ERGODIC MARKOV CHAINS AND END BEHAVIOR

JOSEPH GORMAN

Abstract. This paper's main goal is to investigate Markov chains at their intersection with ergodic theory. We will begin with a probability example and related definitions, prompting our eventual construction of Markov chains. We will then introduce different Markov chain classes, culminating in a discussion of ergodicity. The intitial sections motivate our proof of the ergodic theorem, which we specifically apply to Markov chains as a means of calculating unique stationary distributions. Applications to Markov chains Monte Carlo will be explored in an appendix. Previous knowledge of probability theory is not necessary, and, while an understanding of linear algebra, calculus, and combinatorics is assumed, it is mostly tangential to this paper's central concepts.

Contents

1. A famous example in probability

Example 1.1. Suppose we have a grasshopper, and we want to study his movement. Our grasshopper lives on an infinitely long east-west line, and he spends his life taking randomly directed one meter jumps. With just this information, how can we possibly analyze his movement? One possible first step is to come up with more specific questions to answer about our grasshopper, including the following. (1) What is the probability that our grasshopper returns to the origin after n jumps? (2) What is the probability that our grasshopper makes k consecutive jumps in the same direction? (3) What other interesting patterns and properties arise from our grasshopper's movement? To answer these questions, we can begin by visually representing our grasshopper's movement (see Figures 1 and 2).

Example 1.1 is a famous probability scenario: the random walk. So let us begin with probability. We can introduce some important definitions and useful examples that will be the basis of understanding our grasshopper's movement.

Date: December 20, 2023.

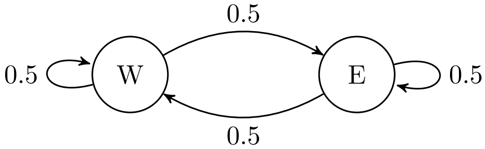

Figure 1. Representation of our grasshopper's directional movement

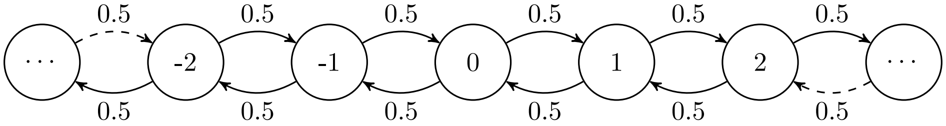

Figure 2. Representation of our grasshopper's positional movement

Definition 1.2. A **sample space** (denoted Ω) is a set of all possible outcomes in a situation or model.

Definition 1.3. A function $P : \Omega \to [0,1] \subset \mathbb{R}$ is called a **probability distribution** if it satisfies the following two conditions:

$$\text{(i) for any } a \in \Omega, P(a) \geq 0, \text{ and (ii) } \sum_{a \in \Omega} P(a) = 1.$$

For the remainder of the paper, P will solely refer to probability distributions. We call the pair (Ω, P) a **probability space**.

The definition of a probability distribution can also be plainly stated: it is a function for which (i) if an event is in the sample space, there is a non-negative chance that it will occur, and (ii) the probability that an event in the sample space occurs is one, meaning an event must come from the sample space. Additionally, mathematicians sometimes study event spaces (i.e. power sets of sample spaces). There are certain situations in which it is important to distinguish between an event space and a sample space—they are generally used in different fields of mathematics—but for this paper's purposes, the difference between the two is not significant. As for probability spaces, they are useful tools for mathematically communicating lexical scenarios, but because they basically act as a container for other information, we will not discuss them further.

Definition 1.4. A **random variable** is a function $X : \Omega \to \mathbb{R}$ that maps randomly occurring events from our sample space to real numbers. The likelihood of each event occurring can be determined by a corresponding probability distribution.

Example 1.5. Consider a game in which you flip a coin. If you flip heads, you receive a point, but if you flip tails, you receive nothing. How can we represent this game? We can construct a random variable and its probability distribution:

$$X = \Big\{ 1 \text{ if heads}, 0 \text{ if tails}. \qquad P(X = 1) = P(X = 0) = 0.5.$$

Definition 1.6. A **discrete-time stochastic process** (denoted $\{X(n)\}_{n \in \mathbb{N}_0}$) is a collection of random variables indexed by $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. This process is discrete because its index is a discrete set. A continuous-time stochastic process would be indexed by $t \in [0, \infty)$, but we will not further analyze these in this paper.

Now that we have introduced some probability-related terminology, we can return to the random walk. There are two key components of our grasshopper's movement: direction and position. We will first consider our grasshopper's directional movement (visualized in Figure 1). He can jump east or west, so our sample space is $\Omega = \{W, E\}$. We can also construct a random variable and its probability distribution for our grasshopper:

$$X = \begin{cases} 1 \text{ if } E, -1 \text{ if } W \end{cases} \quad P(X = 1) = P(X = -1) = 0.5.$$

With the above tools, we can write down different possibilities for the outcomes of our stochastic process. For example, if the grasshopper's jumps were $W, W, E, W, E, E, E, W, E, \cdots$, the process would be $-1, -1, 1, -1, 1, 1, 1, -1, 1, \cdots$.

Notice how, in defining our random variable the way we did, we have a clear link between analyzing the directional and positional movement of our grasshopper. This is because his position increases by one if he jumps to the east and decreases by one if he jumps to the west. So consider our directional stochastic process: if we takes its partial sums, we can find our grasshopper's position at any respective time index. It is important to remember that we cannot simply take the partial sum because the jumps take place in between states. We need to account for the grasshopper's start at the origin, so 0 needs to be the first element of the positional sequence. Let us now compare the different possible representations of our grasshopper's movement (note that $\Omega(n)$ is not mathematically rigorous or completely accurate, but it makes sense in comparing these sequences, so we will use it as notation below):

$$\begin{aligned} \text{Directional movement} &= \{\Omega(n)\}_{n \in \mathbb{N}_0} \to W, W, E, W, E, E, E, W, E, \cdots ; \\ \text{Stochastic process} &= \{X(n)\}_{n \in \mathbb{N}_0} \to -1, -1, 1, -1, 1, 1, 1, -1, 1, \cdots ; \\ \text{State movement} &= \{0\}, \sum_{n \in \mathbb{N}_0} \{X(n)\} \to 0, -1, -2, -1, -2, -1, 0, 1, 0, 1, \cdots . \end{aligned}$$

Now we will more deeply analyze the positional movement of our grasshopper (visualized in Figure 2). He lives along an infinite east-west line, so our sample space can simply be the different states in our figure, which makes it $\Omega = \{\cdots, -2, -1, 0, 1, 2, \cdots\} = \mathbb{Z}$. Deciding our random variable is fairly direct: our sample space is already numerical, so we will have our random variable $X(n)$ be an identity mapping (i.e. $X(n)$ will directly correspond to the grasshopper's position at time n). To determine our probability distribution, we have several things to consider, but the function boils down to the following relationship:

$$P(X(n) = k) = \frac{\text{number of paths to state } k \text{ in } n \text{ steps}}{\text{number of possible paths taken in } n \text{ steps}}.$$

We will first determine an expression to represent the denominator. Our grasshopper has two choices for each jump, east and west, so, after he makes n jumps, there are 2^n possible paths he could have taken. Evaluating the numerator in an explicit, numerical manner is slightly more complicated. Define $\mathcal{E}, \mathcal{W}$ as the number of jumps east or west, respectively. We know that $k = \mathcal{E} - \mathcal{W}$ and $n = \mathcal{E} + \mathcal{W}$, implying that $\mathcal{E} = \frac{n+k}{2}$. Let us also define an n-length set $J \coloneqq \{j_1, \cdots, j_n\}$ where j_i represents the direction of the i-th jump of the grasshopper (note that all elements of J are E or W, and sets cannot have repeating elements, so J is not technically a set, but we are calling it such because we will perform set-like operatives on it).

Given all of this information, we know that, in order to reach state k after n steps, the number of E in our set J must be $\mathcal{E}$. So the number of paths to state k in n steps has to be equal to the number of size-$\mathcal{E}$ subsets of J. This means that our numerator is $\binom{n}{\mathcal{E}} = \binom{n}{\frac{n+k}{2}}$, implying that $\frac{n+k}{2} \in \mathbb{Z}$. Now we can construct the following probability distribution, which, given the grasshopper's starting point as the origin, evaluates the likelihood the grasshopper is at state k after n jumps:

$$P(X(n) = k) = \frac{\binom{n}{\frac{n+k}{2}}}{2^n}. \tag{1.7}$$

The above formula is a useful distribution, but it is not the only condition of our grasshopper's movement. We know that the next state our grasshopper visits must be adjacent to his current state, mirroring the movement and related probabilities of the grasshopper's east-west movement:

$$P(X(n+1) = k-1|X(n) = k) = 0.5 = P(X(n+1) = k+1|X(n) = k). \tag{1.8}$$

We see that (1.7) and (1.8) focus on specific scenarios within our simulation (the first assumes $k = 0$ when $n = 0$, and the second only predicts movement one jump at a time), but we combine we can use these equations to derive a more general formula. Assume that our grasshopper has taken i jumps and is at state j, the likelihood that he reaches state k after n jumps (where $n > i$) can be calculated by the formula below:

$$P(X(n) = k|X(i) = j) = \frac{\binom{n-i}{\frac{(n+k)-(i+j)}{2}}}{2^n} \text{ such that } \frac{n+k}{2}, \frac{i+j}{2} \in \mathbb{Z}.$$

2. Markov chain terminology and distributions

Definition 2.1. Let $i_0, \cdots, i_{n-1}, i, j$ be states in a sample space. The **Markov property** states that

$$\begin{aligned} &P(X(n+1) = j|X(n) = i, X(n-1) = i_{n-1}, \cdots, X(n) = i_0) \\ = \; &P(X(n+1) = j|X(n) = i). \end{aligned}$$

A **Markov chain** is a stochastic process satisfying the Markov property.

Reconsider Example 1.1. The positional and directional movement of our grasshopper can be represented by Markov chains. These fascinating modeling tools are popular because they are relatively simplistic yet powerful stochastic models. This is because of the Markov property, which does not care about the path you took to a current state when determining transition probabilities: Markov chains are memory-free models.

Example 2.2. Board games that use dice are great examples of the Markov property. For example, if you were playing *Chutes and Ladders*, your next state on the board would be entirely dependent on your current state and the roll of a die. It does not matter how you got to your current state, all that matters is that you are there. Conversely, card games do not follow the Markov property. Consider a hand of *Blackjack* in which you hold two Tens, but you see that all of the Aces are in other hands at the table. This means it is impossible for you to draw a card that keeps you under twenty-one, so the only logical choice is to stay, not hit. This choice to stay is not solely based on your current state (the sum of your cards), it is also based on the past states of the players and the specific cards drawn. The path taken (i.e. cards drawn) matters, so card games are Markovian.

In Example 1.1, our grasshopper's future state is entirely dependent on his current state, which can be seen mathematically in (1.8). This formula shows that, no matter how our grasshopper got to state k, his next move is solely dependent on the fact that he is at state k in his current state. Markov chains are generally visually represented weighted digraphs like Figures 1 and 2. Topics in graph theory are not very relevant to this paper, so a formal definition of weighted digraphs is not necessary, and we do not need to dive into their properties. For this paper's purpose, a weighted digraph is a collection of states connected by arrows and probabilities.

Definition 2.3. Consider two states i and j in a sample space. The **transition probability** (denoted P_{ij}) is the probability of directly moving from state i to j. This can be formulaically represented as

$$P_{ij} = P(X(n+1) = j | X(n) = i).$$

A Markov chain is called **time-homogeneous** if its transition probabilities do not change over time. We will exclusively study time-homogeneous Markov chains for the remainder of this paper.

For our purposes, the notion of a probability vector is particularly important. It is quite similar to a probability distribution; a probability vector is any vector that has no negative entries and all entries sum to exactly one. For any state in a Markov chain, there will certainly be a next state visited (even if it is itself). Thus, if we index the states of a Markov chain by $\mathbb{N}$, we can create a probability vector for each state, and its elements will represent the likelihood of it visiting each state on the next step. This is called a transition vector. For example, assume we have a Markov chain with states $\Omega = \{1, \cdots, N\}$. Then for an arbitrary state i, we can create the transition vector $\mathbf{p}_i = \{P_{i1}, \cdots, P_{iN}\}$.

Definition 2.4. Consider a Markov chain with $|\Omega| = n$. Its **transition matrix** (denoted P) is an $n \times n$ matrix such that its ij-th entry is P_{ij}. The transition matrix's contents are n n-length indexed transition vectors.

Definition 2.5. Let $i_0, i_1, \cdots$ be states in a sample space. We write $\phi(i_j) = P(X(0) = i_j)$ as the probability of the Markov chain starting at state i_j. The **initial distribution** of a Markov chain is $\overline{\phi} = (\phi(i_0), \phi(i_1), \cdots)$.

Proposition 2.6. *The product of n transition matrices is also a transition matrix.*

Proof. Let matrices $P, P_1, \cdots, P_n$ be k-dimensional transition matrices and define $\mathbf{1} := (1, \cdots, 1)^T \in \mathbb{R}^k$. Because of the unique nature of $\mathbf{1}$, for every row i the product $P\mathbf{1} = \sum_{j=1}^{k} P_{ij} = 1$, which means $P\mathbf{1} = \mathbf{1}$ because each row of the matrix is a transition vector. Thus, if $(P_1 \circ \cdots \circ P_n)(\mathbf{1}) = \mathbf{1}$, then $(P_1 \circ \cdots \circ P_n)$ is a transition matrix. This can be proven with the associative property: multiply the two rightmost elements over and over until the result is $\mathbf{1}$. □

Theorem 2.7. *Consider a Markov chain with a finite sample space $\Omega = \{1, \cdots, N\}$, a transition matrix P, and an initial distribution $\overline{\phi} = (\phi(1), \cdots, \phi(N))$. Then the probability vector of $X(n)$ is given by*

$$(P(X(n) = 1), \cdots, P(X(n) = N)) = \overline{\phi} P^n.$$

Proof. This can be proved by induction. We know that the probability being at state j when $n = 1$ is based on the probability of us taking any path that contains

$n = 1$. So, we can sum of the probability of taking each of these paths, which we can then simplify by the definition of transition probabilities:

$$P(X(1) = j) = \sum_{i \in \Omega} P(X(0) = i) \cdot P(X(1) = j | X(0) = i) = \sum_{i \in \Omega} \phi(i) \cdot P_{ij}.$$

This final summation essentially does vector-matrix multiplication with the initial distribution and the j-th column of the transition matrix, making this the j-th element in our initial distribution, thus satisfying the $n = 1$ case. We can prove the inductive step using the Markov property.

$$\begin{aligned} P(X(n+1) = j) &= \sum_{i \in \Omega} P(X(n) = i) \cdot P(X(n+1) = j | X(n) = i) \\ &= \sum_{i \in \Omega} P(X(n) = i) \cdot P_{ij}. \end{aligned}$$

Assuming that $P(X(n) = i)$ is given as the i-th element of the row vector $\overline{\phi} P^n$, we observe that the j-th coordinate of the row vector $(\overline{\phi} P^n) P$ coincides with $P(X(n+1) = j)$ obtained above. We deduce that the distribution of $X(n+1)$ is $\overline{\phi} P^{n+1}$. □

Proposition 2.6 is not necessary for completing Theorem 2.7 or any future proof in this paper, but these two proofs side-by-side yield notable information nonetheless. We know that P^n is itself a transition matrix, which makes P^n a useful calculation in so many different simulations and circumstances. In essence, this means P^n_{ij} is the probability of transitioning from state i to j in exactly n steps. Theorem 2.7 is a very applicable theorem because it gives a relatively straightforward way of calculating probability distributions at any time n. This proof will continue to be important as we begin to study other properties of Markov chains, so we will return to it later.

3. Properties of various Markov chain classifications

Many Markov chain classifications are based on the properties of different paths, which are essentially sequences of states that connect one state to another. More formally, there is a path from state i to state j (denoted $i \to j$) if there is a sequence of states $i = i_0 \to i_1 \to \cdots i_n = j$ that connects the two where we have $P_{i_{l-1} i_l} > 0$ for all $l = 1, \cdots, n$ (i.e. there is a positive probability of going between any adjacent states in the sequence). If we have $i \to j$ and $j \to i$, we say that i and j communicate (denoted $i \leftrightarrow j$). Additionally, a communication class is a group of states that all communicate with each other (note that all states communicate with themselves, and every state belongs to exactly one communication class, which can be proved by contradiction).

Definition 3.1. A Markov chain is **irreducible** if it has only one communication class. If this is not true, the Markov chain is **reducible**.

Definition 3.2. The **hitting time** of a state i is $T_i = \inf\{n > 0 | X(n) = i)$. When the Markov chain starts at state i, this is generally called the **return time**.

Definition 3.3. A state is **recurrent** if and only if $P(T_i < +\infty | X(0) = i) = 1$, but if this is not true, the state is called **transient**.

We will return once again to Figure 1. We can draw a path $E \to W$ and $W \to E$ directly, implying that $E \leftrightarrow W$, meaning E and W are in the same communication class. Because these are the only two states in the Markov chain, there is only one communication class, so we can say Figure 1 is irreducible. Assume $X(0) = E$. The only possible way to never return to E is to have the random variables index as $E, W, W, W, \cdots$. This would mean that for every $n \in \mathbb{N}$, $X(n) = W$, and we know that $P(X(n) = W | n \in \mathbb{N}) = 0.5$. So, the probability of this collection of random variables occuring is $\lim_{n\to\infty}(0.5)^n = 0$, which implies that $P(T_i < +\infty | X(0) = i) = 1$, so the state E is recurrent. The same argument follows for the recurrence of state W, so each state is recurrent.

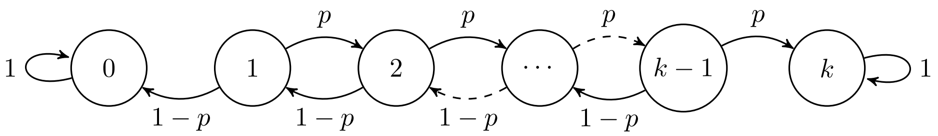

FIGURE 3. Diagram of k-state Gambler's ruin

Example 3.4. The Gambler's ruin simulation depicts two parties, A and B, who bet against each other in a series of hands of some game (oftentimes one party is a casino, betting agency, etc.). Gambler's ruin shows that the odds are always stacked in the house's favor, and the Markov chain's physical nature has some interesting properties. In Gambler's ruin, player A has a probability p, $(0 < p < 1)$ of winning each hand, implying party B has probability $1 - p$. Assume that combined money the two parties are willing to bet is $k \in \mathbb{Z}$; player A has an initial wealth of $i \in \mathbb{Z}$; each player bets one dollar per hand; and the game continues until one of the players is out of money. We can represent this game as a k-state Markov chain with $X(0) = i$. $X(n) = 0$ (resp. $X(n) = k$) implies that player A (resp. player B) is out of money after n hands. Figure 3 depicts the gambler's ruin Markov chain. Consider state 0. There is no path from state 0 to any other state in the chain, so state 0 is in its own communication class. State k is similarly in its own communication class. Because this Markov chain has the paths $1 \to 2 \to \cdots \to k-1$ and $k-1 \to k-2 \to \cdots \to 1$, all of these states exist in a communication class together. Our Markov chain is reducible with the following three communication classes: $\{0\}, \{1, \cdots, k-1\}, \{k\}$. Now let us look at the recurrence and transience of different states. For state 0, $X(0) = 0$ implies that $X(1) = 0$, so this state is recurrent, and state k is similarly recurrent. Consider some arbitrary state $j \in \{1, \cdots, k-1\}$. We are $k - j$ steps from reaching state k, and there is probability p that we will move one step closer to state k during each step. Mathematically, we can show that if $X(0) = j$, then

$$P(X(1) = j + 1 | X(0) = j) = p, \text{ implying } P(X(k - j) = k | X(0) = j) = p^{(k-j)} > 0.$$

This means that, for state j, there is at least a $p^{(k-j)}$ chance that it will never be returned to (note that the actual likelihood is higher, but completely unpacking that probability is not necessary for proving this state's transience), so every state in the communication class $\{1, \cdots, k-1\}$ is transient.

Proposition 3.5. *The states in a communication class are either all recurrent or all transient (i.e. of the same type), and we call this class such.*

Proof. This proof makes use of another theorem about recurrence: a state i is recurrent if and only if

$$\sum_{n=1}^{\infty} P_{ii}^{n} = +\infty. \tag{3.6}$$

While that property will not be proven in this paper, it can be seen in full as Theorem 5.3.1-2 in Durrett (2019). Assume that states i, j are in the same communication class. Then, there exist $l, m \in \mathbb{N}$ such that $P_{ij}^{l}, P_{ji}^{m} > 0$. Further, for any $k \geq 1$, $P_{ij}^{l} \cdot P_{jj}^{k} \cdot P_{ji}^{m}$ is the probability of going from i to j in l steps, looping back to j in k steps, and returning from j to i in m steps. As this is just one specific path of looping back from i to i in $l+k+m$ steps, we know that $P_{ij}^{l} \cdot P_{jj}^{k} \cdot P_{ji}^{m} \leq P_{ii}^{l+k+m}$ (there may be other possible loops back to state i in the same time). Further, the loop back to state i may be possible in any $n \in \mathbb{N}$ steps without looping through state j, making an even further generalization of a return possibility. Thus, we have the following inequality:

$$\sum_{n=1}^{\infty} P_{ii}^{n} \geq \sum_{k=1}^{\infty} P_{ii}^{l+k+m} \geq \sum_{k=1}^{\infty} P_{ij}^{l} \cdot P_{jj}^{k} \cdot P_{ji}^{m} = P_{ij}^{l} \cdot P_{ji}^{m} \cdot \sum_{k=1}^{\infty} P_{jj}^{k}.$$

Assume without loss of generality that j is a recurrent state. Then, (3.6) would imply that $\sum_{k=1}^{\infty} P_{jj}^{k} = +\infty$, which implies that $\sum_{n=1}^{\infty} P_{ii}^{n} = +\infty$, so i must be a recurrent state. Similarly, assume without loss of generality that i is a transient state. Then the above argument works in reverse to show that j must also be a transient state. □

Definition 3.7. A possible **return path** to state i is represented as $D_i = \{n \in \mathbb{N} | P_{ii}^{n} > 0\}$. The **period** of state i is mathematically represented as

$$per(i) = \sup\{n \in N | \frac{x}{n} \in \mathbb{N} \text{ for any } x \in D_i\}.$$

A state i is called **periodic** if $per(i) > 1$ and **aperiodic** if $per(i) = 1$.

Essentially, the period of a state is the largest number to divide the length of every possible return path (i.e. the greatest common factor of all possible return times for a state). This definition is limited, though, because it only refers to individual states. We will expand on this definition soon, but we will first return to Figure 1. It is possible to return to state E after one step, so the period of state E must be one, and state W similarly has period one. We can thus say that this Markov chain is aperiodic because each of its states is aperiodic. Now reconsider Figure 2. In order to return to state 0 after n steps, we need $\mathcal{E} + \mathcal{W} = n$ and $\mathcal{E} + \mathcal{W} = 0$, which implies that $\mathcal{E} = \mathcal{W}$ and n must be even, so the period of state 0 is two. Similarly, every other state in this Markov chain has period two.

Proposition 3.8. *The states in a communication class all have the same period, and if that period is one, the communication class is aperiodic.*

Proof. Assume that states i and j are in the same communication class. Then, there exist $l, k \in \mathbb{N}$ such that $P_{ij}^{k} > 0, P_{ji}^{l} > 0$. We know that $m = l + k \in D_i, D_j$ because there is a path for i through j and vice versa. Assume there exists $D_i = n$, then there must also exist $D_j = l + n + k = n + m$ from an l-step trip from j to i, an n-step loop at i, and a k-step return from i to j. Because of this, we know that $per(j)$ divides n, and $per(j)$ divides $n + m$, so $per(j)$ must divide m. Extending this argument to any size $x \in D_i$ return time, it must follow that $per(j)$ divides x. Because $per(j)$ divides x for any $x \in D_i$, and $per(i)$ is the supremum of all possible numbers that divide any $x \in D_i$, it follows that $per(i) \geq per(j)$. The same logical argument follows to show that $per(j) \geq per(i)$, so $per(i) = per(j)$. □

4. Ergodic Markov chains and stationary distributions

Definition 4.1. We call a Markov chain **ergodic** if and only if it is irreducible, recurrent, and aperiodic.

It may seem difficult to prove that a Markov chain is ergodic, but, because of our work in Section 3, we have some shortcuts. To prove ergodicity, we would first prove need to prove irreducibility, which can be most easily done by finding looping paths that pass through every state in both directions. After this, we only have to prove that one state is recurrent because, based on Proposition 3.5, the rest of the states would then also be recurrent. Similarly, we only need to show that one state is aperiodic because, by Proposition 3.8, the rest of the states would then also be aperiodic.

For the remainder of this paper, we will study some fascinating properties of ergodic Markov chains. This paper is too short to unpack every interesting aspect of ergodic Markov chains (e.g. each has some $t \in \mathbb{N}$ such that, for any $i, j \in \Omega$, $P^t_{ij} > 0$), but there are interesting references in the bibliography. This section's primary goal is to motivate and prove the ergodic theorem. Then we will show how to calculate the unique stationary distributions of ergodic Markov chains. But before we prove the ergodic theorem, there are several definitions and proofs we must introduce, including lemmas related to dominated convergence and hitting times between arbitrary states.

Definition 4.2. A recurrent state i is called **positive recurrent** if and only if the expected return to time state i is finite, meaning $\mathbb{E}[T_i : X(0) = i] < \infty$. If this is not true, we call the state **null recurrent** (note that all finite ergodic Markov chains are positive recurrent).

Lemma 4.3. *Let $A_1, A_2, \cdots, A_m, \cdots, B : \mathbb{N} \to \mathbb{R}$ be sequences. Suppose that for any $l, m \in \mathbb{N}$, $|A_m(l)| \leq B(l)$ where $\sum_{l=1}^{\infty} B(l) < \infty$. Also assume that there exists $A : \mathbb{N} \to \mathbb{R}$ such that $\lim_{m\to\infty} A_m(l) = A(l)$. Then this implies that*

$$\lim_{m\to\infty} \sum_{l=1}^{\infty} A_m(l) = \sum_{l=1}^{\infty} A(l).$$

Proof. We can prove that these two summations are equal by proving that the upper limit of their difference is less than every $\varepsilon > 0$. Because the summation of B is finite, we know that $\lim_{l\to\infty} B(l) = 0$, which is equivalent to saying that, for any $\varepsilon > 0$, there exists some $N \in \mathbb{N}$ such that

$$\sum_{l=N+1}^{\infty} B(l) < \frac{\varepsilon}{4}.$$

Consequently, we can also say that, by our second condition, the infinite sums corresponding to A_m and A must also be finite. Thus, in order to prove the equivalence of these two summations, we want to show that the difference between them heads to zero as $m \to \infty$.

$$\begin{aligned}
&\limsup_{m\to\infty}[|\sum_{l=1}^{\infty} A_m(l) - \sum_{l=1}^{\infty} A(l)|] \\
\leq &\limsup_{m\to\infty}[|\sum_{l=1}^{\infty}(A_m(l) - A(l))|] \\
\leq &\limsup_{m\to\infty}[\sum_{l=1}^{\infty} |A_m(l) - A(l)|] \\
= &\limsup_{m\to\infty}[\sum_{l=1}^{N} |A_m(l) - A(l)| + \sum_{l=N+1}^{\infty} |A_m(l) - A(l)|].
\end{aligned}$$

Because we know that $|A_m(l)|$ and $|A(l)|$ are bounded by $B(l)$, it follows that

$$\sum_{l=N+1}^{\infty} |A_m(l) - A(l)| < 2 \sum_{l=N+1}^{\infty} B(l) < \frac{\varepsilon}{2}.$$

Also, because of our previously set conditions, $\lim_{m\to\infty} A_m(l) = A(l)$, we know for every $i \in \mathbb{N}$ such that $1 \leq i \leq N$:

$$|A_i(l) - A(l)| < \frac{\varepsilon}{2N}, \text{ implying } \sum_{l=1}^{N} |A_m(l) - A(l)| < \frac{\varepsilon}{2}.$$

From this, we can say that

$$\limsup_{m\to\infty} |\sum_{l=1}^{\infty} A_m(l) - \sum_{l=1}^{\infty} A(l)| < \varepsilon.$$

Consequently, these two summations must be equal, concluding our proof. □

Lemma 4.4. *In a communication class containing states i and j, if i is positive recurrent, then j is positive recurrent and $\mathbb{E}[T_j|X(0) = i] < \infty$. This can be expressed mathematically as*

$$\sum_{k=1}^{\infty} P(T_j = k|X(0) = i) = 1.$$

Proof. Consider the value $T_{ij} = \inf\{n \geq 0|X(0) = i, X(n) = j\}$. In other words, T_{ij} is the first visit to state j provided we start at state i. To complete this proof, we will show that T_{ij} must be finite (i.e. we will show that $P(T_{ij} = \infty) = 0$). We can simply manipulate the notation we have constructed to show that

$$\mathbb{E}[T_{ij}|X(0) = i] = \mathbb{E}[T_{ij}] = \sum_{n=1}^{\infty} n \cdot P(T_{ij} = n) + (\infty) \cdot P(T_{ij} = \infty).$$

We want to show that this expected value is finite, which would imply that $P(T_{ij} = \infty) = 0$. Choose the smallest $L \in \mathbb{N}$ such that $P_{ji}^L > 0$ (i.e. L is the lengths of the shortest path from j to i). Now consider a specific type of path:

$$A = \{X(0) = j, X(1) \neq j, X(2), \neq j, \cdots, X(L) = i\}.$$

In essence, these are paths that start at j and go to i in L steps without returning to j. A encompasses all paths from j to i of length L because, if the path returned to j in s steps, then it would take at least $s+L > L$ steps to reach i, which would already fail to meet the conditions of set A. However, this is a special type of path from j to i because there may be paths between the two that are not length L. Thus, we can say that

$$\mathbb{E}[T_{jj}] \geq \sum_{m=1}^{\infty} m \cdot P(T_{ij} = m|A) \cdot P(A).$$

Recall that $\mathbb{E}[T_{ij}] = \sum_{l=0}^{\infty} l \cdot P(T_{ij} = l)$. We can now manipulate our summations:

$$\begin{aligned}
\mathbb{E}[T_{jj}] &\geq \sum_{m=1}^{\infty} m \cdot P(T_{ij} = m|A) \cdot P(A) \\
&= \sum_{m=1}^{\infty} m \cdot P(T_{jj} = m|A) \\
&= \sum_{k=1}^{\infty} (L+k) \cdot P(T_{jj} = L+k|A) \\
&= L \cdot \sum_{m=L+1}^{\infty} m \cdot P(T_{jj} = m|A) + \sum_{k=1}^{\infty} k \cdot P(T_{ij} = k|X(0) = i).
\end{aligned}$$

Thus, we can say

$$\infty > \mathbb{E}[T_{jj}] \leq \mathbb{E}[T_{jj}|A] \cdot P(A) = (L + \mathbb{E}[T_{ij}]) \cdot P(A) \implies \mathbb{E}[T_{ij}] < \infty.$$

This completes our proof, showing that elements in the same communication class will reach each other in finitely many steps. □

Definition 4.5. A sequence of random variables is called **independent and identically distributed** (denoted i.i.d.) if each random variable has the same probability distribution and is independent from every other random variable. Consider a sequence $T, T_1, T_2, \cdots$ of i.i.d. random variables with a distribution function $f(x) = P(T \leq x)$. A sequence $S_0, S_1, S_2, \cdots$ is called a **renewal sequence** if it has $S_0 = 0$ and $S_n = T_1 + \cdots + T_n$ for $n \in \mathbb{N}$.

Theorem 4.6. *Suppose we have a state i in a finite ergodic Markov chain. For any initial distribution, it holds that*

$$\lim_{n\to\infty} P(X(n) = i) = \frac{1}{\mathbb{E}[T_i|X(0) = i]}. \tag{4.7}$$

Proof. We have proven two of the lemmas we will use in this paper, but there is one more which we will not prove—see Proposition 1.2.2 from Asmussen (2003) for a complete proof. This states that, for an aperiodic renewal sequence (u_n) generated by probability vector (f_k), it holds that

$$\lim_{n\to\infty} u_n = \frac{1}{\sum_{k=1}^{\infty} k \cdot f_k}.$$

Now consider a fixed $j \in \Omega$ and define $f_k := P(T_j = k|X(0) = j)$. We can split the event $(X(n) = j)$ into cases by their hitting time at state j. We can thus show by the Markov property that

$$\begin{aligned}u_n &= P(X(n) = j|X(0) = j)\\ &= \sum_{k=1}^{n} P(T_j = k|X(0) = i) \cdot P(X(n-k) = j|X(0) = j)\\ &= \sum_{k=1}^{n} f_k \cdot u_{n-k}.\end{aligned}$$

Additionally, since j is an aperiodic state, it follows that

$$P_{jj}^n = u_n \to \frac{1}{\sum_{k=1}^{\infty} k \cdot f_k}.$$

Consider starting at an arbitrary i and returning to j. By Lemma 4.4, we know

$$\begin{aligned}P_{ij}^n &= P(X(n) = j|X(0) = i)\\ &= \sum_{k=1}^{n} P(T_j = k|X(0) = i) \cdot P(X(n-k) = j|X(0) = j)\\ &= \sum_{k=1}^{\infty} P(T_j = k|X(0) = i) \cdot u_{n-k}.\end{aligned}$$

In order to complete the proof, we need to apply Lemma 4.3 to show dominated convergence's relevance to the ergodic theorem. Consider the function:

$$A_n(k) = \begin{cases} P(T_j = k|X(0) = i) \cdot u_{n-k}, k \leq n \\ 0, k > n. \end{cases}$$

Note that

$$\begin{aligned}\lim_{n\to\infty} A_n(k) &= \lim_{n\to\infty} (P(T_j = k|X(0) = i) \cdot u_{n-k})\\ &= P(T_j = k|X(0) = i) \cdot \lim_{n\to\infty} u_{n-k}\\ &= P(T_j = k|X(0) = i) \cdot \frac{1}{\sum_{k=1}^{\infty} k \cdot f_k} = A(k).\end{aligned}$$

We also know that u_n is a converging sequence, so it must have a supremum:

$$|P(T_j = k|X(0) = i) \cdot u_{n-k}| \leq P(T_j = k|X(0) = i) \cdot \sup u_n.$$

Thus we can define a function

$$B(k) \coloneqq P(T_j = k|X(0) = i) \cdot \sup\{u_n\}$$

to meet another condition of dominated convergence. Note that

$$\sum_{k=1}^{\infty} B(k) = \sup|\{u_n\}| \cdot \sum_{k=1}^{\infty} P(T_j = k|X(0) = i) = \sup|\{u_n\}|.$$

Consequently, we know that the summation of all $B(k)$ must be finite. This means that all of the conditions of dominated convergence are satisfied, so, based on both of our lemmas, we can say:

$$\begin{aligned}\lim_{n\to\infty}(\sum_{k=1}^{\infty} A_n(k)) &= \sum_{k=1}^{\infty} A(k)\\ &= \sum_{k=1}^{\infty} P(T_j = k|X(0) = i) \cdot \frac{1}{\sum_{l=1}^{\infty} l \cdot f_l}\\ &= \frac{1}{\sum_{l=1}^{\infty} l \cdot f_l} \cdot \sum_{k=1}^{\infty} P(T_j = k|X(0) = i) = \frac{1}{\sum_{l=1}^{\infty} l \cdot f_l}.\end{aligned}$$

Because f_l is a probability vector and it is being multiplied by l within its summation, we are generating the reciprocal of a weighted average (i.e. the reciprocal of our desired expected value). This proves the ergodic theorem. □

While the proof of the ergodic theorem itself is not necessarily the most digestible, there are many ways in which we can directly apply it to the end behavior of ergodic Markov chains. Let us consider a probability vector π, which is defined below:

$$\begin{aligned}\pi(j) &\coloneqq \lim_{n\to\infty} P(X(n+1) = j)\\ &= \sum_{l\in\Omega}\{\lim_{n\to\infty} P(X(n) = l)\} \cdot P_{lj}\\ &= \sum_{l\in\Omega} \pi(l)P_{lj}.\end{aligned}$$

Definition 4.8. Consider the above probability vector π and define $\overline{\pi} = (\pi(j))_{j\in\Omega}$, then $\overline{\pi}$ is a **stationary distribution**.

To further unpack what this stationary distribution looks like, consider what is happening with each element. Take the i-th element of $\overline{\pi}$. This element is a summation of products: it multiplies each element of itself by its respective element in the i-th column of the transition matrix (i.e. the stationary distribution is equivalent to vector-matrix multiplication of itself by the transition matrix). Thus, we can say that

$$\overline{\pi} = \overline{\pi}P. \tag{4.9}$$

Separately, notice the essence of the stationary distribution. As we see in (4.9), the stationary distribution is resistant to the transition matrix, hence its name. No matter what happens over time with the transition matrix, the probabilities have reached their equilibrium state at the stationary distribution. Also notice when the stationary distribution is reached. By our definition of $\pi(j)$, the stationary distribution occurs as time heads to infinity, which is the same time that we assess for state i in the (4.7). Consequently, with the ergodic theorem, we are calculating the value of state i in a stationary distribution.

There is one complication in the above logic, however. How many stationary distributions does a Markov chain have? When is a stationary distribution unique? We can show that it is possible for a Markov chain to have more than one stationary distribution with Figure 3. If our initial distribution starts at state 0, it follows that our stationary distribution would end up being $\pi = (1, 0, 0, \cdots, 0)$. But if our initial distribution starts at state k, it follows that our stationary distribution would end up being $\pi = (0, 0, \cdots, 0, 1)$. This means we have at least two stationary distributions for Figure 3. So what is necessary for a unique stationary distribution?

To answer this question, let us return to Theorem 2.7 and extend this equation to the case where $n \to \infty$. Then we can say that

$$\lim_{n\to\infty} (P(X(n) = 1), \cdots, P(X(n) = N)) = \overline{\pi} = \lim_{n\to\infty} \overline{\phi} P^n.$$

From this, we know that, just like any other probability vector for a Markov chain, a stationary distribution, if it exists, for a Markov chain is uniquely determined by its initial distribution and transition matrix.

How does this relate to ergodic Markov chains? Return to Theorem 4.6: to find the probability of visiting a state i as $n \to \infty$, we calculate the return time to i with the assumption that the initial distribution starts at i. First, because an expected return time exists, we know that any state in an ergodic chain has a stationary limit. Second, because the initial distribution is fixed, the stationary distribution of an ergodic Markov chain is entirely dependent on its transition matrix. Consequently, we can say that **the stationary distribution of an ergodic Markov chain is unique**. Moreover, we can calculate the infinite step transition matrix, and, because of the initial distribution in 4.6, we know that our stationary distribution is equivalent to the sequence of diagonal elements in our resulting matrix. In other terms, the stationary distribution of a k-state ergodic Markov chain is

$$\overline{\pi} = \lim_{n\to\infty} (P^n_{11}, P^n_{22}, \cdots, P^n_{kk}).$$

5. Appendix: Introduction to Markov chain Monte Carlo

Ergodic Markov chains are noteworthy for the relative ease of calculating their unique stationary distributions. But there are other fascinating applications of Markov chains (especially ergodic ones) in other fields, specifically data science. These processes can be applied in Markov chain Monte Carlo (MCMC) simulation. Before we dive into MCMC, we need to understand Monte Carlo simulations.

Definition 5.1. A **Monte Carlo** simulation is a technique to model the probability of various events that are otherwise difficult to predict due to randomness.

Example 5.2. One common type of Monte Carlo simulation is Monte Carlo integration, which aims to be able to determine the area of different integrals and spaces. In this example, we are going to use Monte Carlo integration to approximate π. We can first inscribe a circle of radius one inside a square (implying that the square has side length two and area four). We can then randomly shoot points into the square and see the proportion that land within the circle. In multiplying this proportion by four, we can approximate π (sample python simulation shown):

```
import numpy as np
num_sims = _____
num_inside_circ = 0
for i in np.arange(num_sims):
    x = np.random.uniform(-1,1)
    y = np.random.uniform(-1,1)
    if x**2 + y**2 <= 1:
        num_inside_circ += 1
4 * num_inside_circ / num_sims
```

There are notable upsides and downsides Monte Carlo simulations. They are less complicated than many other types of algorithms, but, in order to generate accurate results, they take time. Consider possible simulation sizes for Example 5.2. When I ran one billion trials, π was approximated as `3.141501324`, but the simulation took a great deal of computer power. Conversely, when I ran one million trials, the simulation only needed a few seconds to run, but the approximation was `3.140452`: a much worse estimate of π. We will now turn to MCMC.

Definition 5.3. A **Markov chain Monte Carlo** simulation is a method for producing an ergodic Markov chain whose stationary distribution $\overline{\pi}$ closely resembles some target density $p(x)$.

MCMC simulations are based on accept-reject sampling—the process of randomly sampling within an overarching distribution to more accurately model a smaller one—but they are more dynamic in their abilities. Because of Markov chains, MCMC is rooted in its resampling of a distribution based on its last distribution (i.e. $X(0)$ is the first sample, and based on its results, the sample goes to the next state, which may or may not be equivalent to $X(0)$). But how do we determine the transition probabilities between states in MCMC? We use a property called detailed balance, which connects an equation resembling Bayes's theorem to our representation of stationary distributions in (4.9):

$$\begin{aligned}
&\pi(i) \cdot P_{ij} = \pi(j) \cdot P_{ji} \text{ for all } i, j \in \Omega \\
\implies &\sum_{i \in \Omega} \pi(i) \cdot P_{ij} = \sum_{i \in \Omega} \pi(j) \cdot P_{ji} \\
\implies &\pi(j) = \sum_{i \in \Omega} \pi(i) \cdot P_{ij} \implies \pi P = \pi.
\end{aligned}$$

There are many different types of MCMC simulations, but one of the most famous is the Metropolis-Hasting algorithm. In Metropolis-Hastings, our goal is to sample from an unknown target distribution $p(x)$. What we do know, though, is the formula of a function $f(x) \propto p(x)$. Given this, we design a Markov chain with a burn-in period (the beginning states we will later disregard) that may or may not resemble $p(x)$, but we will eventually reach a state and stationary distribution which is treated as $p(x)$.

There are many steps to conducting Metropolis-Hastings. We first randomly choose some starting $X(0)$, and we draw a random sample j from our proposed distribution Q, and we assess Q_{ij}. We then compute a value

$$A_{ij} = \min\{1, \frac{p(j)Q_{ji}}{p(i)Q_{ij}}\}.$$

This is essentially comparing the ratios of the true probabilities (which we can do because of f) and the proposed probabilities. Now we sample $u \in U[0,1]$ to decide what acceptance threshold we will choose for our proposed A_{ij}. If our $u < A_{ij}$, then we transition from $X(0) = i$ to $X(1) = j$, otherwise, we stay and have $X(1) = i$. Then we draw another random sample k from our proposed distribution and repeat all of the above steps once again. Over time, the proposed sample will begin to change and look more and more like our target distribution. This makes Metropolis-Hastings a powerful tool at the intersection of statistics and data science and an amazing application of Markov chains.

Acknowledgements

I would like to give an enormous thank you to my mentor, Elizaveta Shuvaeva! She was so helpful in determining resources for my research, and our meetings were useful opportunities to gain insight on proofs: I would have never been able to complete this paper without her extensive support. She was able to answer my many questions, and, because of our conversations, I have a newfound interest in probability theory and its applications. I would also like to thank Prof. Peter May for a terrific REU, and I would like to thank Prof. Daniil Rudenko and Prof. Laszlo Babai for their interesting and insightful lectures. I am very glad that I participated in this REU; it has helped me find and study my new favorite areas of mathematics!

References

[1] Asmussen, Soren. "Applied Probability and Queues." Arhaus University, 2003.

[2] Babai, Laszlo. "Discrete Mathematics: Lecture Notes." University of Chicago, 2023.

[3] Durrett, Rick. "Probability: Theory and Examples." Duke University, 2019.

[4] Dynkin, E.B. "Markov Processes: Theorems and Problems." University of Moscow, 1969.

[5] Lee, Yoonsang. "Data-driven Uncertainty Quantifications." MATH 106: Topics in Applied Mathematics. Dartmouth College, Winter 2021.

[6] Tolver, Anders. "An Introduction to Markov Chains." Lecture Notes for Stochastic Processes. University of Copenhagen, 2016.

[7] Xing, Eric P. "Approximate Inference: Markov Chain Monte Carlo." CS 10-708: Probabilistic Graphical Models. Carnegie Mellon University, Spring 2017.