chain can be *reduced* into smaller parts. Thus, it would make sense for us to study systems which we can not be reduced further.

Definition 2.10. A Markov Chain is considered **irreducible** if for any two states $x, y \in \chi$ there exists a $t > 0$ such that $P^t(x, y) > 0$.

Notice that irreducible chains eliminate the possibility of having an absorbing state in the system. This is because it should always be possible to get from any state to any other state. Irreducible chains are also just Markov Chains under one essential communication class.

Proposition 2.11. *The transition matrix P has a unique stationary distribution if and only if there is a unique essential communicating class.*

Proof. Suppose there is one unique essential communicating class C. The transition matrix P of a Markov Chain restricted under this communicating class is P_C. Assume that it has an unique stationary probability distribution π_C. Let π be another probability distribution on χ such that $\pi = \pi P$. By Proposition 2.9, we know that for all $y \notin C$, $\pi(y) = 0$. Therefore for $x \in C$,

$$\pi(x) = \sum_{y\in\chi} \pi(x)P(x, y) = \sum_{y\in C} \pi(x)P(x, y) = \sum_{y\in\chi} \pi(x)P_C(x, y).$$

In the above, this shows that π restricted to C is stationary for P_C. However, it is also the same as the following unique distribution

$$\pi(x) = \begin{cases} \pi_c(x), & x \in C \\ 0, & x \notin C. \end{cases}$$

For the sake of argument, suppose that there are distinct communication classes C_1, C_2 for a transition matrix P with a unique stationary distribution π. Once we restrict the matrix to either of these classes, we know that P restricted to that particular class is irreducible. This means that there exists a $r_1 > 0$ of $P_{C_1}^{r_1}(x, y) > 0$ for all $x, y \in C_1$. On the other hand, if we know that $x \in C_1$ and $y \in C_2$ then, $P_{C_1}^{k}(x, y) = 0$ for all $k \geq 1$. However, this implies that the transition matrix has multiple stationary distributions since $\pi = \pi P_{C_1}$ and $\pi = \pi P_{C_2}$. However, there are terms in the transition matrices which are not the same. Therefore, the stationary distribution also can not be the same. Via contradiction, we know that there can not be two communicating classes. ■

With irreducibility, we were able to prove the uniqueness of stationary distribution. However, all of this work would be useless unless we can prove the existence of stationary distributions. To build up to this idea, we introduce another concept of a stationary measure. In essence, it is a building block for a stationary distribution, which is forced to have all its components sum to one. However, a stationary measure does not need to have all of its terms sum to one. It is a solution to $\pi P = \pi$.

Definition 2.12. A **stationary measure** of a Markov Chain (X_t) on state space χ is the expected number of visits to a state y before returning to another state x. More precisely, $\tilde{\pi}$ is a stationary measure when for all $y \in \chi$,

$$\tilde{\pi}(y) = \sum_{t=0}^{\infty} \mathbb{P}_x \Big(X_t = y, \tau_x^+ > t \Big).$$

Proposition 2.13. *Let $\tilde{\pi}$ be the stationary measure on χ. Then,*

(a) If $\mathbb{P}(\tau_x^+ < \infty)_x = 1$, then $\tilde{\pi}$ satisfies $\tilde{\pi} P = \tilde{\pi}$.

(b) If $\mathbb{E}\left[\tau_x^+\right]_x < \infty$, then $\pi = \frac{\tilde{\pi}}{\mathbb{E}\left[\tau_x^+\right]_x}$ is a stationary distribution.

Proof. Let's first check part (a) of this proposition. Using Definition 2.12, for all $y \in \chi$

$$\sum_{x \in \chi} \tilde{\pi}(x) P(x,y) = \sum_{x \in \chi} \sum_{t=0}^{\infty} \mathbb{P}_z \Big(X_t = x, \tau_z^+ > t \Big) P(x,y).$$

Notice that $\{\tau_z^+ \geq t+1\} = \{\tau_z^+ > t\}$ and

$$\mathbb{P}_z(X_t = x, X_{t+1} = y, \tau_z^+ \geq t+1) = \mathbb{P}(X_t = x, \tau_z^+ \geq t+1) P(x,y).$$

Carefully flipping the order of the summation,

$$\sum_{x \in \chi} \tilde{\pi}(x) P(x,y) = \sum_{t=0}^{\infty} \mathbb{P}_z \Big(X_{t+1} = y, \tau_z^+ \geq t+1 \Big).$$

$$\sum_{x \in \chi} \tilde{\pi}(x) P(x,y) = \sum_{t=1}^{\infty} \mathbb{P}_z \Big(X_t = y, \tau_z^+ \geq t \Big). \tag{2.10}$$

This Equation 2.10 can be simplified further as follows.

$$\begin{aligned} \sum_{t=1}^{\infty} \mathbb{P}_z \Big(X_t = y, \tau_z^+ \geq t \Big) &= \tilde{\pi}(y) - \mathbb{P}_z(X_0 = y, \tau_z^+ > 0) + \sum_{t=1}^{\infty} \mathbb{P}(X_t = y, \tau_z^+ > t)_x && (2.11) \\ &= \tilde{\pi}(y) - \mathbb{P}_z(X_0 = y, \tau_z^+ > 0) + \mathbb{P}_z(X_{\tau_z^+} = y) && (2.12) \\ &= \tilde{\pi}(y). && (2.13) \end{aligned}$$

The final equality is reached when one considers the following two cases: $y = z$ and $y \neq z$. When $y = z$, the last two terms are both 1 and cancel each other out. When $y \neq z$, both of the terms are zero. Therefore, Equation 2.10 with the previous result can be combined to show that $\tilde{\pi} P = \tilde{\pi}$. Notice that $\mathbb{E}\left[\tau_z^+\right]_z = \sum_{x \in \chi} \tilde{\pi}(x)$. Therefore, we can normalize a stationary measure to get a stationary probability distribution.

$$\pi(x) = \frac{\tilde{\pi}(x)}{\mathbb{E}\left[\tau_z^+\right]}. \tag{2.14}$$

■

The conditions displayed in the proposition can also be proven to be true for irreducible chains.

Lemma 2.14. *For any states $x, y \in \chi$, of an irreducible chain $\mathbb{E}\left[\tau_y^+\right]_x < \infty$ and $\mathbb{P}_x(\tau_y^+ < \infty) = 1$.*

Proof. By the definition of irreducibility, there exists a $r > 0$ and a real $\epsilon > 0$ such that for any states $a, b \in \chi$, there exists a $j \leq r$ with $P^j(a,b) \leq \epsilon$. This is the equivalent of making an upper bound that represents moving from a current state to any other state. We can partition the chain into several parts based on its current location $t \geq 0$. For any value of X_t, the probability of hitting a state y between t and $t + r$ is at least ϵ. Therefore, for a finite $k \geq 0$

$$\mathbb{P}_x(\tau_y^+ > kr) = \mathbb{P}_x(\tau_y^+ > kr \mid\mid \tau_y^+ < (k-1)r)\ \mathbb{P}_x(\tau_y^+ > (k-1)r) \leq (1-\epsilon)\ \mathbb{P}_x(\tau_y^+ > (k-1)r).$$

However this implies that,

$$\mathbb{P}_x(\tau_y^+ > kr) \leq (1-\epsilon)^2 \mathbb{P}_x(\tau_y^+ > (k-2)r).$$

Therefore,

$$\mathbb{P}_x(\tau_y^+ > kr) \leq (1-\epsilon)^k.$$

Recall that whenever Y is an integer-valued non-negative random variable, $\mathbb{E}[Y] = \sum_{t \geq 0} \mathbb{P}(Y > t)$. Thus,

$$\mathbb{E}\left[\tau_y^+\right]_x = \sum_{t \geq 0} P_x(\tau_y^+ > t) \leq r \sum_{k \geq 0} P_x(\tau_y^+ > kr) \leq r \sum_{k \geq 0} (1-\epsilon)^k < \infty.$$

Since for all $t \geq 0$ we know that $P_x(\tau_y^+ > t) \leq 1$, and the probability $P_x(\tau_y^+ > kr) \leq (1-\epsilon)^k$,

$$1 \geq P_x(\tau_y^+ < kr) \geq 1 - (1-\epsilon)^k.$$

Notice that as $k \to \infty$, the right-hand inequality becomes smaller. Therefore,

$$\lim_{k \to \infty} 1 \geq \lim_{k \to \infty} P_x(\tau_y^+ < kr) = P_x(\tau_y^+ < \infty) \geq \lim_{k \to \infty} 1 - (1-\epsilon)^k = 1.$$

■

With all of the above being conducted, we need to show one final piece of the puzzle so that we can prove the existence of stationary distributions. Now we will present harmonic functions.

Definition 2.15. A function $h : \chi \to \mathbb{R}$ is **harmonic** at x if

$$h(x) = \sum_{y \in \chi} P(x,y) h(y).$$

Similarly, a function is considered harmonic on $D \subset \chi$ if it is harmonic at every state $x \in D$. If h is a column vector, then the harmonic function h, which is harmonic on all of χ, also satisfies $h = Ph$ for transition matrix P of a Markov Chain (X_t) on χ.

Lemma 2.16. *Let χ be a finite state space. Suppose that P is a transition matrix of an irreducible chain (X_t) on χ. Then, a function h which is harmonic at every point in state space χ is constant.*

Proof. Since χ is finite, there must be a maximum state $x \in \chi$ such that $h(x) = M$ is the largest value of the function. That is, $h(x) \geq h(z)$ for all $z \in \chi$. If for any state z there is a real $\delta > 0$ such that $P(x,z) > \delta$ and $h(z) < M$, then

$$h(x) = P(x,z)h(z) + \sum_{y \neq z} P(x,y)h(y) \leq P(x,z)h(z) + \sum_{y \neq z} P(x,y)h(x) \leq \delta h(z) + (1-\delta)h(x) < M.$$

Notice that the last part of the inequality is true because it can be rewritten as

$$h(x) - \delta(h(x) - h(z)) = M - \delta(h(x) - h(z)) < M$$

where $\delta(h(x) - h(z)) > 0$. However, this is a contradiction because we assumed that $h(x) = M$. Therefore, $h(z) = M$ for all states z such that $P(x,z) > 0$.

Since the chain is irreducible, if there is another state than z in χ then, there should be at least one state $y \in \chi$ such that $P(z, y) > 0$. In fact, we know there must be a chain $X_0, X_1, \dots, X_t = y$ where $P(X_t, X_{t+1}) > 0$. Starting from the end of the Markov Chain, we can inductively travel backward to check all

$$X_t = X_{t-1} = \cdots = X_0 = h(y) = M$$

with the same reasoning as above. Therefore, the harmonic function must be constant. ■

With all of the above work, we were able to gain an understanding of different types of Markov Chains. We looked at pitfalls of certain chains and categorized them based on their properties. We also converted some of the observations from our examples into concrete understanding by proving when those observations hold. We now have enough to prove the existence of unique stationary distributions for a Markov Chain.

Corollary 2.17. *Let P be a transition matrix of an irreducible Markov Chain. There exists a unique probability distribution π such that $\pi = \pi P$.*

Proof. By Lemma 2.14, we know that the hypothesis for Proposition 2.13 is true for all irreducible Markov Chains. Therefore, a stationary measure exists and thus a stationary distribution can be constructed following the method described in Proposition 2.13. Lemma 2.16 implies that the kernel of $P - I$ has a dimension of 1, so the column rank of $P - I$ is $|\chi| - 1$. In this situation, I represents the identity matrix. Let's suppose there is a probability distribution ν such that $\nu = \nu P$ and a probability distribution $\pi = \pi P$ then $\pi = \nu$. This is because the equation $\nu = \nu P$ has a one-dimensional space of solutions and only contains one vector whose entries sum to one. Since the space only has one unique solution, any distribution $\pi = \pi P$ must also contain the same vector whose entries sum to one. ■

Harmonic functions are also useful for studying other types of chains. Functions on finite state spaces have led to amazing results. One interesting example is Kirchhoff's Node Law, which is discussed in [7] under the context of Markov Chain theory.

For irreducible chains, we have already seen instances in which the transition between any state is possible. However, we could generalize this to a greater extent with Markov Chains which can travel from any state to any other state for a fixed number of steps. In other words, starting from any given state there should be some $t \geq 0$ number of steps one could take so that it is possible to move from one state to another state.

Definition 2.18. Let $\mathcal{T}(x) = \{t \geq 1 : P^t(x, x) > 0\}$ be the set of all possible return times for a Markov Chain with transition matrix P. The **period** of the state x is $\gcd\{\mathcal{T}(x)\}$. A chain is considered **aperiodic** if all states in the chain have a period of 1. If the chain is not aperiodic then, the chain is considered periodic.

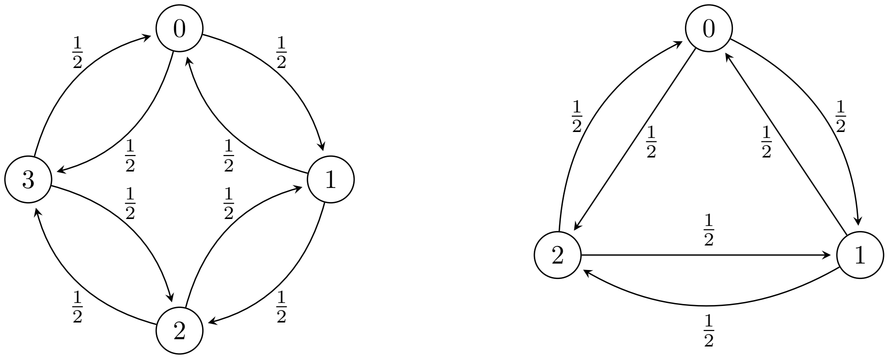

Figure 6: The above are examples of N-cycle random walks on $\mathbb{Z}_n$, where $\mathbb{Z}_n$ are the integers mod n. This figure displays the walks for $n = 4$ on the left and $n = 3$ on the right. To move from one state to another, flip a coin. If we start on state k then, we move to state $(k + 1) \mod n$ on a heads and move to state $(k - 1) \mod n$ on a tails. Notice that $n = 3$ is aperiodic and $n = 4$ has a period of 2.

Proposition 2.19. *If P is an irreducible transition matrix of a Markov Chain on state space χ, then the period of any two states $x, y \in \chi$ are the same.*

Proof. Fix two states x and y. Since the transition matrix is irreducible, there exist integers $j > 0$ and $k > 0$ such that $P^j(x, y) > 0$ and $P^k(y, x) > 0$. Then, $P^{j+k}(x, x) > 0$ and $P^{j+k}(y, y) > 0$. Therefore, $(j + k) \in (\mathcal{T}(x) \cap \mathcal{T}(y))$. For any $i \in \mathcal{T}(x)$, we also know that $(i + j + k) \in \mathcal{T}(x)$. However, we also know that $(j + i + k) \in \mathcal{T}(y)$. Let $\mathcal{T}'(y) = \{x_0 - (j + k) : x_0 \in \mathcal{T}(x)\}$. Notice that $\mathcal{T}(x) \subseteq \mathcal{T}'(y)$, which implies that $\gcd(\mathcal{T}(y)) \leq \gcd(\mathcal{T}(x))$. By a similar argument, we find that $\gcd(\mathcal{T}(x)) \leq \gcd(\mathcal{T}(y))$. Therefore, the period of any state is the same in an irreducible chain. ■

Proposition 2.20. *Let χ be a finite state space. If P is an aperiodic and irreducible transition matrix, then there is an integer $r_0 > 0$ such that $P^r(x, y) > 0$ for all states $x, y \in \chi$ where $r \geq r_0$.*

Proof. Since P is an irreducible transition matrix, for any $x, y \in \chi$ there exists a $m > 0$ such that $P^m(x, y) > 0$. In addition, by Proposition 2.19, we also know that the period of all states must be 1. Since P is also aperiodic, the collection $\mathcal{T}(x)$ has at least two relatively prime return times. Take $p > 0$ and $q > 0$ to be these return times. Note that all of the linear combinations $(np + lq) \in \mathcal{T}(x)$ for all $n > 0$ and $l > 0$. By the Chicken McNugget Theorem, take $r_0 = \max(pq - p - q + 1)$ for all p, q for every unordered finite pair (x, y). and any $r \geq r_0$ implies that $P^r(x, y) > 0$, because $r \in \mathcal{T}(x)$. ■

3 The Main Show

This entire section will be devoted to proving the main buildup to this paper. First we present the Strong Law of Large numbers, which will be important in the later proof of the Convergence and Ergodic theorems.

Theorem 3.1 (Strong Law of Large Numbers). *For a sequence of random variables $X_1, X_2, \ldots$ such that $\mathbb{E}[X_i] = 0$ and*

$$\mathrm{Var}[X_{i+1} + \cdots + X_{i+k}] \leq Ck$$

for some $C > 0$ and all i and k, then

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} X_i = \mathbb{E}[X_i] = 0\right) = 1. \tag{3.1}$$

Proof. Notice that for a finite value $z > 0$, this sum can be split into different parts

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} X_i = 0\right) = \mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{z^2-1} X_i + \lim_{t\to\infty} \frac{1}{t} \sum_{i=z^2}^{t-1} X_i = 0\right).$$

For a given t, consider m such that $m^2 \leq t \leq (m+1)^2$. Let $A_t = \frac{1}{t}\sum_{i=0}^{t-1} X_i$. Notice this average can be split up similarly as above. Let $B_t = \frac{1}{t}\sum_{i=m^2}^{t-1} X_i$. Now rewriting

$$A_t = \frac{1}{t}\left(m^2 A_{m^2} + B_t\right).$$

Looking at the expected value of A_{m^2}, it is enough to show that A_t is zero by showing that the individual parts of this sum are always zero. This can be done by studying the expected values of the random variables. Recall that $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ for a random variable X.

$$\mathbb{E}\left[A_{m^2}^2\right] = \frac{\mathbb{E}\left[(\sum_{i=0}^{m^2-1} X_i)^2\right]}{t^2} = \frac{\mathrm{Var}\left[(\sum_{i=0}^{m^2-1} X_i)^2\right]}{t^2} \leq \frac{Cm^2}{t^2} \leq \frac{Ct}{t^2} = \frac{C}{t}.$$

By constructing this upper bound, it is shown that $\mathbb{E}\left[\sum_{m=0}^{\infty} A_{m^2}^2\right] < \infty$ since each of the terms has a decreasing upper bound. In other words, as t becomes larger we notice that the upper bound becomes smaller.

Since this expected value of the sum is finite, it is implied that

$$\mathbb{P}\left(\sum_{m=1}^{\infty} A_{m^2}^2 < \infty\right) = 1, \quad \mathbb{P}\left(\lim_{m\to\infty} A_{m^2} = 0\right) = 1.$$

Notice that $\lim_{t\to\infty} \frac{m^2}{t} = 1$. This is true because we could consider the inequality

$$m^2 \leq t \leq (m+1)^2$$

$$\frac{m^2}{t} \leq \frac{t}{t} = 1 \leq \frac{(m+1)^2}{t}$$

$$\lim_{t\to\infty} \frac{m^2}{t} \leq \lim_{t\to\infty} 1 = 1 \leq \lim_{t\to\infty} \frac{(m+1)^2}{t}$$

$$\lim_{t\to\infty} \frac{m^2}{t} \leq \lim_{t\to\infty} 1 = 1 \leq \lim_{t\to\infty} \frac{(m^2+2m+1)}{t}$$

$$\lim_{t\to\infty} \frac{m^2}{t} \leq \lim_{t\to\infty} 1 = 1 \leq \lim_{t\to\infty} \frac{(m^2+2m+1)}{t} \leq \lim_{t\to\infty} \frac{(m^2+2\sqrt{t}+1)}{t}$$

$$\lim_{t\to\infty} \frac{m^2}{t} \leq \lim_{t\to\infty} 1 = 1 \leq \lim_{t\to\infty} \frac{(m^2+2m+1)}{t} \leq \lim_{t\to\infty} \frac{(m^2)}{t}$$

The left-hand inequality happens to collapse into the same limit as we see on the right-hand side. Since the same limit bounds each side, we get that the limit evaluates to 1.

Notice that larger values of t imply larger and larger values of m due to the restriction placed upon m and therefore,

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t} m^2 A_{m^2} = 0\right) = 1.$$

Consider B_t similarly as seen above.

$$\mathbb{E}\left[{B_t}^2\right] = \frac{\mathbb{E}\left[\sum_{i=m^2}^{t-1} X_i\right]}{t^2} = \frac{\operatorname{Var}\left[\sum_{i=m^2}^{t-1} X_i\right]}{t^2} \leq \frac{2Cm}{t^2} \leq \frac{2C}{t^{\frac{3}{2}}}.$$

Therefore, $\mathbb{E}\left[\sum_{t=0}^{\infty} B_t^2\right] < \infty$. In the same way as before, this implies that the individual terms in this sum are heading towards zero, otherwise, this infinite sum would not have a finite value. Thus,

$$\mathbb{P}\left(\lim_{t\to\infty} B_t = 0\right) = 1.$$

$$\blacksquare$$

As previously mentioned in section 2, we were interested in studying the space of a probability distribution. One aspect of this is defining a real-valued function f over this state space and inspecting how this function behaves over the entire space. What we want to study is the expected value of this function over a probability distribution π, which is denoted as $\mathbb{E}\left[f\right]_\pi$. Formally, on a finite state space χ,

$$\mathbb{E}\left[f\right]_\pi = \sum_{x\in\chi} f(x)\pi(x).$$

Theorem 3.2 (Ergodic Theorem)**.** *Let $f : \chi \to \mathbb{R}$ be a function on the finite state space χ. If (X_i) is an irreducible chain with stationary distribution π, then for any starting distribution μ*

$$\mathbb{P}\left(\lim_{t\to\infty} \frac{1}{t} \sum_{i=0}^{t-1} f(X_i) = \mathbb{E}\left[f\right]_\pi\right) = 1.$$

Proof. Without loss of generality, assume that the Markov Chain (X_i) starts at the state $x \in \chi$.

Define $\tau_{x,0}^+ = 0$. Consider the hitting times for the starting state $\tau_{x,k-1}^+ = \min\{t > \tau_{x,k-1}^+ : X_t = x\}$. Notice that the terms that come between these times are identical to each other. This is because the chain comes back to the same state that it originally started on. Since this chain is also a Markov Chain, we notice that every time this chain "starts again" it is going to be identical and equally distributed as it was previously. This partitions the chain into different pieces, based on the hitting times of state x. Thus, if

$$Y_k = \sum_{i=\tau^+_{x,k-1}}^{\tau^+_{x,k}-1} f(X_i),$$

then (Y_k) is independent and identically distributed to $(Y_{k-1}), (Y_{k-2}), \ldots, (Y_1)$. We have previously seen that the $\mathbb{E}\left[\tau^+_{x,k}\right]_x < \infty$ and know that χ is a finite state space. Therefore, there must exist a finite upper bound $U = \max_{y\in\chi} |f(y)| < \infty$ such that the $\mathbb{E}\left[|Y_1|\right] \leq U\mathbb{E}\left[\tau^+_{x,1}\right]_x < \infty$. Let $S_t = \sum_{i=0}^{t-1} f(X_i)$ and $S_{\tau^+_{x,k}} = \sum_{i=\tau^+_{x,k-1}}^{\tau^+_{x,k}-1} Y_i$. By Theorem 3.1,

$$\mathbb{P}_\mu\left(\lim_{n\to\infty} \frac{S_{\tau^+_{x,n}}}{n} = \mathbb{E}\left[Y_1\right]_x\right) = 1.$$

Then, also notice that the hitting time $\tau^+_{x,n} = \sum_{k=1}^{n}(\tau^+_{x,k} - \tau^+_{x,k-1})$. Therefore, we can again use Theorem 3.1 to find that

$$\mathbb{P}_\mu\left(\lim_{n\to\infty} \frac{\tau^+_{x,n}}{n} = \mathbb{E}\left[\tau^+_{x,1}\right]_x\right) = 1.$$

Therefore, we also know that

$$\mathbb{P}_\mu\left(\lim_{n\to\infty} \frac{S_{\tau^+_{x,n}}}{\tau^+_{x,n}} = \frac{\mathbb{E}\left[Y_1\right]_x}{\mathbb{E}\left[\tau^+_{x,1}\right]_x}\right) = 1.$$

Notice that the left-hand side of this limit is the same as proposed in the statement of this theorem. For the right-hand side, realize that

$$\begin{aligned}
\mathbb{E}\left[Y_1\right]_x &= \mathbb{E}\left[\sum_{i=0}^{\tau^+_{x,1}-1} f(X_i)\right]_x \\
&= \mathbb{E}\left[\sum_{y\in\chi} f(y) \sum_{i=0}^{\tau^+_{x,1}-1} \mathbf{1}_{\{y=x\}}\right]_x \\
&= \sum_{y\in\chi} f(y)\mathbb{E}\left[\sum_{i=0}^{\tau^+_{x,1}-1} \mathbf{1}_{\{y=x\}}\right]_x \\
&= \sum_{y\in\chi} f(y)\tilde{\pi}(y) \\
&= \mathbb{E}\left[f\right]_\pi \mathbb{E}\left[\tau^+_{x,1}\right]_x .
\end{aligned}$$

Thus,

$$\mathbb{P}_\mu\left(\lim_{n\to\infty} \frac{S_{\tau^+_{x,n}}}{\tau^+_{x,n}} = \frac{\mathbb{E}\left[Y_1\right]_x}{\mathbb{E}\left[\tau^+_{x,1}\right]_x} = \mathbb{E}\left[f\right]_\pi\right) = 1.$$

■

Corollary 3.3. *Taking $f(y) = \mathbf{1}_{\{y=x\}}$ in Theorem 3.2, shows that*

$$\mathbb{P}_\mu = \left(\lim_{t\to\infty} \frac{1}{t}\sum_{i=0}^{t-1} \mathbf{1}_{\{y=x\}} = \pi(x)\right) = 1.$$

This explains why traveling along a Markov Chain, again and again, will become the stationary distribution when considering starting at an arbitrary state. In addition, we found that this limiting distribution will always exist for chains that satisfy the assumption made for the Ergodic Theorem. In actual simulations, we use a finite approximation to the actual limiting distributions because we can not sample infinitely many times. Therefore, it becomes important for us to also show that taking more steps along this chain will get us progressively closer to the limiting distribution. This requires another idea, the Convergence Theorem. Let's first define what "getting closer" means for our distribution with respect to another distribution, which will be the limiting distribution for our purposes.

Definition 3.4. Intuitively, we define the **Total Variation Distance** between two probability distributions μ and ν on a state space χ by

$$d(\mu, \nu)_{TV} = \max_{A \subseteq \chi} ||\mu(A) - \nu(A)||.$$

Informally, we are looking at the largest difference between these probability distributions.

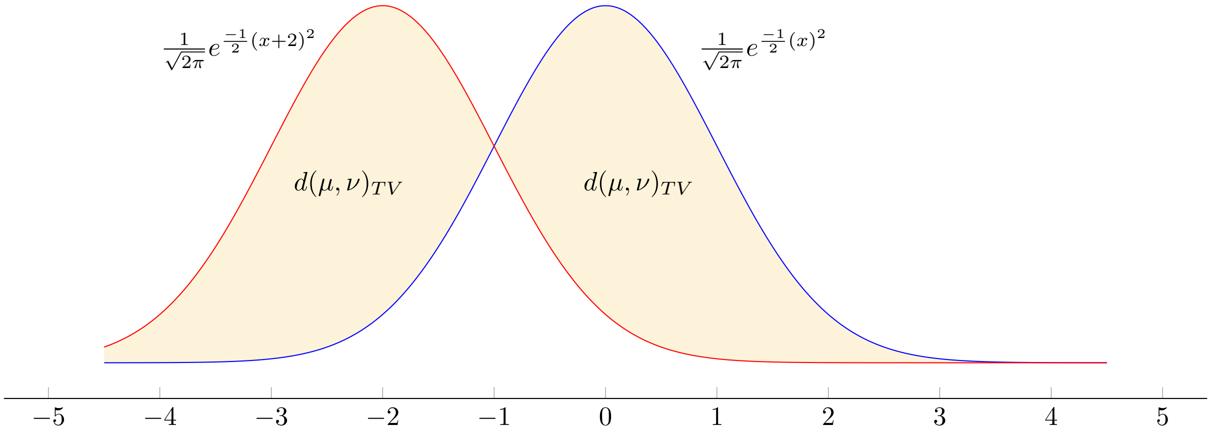

Figure 7: Maximum Distance between Probability Distributions
Looking at the maximum distance between probability distribution between two normal distributions. Due to its unique symmetry structure, this function has a more symmetric outline. The area witnessed between these two curves and the overlapping area must be the same area, which represents the total variation distance between these two probability distributions.

This is a very intuitive way to define the distance between two distributions. However, this is not always computationally the method in which we want to look at this distance between distributions. Therefore, it becomes important to look at alternate, equivalent definitions of this distance which are easier to compute.

Lemma 3.5. *For two probability distributions μ and ν on a state space χ,*

$$d(\mu, \nu)_{TV} = \frac{1}{2} \sum_{x \in \chi} |\mu(x) - \nu(x)|.$$

Proof. Let $B = \{x : \mu(x) \geq \nu(x)\}$. Let $A \subset \chi$ be an event in the probability distributions. For any $x \in A \cap B^c$, notice that the probability must be $\mu(x) - \nu(x) < 0$. This means that the total contribution from the event A is being decreased here. Since $A = (A \cap B) \cup (A \cap B^c)$, we can remove the terms from the complement and receive the following inequality,

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B).$$

However, also notice that $\mu(B) - \nu(B) > 0$. Therefore, adding more elements from this set cannot decrease the distance between these distributions. Thus,

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B).$$

The same argument implies $\nu(A) - \mu(A) \leq \nu(B) - \mu(B)$. Next, notice that $\sum_{x\in\chi} \nu(x) - \mu(x) = 0$. This is because both of these are probability distributions that sum to one over the entire state space. However, we also know that

$$\begin{aligned}\sum_{x\in\chi} \mu(x) - \nu(x) = \sum_{x\in B} \mu(x) - \nu(x) + \sum_{x\in B^c} \mu(x) - \nu(x) = 0\\ \mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)\\ |\mu(B) - \nu(B)| = |\mu(B^c) - \nu(B^c)|.\end{aligned}$$

This fact is the formalization of our observation in the figure above. Now, let $A = B$ (or B^c), then

$$\begin{aligned}|\mu(A) - \nu(A)| &= \frac{1}{2}|2(\mu(B) - \nu(B))|\\ &= \frac{1}{2}|\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)|\\ &= \frac{1}{2}\sum_{x\in\chi} |\mu(x) - \nu(x)|.\end{aligned}$$

■

Theorem 3.6 (Convergence Theorem)**.** *Suppose P is an irreducible and aperiodic transition matrix, with stationary distribution π on a finite state space χ. Then for all $t > 0$, there exist constants $a \in (0, 1)$ and $C > 0$ such that*

$$\max_{x\in\chi}\{d(P^t(x, \cdot), \pi)_{TV}\} \leq Ca^t.$$

Proof. Since P is irreducible and aperiodic, by Proposition 2.20, there exists an r such that $P^r(x, y) > 0$ for any $x, y \in \chi$. Let Π be a stochastic matrix with $|\chi|$ rows of π. Notice there exists a sufficiently small $\delta > 0$,

$$P^r(x, y) \geq \delta\pi(y)$$

for all $x, y \in \chi$. Let $\theta = 1 - \delta$. Define a stochastic matrix Q, such that $P^r(x, y) = (1 - \theta)\Pi + \theta Q$. For any stochastic matrix M, $M\Pi = \Pi$. Also, for any Matrix such that $\pi M = \pi$ we find that $\Pi M = \Pi$. These conditions follow directly from the assumptions about the matrices, so proving this fact is left to the reader.

We will next aim to use induction to prove that $P^{rk} = (1 - \theta^k)\Pi + \theta^k Q^k$. Notice that $k = 1$ holds by the previous construction. Now assume that $k = n$ implies that $P^{rn} = (1 - \theta^n)\Pi + \theta^n Q^n$.

Then, we know that

$$\begin{aligned}P^{r(n+1)} &= P^{rn}P^r\\ &= ((1 - \theta^n)\Pi + \theta^n Q^n)P^r\\ &= \Pi P^r - \theta^n\Pi P^r + \theta^n Q^n((1 - \theta)\Pi + \theta Q)\\ &= \Pi - \theta^n\Pi + \theta^n Q^n(\Pi - \theta\Pi + \theta Q)\\ &= \Pi - \theta^n\Pi + \theta^n\Pi - \theta^{n+1}\Pi + \theta^{n+1}Q^{n+1}\\ &= \Pi - \theta^{n+1}\Pi + \theta^{n+1}Q^{n+1}\\ &= (1 - \theta^{n+1})\Pi + \theta^{n+1}Q^{n+1}.\end{aligned}$$

By induction, we have now shown that $P^{rk} = (1 - \theta^k)\Pi + \theta^{k+1}Q^{k+1}$. Therefore,

$$\begin{aligned}P^{rk+j} &= ([1 - \theta^n]\Pi + \theta^n Q^n)P^j\\ &= \Pi P^j - \theta^n\Pi P^j + \theta^n Q^n P^j\\ &= \Pi + \theta^n(Q^n P^j - \Pi).\end{aligned}$$

Thus, $P^{rk+j} - \Pi = \theta^n(Q^n P^j - \Pi)$. This means that for any $x \in \chi$,

$$\frac{1}{2}|P^{rk+j}(x,\cdot) - \pi(x)| = \frac{1}{2}|\theta^n(Q^n P^j(x,\cdot) - \Pi))|.$$

We can take the sum on both sides to signify the same sum since we are adding the same terms together.

$$\frac{1}{2}\sum_{x\in\chi}|P^{rk+j}(x,\cdot) - \pi(x)| = \frac{1}{2}\sum_{x\in\chi}|\theta^n(Q^n P^j(x,\cdot) - \Pi))|.$$

However, $|Q^n P^j(x,\cdot) - \pi| \leq 1$ since that is the largest distance between two probability distribution values. Therefore,

$$d(P^{rk+j}(x,\cdot) - \pi)_{TV} \leq \theta^k.$$

This is the equivalent of taking $C = (\frac{1}{\theta})^j$ and $a = \theta^{\frac{1}{r}}$. ■

4 Markov Chains Monte Carlo

So far we have been discussing Markov Chains to sample from a system so that we could better understand that system. Now we will proceed to look at Markov Chains with a different perspective, one aimed to create a given distribution stationary by changing the transition matrix. What does that mean? What does that look like? Why is that useful? This section will continue onwards with these essential questions in mind.

In the sections prior, Markov Chains have been used on systems whose transition matrices we previously knew. However, this is an unlikely scenario in many real-world applications. In fact, we may even be able to calculate or have a grasp of a potential stationary distribution. The idea of Monte Carlo is to use Markov Chains to sample according to non-uniform distributions. This algorithm is especially efficient in spaces with higher dimensions because it is independent of the dimension of the space. Let π be the distribution on the finite state space χ whose transition matrix we want to find.

1. Start with an initial transition matrix P. The closer this transition matrix has a stationary distribution to π, the less computational resources will be required for this algorithm.

2. Choose an initial state $x \in \chi$.

3. Take samples from the chain and consider every candidate z with an acceptance probability $a(x,z)$. This acceptance probability of the next random variable $a(x,z) = \min\{1, \frac{\pi(x)P(y,x)}{\pi(y)P(x,y)}\}$.

4. Whenever z is not accepted as the next candidate, we take the original state x as the next random variable in the sequence.

Repeating this process many, many times would lead us to a transition matrix such that for all $x, z \in \chi$ the new transition matrix P' is such that

$$P'(x,z) = \begin{cases} P(x,z)\min\left(1, \dfrac{\pi(z)P(z,x)}{\pi(x)P(x,z)}\right), & x \neq z \\ 1 - \sum_{y\neq x} P'(x,y), & x = z. \end{cases}$$

We will also check if the probability distribution to this Markov Chain should be stationary. This algorithm is known as the **Metropolis-Hastings Algorithm**.

Lemma 4.1. *A probability distribution π is stationary for a transition matrix P' generated via the Metropolis-Hastings Algorithm.*

Proof. Let χ be the state space for the Markov Chain. With loss of generality, for $x, y \in \chi$, assume that $\pi(y)P(y,x) \geq \pi(x)P(x,y)$. Then, the probability $P(x,y) = P'(x,y)$. This implies that

$$\begin{aligned}\pi(x)P'(x,y) &= \pi(x)P(x,y)\\ &= \pi(y)P(y,x)\frac{\pi(x)P(x,y)}{\pi(y)P(y,x)}\\ &= \pi(y)P'(y,x).\end{aligned}$$

Therefore, we also find that this probability distribution must be stationary.

$$\begin{aligned}\sum_{y\in\chi}\pi(y)P(y,x) &= \sum_{y\in\chi}\pi(x)P(x,y)\\ &= \pi(x)\sum_{y\in\chi}P(x,y)\\ &= \pi(x).\end{aligned}$$

■

Some problems involving MCMC require completely independent sampling from the desired distribution. Therefore, a technique employed is to run a Markov Chain enough times so that the probability of drawing any state is that of the desired distribution, take note of that state, and then run the chain many more times before recording the next data point, so that the new sample is practically independent of the previous one and with a probability according to the stationary distribution. However, when using MCMC for integration, the Ergodic Theorem tells us that we can record every data point (even if they are all directly related to the previous data point, and therefore not necessarily independent), and the overall mean will still equal the expectation of the function according to the desired distribution, with probability 1. Consequently, it is unnecessary to use expansive computational resources while not collecting data, thus making the computation of the integral much faster. This result will be employed in the example code explained in the next section, which also describes a potential usage of the Metropolis-Hastings Algorithm.

5 Monte Carlo Integration

An interesting extension of the previous discussion on Markov Chain theory is the usage of Markov Chains with integration. When looking at functions that are well-defined over multidimensional regions, there are many instances when it becomes difficult to calculate the integral of the function analytically. Recall that

$$\mathbb{E}\left[f\right] = \frac{1}{\text{vol(V)}}\int_V f.$$

where the expected value of the function represents the average value of the function over a region $V \subseteq R^n$ for a n dimensional object. Rearranging the terms from this equation leads to the following identity

$$\int_V f = \text{vol(V)}\ \mathbb{E}\left[f\right]. \tag{5.1}$$

Definition 5.1. The **Monte Carlo Estimator** of a function sampled over an integer $q > 0$ points is

$$F_q = \frac{\text{vol(v)}}{q}\sum_{i=1}^{q} f(x_i),$$

where x_i are all independent data points sampled over a uniform distribution.

The idea is to sample N points from a uniform distribution of all points inside the region V. The region V itself can be an arbitrary volume, however, it should be chosen to be volumes that are already known or can be easily computed. Each time we sample a new point from our distribution, we check whether any randomly selected point will be in the region. An analogue to the limit definition, we have a finite sense of what an integral can be, where sampling points in a finite sense gets us closer to the actual area of the curve. Notice that however we finitely approximate this value it should directly follow that

$$\mathbb{P}\left(\lim_{N\to\infty} F_N = \int_V f\right) = 1.$$

If we were to look back at Theorem 3.1, we find that looking at the following finite estimation meets exactly this criteria. By looking at which property we wanted our estimator to contain, we fond an exact definition that we could use from a structure which we had already studied.

However, this definition is more intuitive and only focused on the uniform distribution of the points. This can be useful in many circumstances but is not enough to cover non-uniform distributions. The estimator over any distribution μ is defined to be

$$F_q^\mu = \frac{1}{q}\sum_{i=1}^{q} \frac{f(x_i)}{\mu(x_i)},$$

where all points x_i are sampled independently according to the distribution μ. Note that on a continuous plane over a uniform distribution $\mu(x_i) = \frac{1}{\text{vol}(v)}$ for all x_i. For the sake of completion, let us make sure that this does follow the property that we need.

Lemma 5.2. *The expected value of F_q^μ is the $\int_V f$ for the function f on the region $V \subseteq R^n$.*

Proof.

$$\begin{aligned}\mathbb{E}\left[F_q^\mu\right]_\mu &= \mathbb{E}\left[\frac{1}{q}\sum_{i=1}^{q}\frac{f(x_i)}{\mu(x_i)}\right]_\mu \\ &= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{q}\frac{f(x_i)}{\mu(x_i)}\right]_\mu \\ &= \frac{N}{N}\int_V\left[\frac{f(x_i)}{\mu(x_i)}\mu(x_i)\right]dx_i \\ &= \int_V f.\end{aligned}$$

■

This paper will end with an example of the usage of this algorithm. To make it easier to follow, we will use the Monte Carlo Integration technique to approximate the value of π. Typically, these methods are implemented in higher dimensions and for curves that are harder to solve analytically. However, this presents a good first-time introduction to this topic.

Example 5.1. Let $f(x) = \sqrt{1-x^2}$. Through Monte Carlo Integration, we aim to approximate

$$\int_{-1}^{1}\sqrt{(1-x^2)}dx.$$

Let's first draw a boundary box around the intended area.

Now, we can repeatedly sample from our larger rectangle from a uniform distribution. The more we sample, the closer we get to the actual area of this curve. After a sample of $N = 100$ points, we find our technique approximates the area at around 1.64. After $N = 1000$ points, the area under the curve is approximated at 1.554. If we make this sampling much larger, at $N = 10^6$, we find that the area under the curve is approximately 1.56978. In comparison, note that the actual area of this curve is $\frac{\pi}{2} \approx 1.5707$. In fact, we could also use the value we obtain from this sampling method to approximate the value of π, since we know that $\frac{\pi}{2}$ is the actual area of this curve. With the highest sample size, we find that the $1.56978 \times 2 = 3.13956 \approx \pi$. If we were to take even more samples, we could get even closer to the value of π from sampling alone. There was no heavy calculation required, but just sampling and classifying many many times. The code used for this approximation can be found here. It also includes an approximation directly for π using the fact that $x^2 + y^2 = 1$ creates the unit circle. Then, we sample over the unit circle to approximate the area of π.

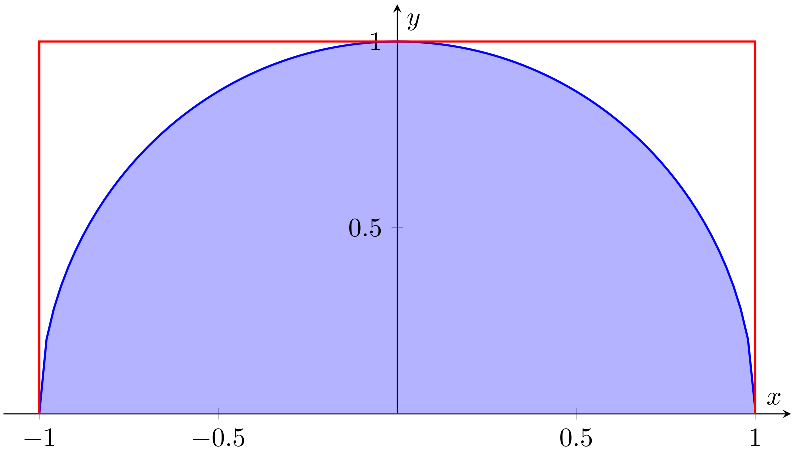

Figure 8: Graph of $f(x) = \sqrt{1-x^2}$ enclosed in larger rectangle

Briefly, also consider a situation where the sampling distribution is not uniform. This means that we would be picking points in our space from a probability distribution $\mu(x)$. Typically, we pick a probability distribution that decreases the variance in our sampling when compared to a uniform distribution. As we had previously shown in Lemma 5.2, we will still be able to approximate the integral by using our probability distribution. An efficient way of doing this is to generate a uniform distribution out of the probability distribution that we have, using the Metropolis-Hastings Algorithm.

Much more complicated applications of this have been studied to find solutions to many problems. An example that the author recently looked at was from the paper Monte Carlo Complexity of Global Solution of Integral Equations by S. Heinrich from 1998. For a much deeper analysis of this problem, check out [5].

6 Acknowledgements

This paper was written under the Euler Circle independent study program. I would like to thank Simon for the advice that was provided for this paper as well as pointing to resources used to study this topic. In addition, I would like to thank my mentor Thomas K. for his suggestions on this paper and the advice provided about this paper during Euler Circle sessions.

References

[1] Christian Borgs, Jennifer T Chayes, Alan Frieze, Jeong Han Kim, Prasad Tetali, Eric Vigoda, et al. Torpid mixing of some monte carlo markov chain algorithms in statistical physics. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 218–229. IEEE, 1999.

[2] Ed Kamya Kiyemba Edris, Mahdi Aiash, Mohammad Ali Khoshkholghi, Ranesh Naha, Abdullahi Chowdhury, and Jonathan Loo. Performance and cryptographic evaluation of security protocols in distributed networks using applied pi calculus and markov chain. *Internet of Things*, 24:100913, 2023.

[3] Victor Ginting, Felipe Pereira, and Arunasalam Rahunanthan. Multi-physics markov chain monte carlo methods for subsurface flows. *Mathematics and Computers in Simulation*, 118:224–238, 2015.

[4] JO Giordano, AS Kalantari, PM Fricke, MC Wiltbank, and VE Cabrera. A daily herd markov-chain model to study the reproductive and economic impact of reproductive programs combining timed artificial insemination and estrus detection. *Journal of dairy science*, 95(9):5442–5460, 2012.

[5] Stefan Heinrich. Monte carlo complexity of global solution of integral equations. *Journal of complexity*, 14(2):151–175, 1998.

[6] Aladár Kollár. Betting models using ai: A review on ann, svm, and markov chain. *Munich Personal Research Association*, 2021.

[7] Steven P. Lalley. Electrical networks and reversible markov chains. *UChicago Publication*, 2023.

[8] Radford M Neal. Probabilistic inference using markov chain monte carlo methods. *Toronto Press*, 1993.

[9] P Ravi Kumar, KL Alex Goh, and KS Ashutosh. Application of markov chain in the pagerank algorithm. *Pertanika Journal of Science & Technology*, 21(2), 2013.

[10] Lauren K Williams. The combinatorics of hopping particles and positivity in markov chains. *arXiv preprint arXiv:2202.00214*, 2022.