Markov Chains: A Random Walk Through Particles, Cryptography, Websites, and Card Shuffling

Mike McCaffrey
Department of Mathematics
University of Notre Dame

Professor Liviu Nicolaescu
Advisor

May 1, 2017

MARKOV CHAINS: A RANDOM WALK THROUGH PARTICLES, CRYPTOGRAPHY, WEBSITES, AND CARD SHUFFLING

MIKE MCCAFFREY

Contents

Date: May 1, 2017.

Abstract

In this paper, I will discuss the origins of Markov chains, the theory behind them, and their convergent quality seen in the Ergodic theorem. From there, I will outline the Metropolis-Hastings algorithm, one of the most important applications of Markov chains, and give examples of its effectiveness and applicability in various areas.

"Life calls the tune, we dance."
- John Galsworthy

Notation

- $\mathbb{N}$ is the set of nonnegative integers.

1. Introduction

We begin our story in 1856, St. Petersburg, Russia. Andrei Andreyevich Markov was born, fell in love with mathematics, and became prominent in the Academy of Sciences, established in St. Petersburg by Peter the Great (1682-1725). Markov was born at a time when the study of probability was thriving in Europe. During this time Jacob Bernoulli had proved one of the first versions of the Law of Large Numbers. He formally proved that the proportion of heads in repeated tossings of a fair coin converged to the expected value of the process. Coin flips are independent events, meaning that the outcome of a current coin flip does not depend on any previous coin flips. Thus Bernoulli proved that independent events have a convergence property. This inspired a both moral and mathematical argument from a man named Pavel Nekrosov.

Nekrosov was a Russian theologian turned mathematician, and argued that Bernoulli's discovery was proof of free will. He noted that social data, such as crime rates, converge to a probabilistic average by the law of large numbers. Therefore he argued individual acts, such as the commission of a crime, must be independent. In other words individual acts are voluntary and done out of free will. Nekrosov happened to be one of Markov's social enemies, a man that Markov referred to as an abuse of mathematics. Markov claimed that independence was not needed for such a convergence. In order to prove Nekrosov's claims false, Markov laid the foundation of what are now known as Markov chains, probability objects that are dependent on a current state for a future state. These chains, with certain assumptions, are able to converge in a way similar to coin flips.

Many events in the natural world are not independent. The weather today cannot possibly be independent of yesterday's weather; the probability of passing a test is somehow related to how much one has studied the night before. Markov's constructions demonstrate that even though the future of the natural world is dependent on elements of the past, there may still exist some higher natural order or convergence. But before we can see Nekrosov proven wrong, we must delve into the construction of Markov chains, and the properties that underlie the necessary assumptions (History provided by [10]).

This paper assumes an introductory understanding of basic aspects of probability theory. The next section will review various concepts, and further discussion can be found in [2, 7, 8]. A majority of the theory discussed in Markov chains is greatly inspired by [3], in addition to all of the other references listed.

2. A Few Probabilistic Facts.

2.1. **State spaces and random objects.** In this paper we define a *state space* to be a pair $(S, \mathscr{S})$ where S is a set and $\mathscr{S}$ is a σ-algebra of subsets of S. The subsets in $\mathscr{S}$ are called the *measurable* or *observable* subsets of the state space.

A *probability* measure on $(S, \mathscr{S})$ is a measure $\mathbb{P} : \mathscr{S} \to [0, \infty]$ such that $\mathbb{P}(S) = 1$. A *probability space* is a triplet $(\Omega, \mathscr{O}, \mathbb{P})$, where $(\Omega, \mathscr{O})$ is a state space and $\mathbb{P} : \mathscr{O} \to [0, 1]$ is a probability measure. In this case, the subsets $E \in \mathscr{O}$ are called (observable) *events*.

Example 2.1 (Fundamental Examples)**.** (a) A finite or countable set $\mathbb{I}$ is naturally a state space in which any subset is measurable. A *distribution* on $\mathbb{I}$ is a function $\mu : \mathbb{I} \to [0, \infty)$ such that

$$\sum_{i \in \mathbb{I}} \mu(i) = 1.$$

A distribution defines a probability measure $\mathbb{P}_\mu$ on $\mathbb{I}$ by setting

$$\mathbb{P}_\mu(J) = \sum_{j \in J} \mu(j) \ \ \forall J \subset \mathbb{I}.$$

Conversely, any probability measure $\mathbb{P}$ on $\mathbb{I}$ defines a distribution $\mu : \mathbb{I} \to [0, 1]$,

$$\mu(i) = \mathbb{P}(\{i\}), \ \ \forall i \in \mathbb{I}.$$

The *Dirac distribution* concentrated at $i \in \mathbb{I}$ is the distribution $\delta_i : \mathbb{I} \to [0, 1]$ defined by

$$\delta_i(j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i. \end{cases} \tag{2.1}$$

(b) The real axis $\mathbb{R}$ has a natural structure of state space in which the measurable subsets are the Borel subsets of $\mathbb{R}$. □

Fix a probability space $(\Omega, \mathscr{O}, \mathbb{P})$. Suppose that $(S, \mathscr{S})$ is a state space. An *S-valued random object* or *S-valued random variable* is a measurable map $X : \Omega \to S$, i.e., a map such that

$$X^{-1}(A) \in \mathscr{O}, \ \ \forall A \in \mathscr{S}.$$

In the special case when S is the canonical state space $\mathbb{R}$ we will refer to an $\mathbb{R}$-valued random variable X simply as a *random variable*. Thus, a random variable X is a function $X : \Omega \to \mathbb{R}$ such that the set $\{X < r\} \subset \Omega$ is measurable for any $r \in \mathbb{R}$.

The *distribution* of a random variable is the probability measure $\mathbb{P}_X$ on the state space $\mathbb{R}$ uniquely determined by the equalities

$$\mathbb{P}_X\big((-\infty, r) \big) = \mathbb{P}(X < r), \ \ \forall r \in \mathbb{R}.$$

If X is an integrable random variable, then its *expectation* is the real number

$$\mathbb{E}[X] := \int_\Omega X(\omega) \mathbb{P}(d\omega) = \int_{\mathbb{R}} x \mathbb{P}_X(dx).$$

Definition 2.2. (a) A collection of events $(A_i)_{i \in I} \subset \mathscr{A}$ is called *independent* if, for any *finite* subset $J \subset I$, we have

$$\mathbb{P}\Big(\bigcap_{j \in J} A_j \Big) = \prod_{j \in J} \mathbb{P}\big(A_j \big).$$

(b) Let $(\mathbb{S}, \mathscr{A}_\mathbb{S})$ be a state space. A collection of $\mathbb{S}$-valued random variables $(X_i)_{i \in I}$ is called *independent* if, for any collection of measurable subsets $S_i \in \mathscr{A}_\mathbb{S}$, $i \in I$, the collection of events $\big(\{X_i \in S_i\} \big)_{i \in I}$ is independent.

(c) A collection of S-valued random variables $(X_i)_{i\in I}$ is called *identically distributed* if

$$\mathbb{P}(X_i \in A) = \mathbb{P}(X_j \in A), \ \ \forall i,j \in I, \ \ A \in \mathscr{S}.$$

(d) A collection of S-valued random variables $(X_i)_{i\in I}$ is called *independent, identically distributed* or *i.i.d.* if it is both independent and identically distributed. □

Definition 2.3. An event $A \in \mathscr{O}$ happens *almost surely* (*a.s.* for brevity) if

$$\mathbb{P}(A) = 1.$$

A sequence $\{X_n\}_{n\geq 0}$ of random variables converges almost surely to a random variable X if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

□

Theorem 2.4 (Strong Law of Large Numbers)**.** *If $\{X_n\}_{n\geq 1}$ is an i.i.d. sequence of random variables such that*

$$\mathbb{E}[X_1] < \infty$$

then, almost surely,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X_1].$$

□

Example 2.5. If X_1 is a fair coin flip, that is it takes on value 0 with probability $\frac{1}{2}$ and value 1 with probability $\frac{1}{2}$, we have that $\{X_n\}_{n\geq 1}$ defines an i.i.d. sequence of random variables, where $\mathbb{E}[X_1] = 1/2$. Thus, almost surely,

$$\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \frac{1}{2}.$$

□

2.2. **Conditional Probability and Independence.** Given two events $A, B \in \mathscr{O}$, we define the conditional probability of A given B to be the the number

$$\mathbb{P}(A|B) := \begin{cases} \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}, & \mathbb{P}(B) \neq 0, \\ 0, & \mathbb{P}(B) = 0. \end{cases} \tag{2.2}$$

Suppose that $B \in \mathscr{O}$ has positive probability. Denote by $\mathscr{F}_B$ the σ-algebra of subsets of B

$$\mathscr{F}_B := \big\{ S \in \mathscr{F}; \ \ S \subset B \big\}.$$

Then the map

$$\mathbb{P}_B : \mathscr{F}_B \to [0,1], \ \ \mathbb{P}_B(S) = \mathbb{P}(S|B), \ \ \forall S \in \mathscr{F}_B \tag{2.3}$$

is a probability measure.

A *measurable partition* of $(\Omega, \mathscr{O})$ is a countable subfamily $\mathscr{F} \subset \mathscr{O}$ consisting of pairwise disjoint events whose union is Ω. The σ-algebra generated by $\mathscr{F}$ is the subcollection $\mathscr{F}^\sigma \subset \mathscr{O}$ consisting of unions of subsets of $\mathscr{F}$.

Example 2.6. (a) Suppose that $X : (\Omega, \mathscr{O}, \mathbb{P}) \to \mathbb{I}$ is a random object whose range is a countable set $\mathbb{I}$. The partition determined by X is the measurable partition $\mathscr{F}_X$ consisting of the events

$$\big\{ X = i \big\}, \ \ i \in \mathbb{I}.$$

The σ-algebra $\mathscr{F}_X^\sigma$ consists of the events

$$\big\{\, X \in J \,\big\},\ \ J \subset \mathbb{I}.$$

(b) More generally, to a finite number of random objects $X_1, \ldots, X_n : \Omega \to \mathbb{I}$, $\mathbb{I}$ countable, we can associate the measurable partition $\mathscr{F}_{X_1,\ldots,X_n}$ consisting of the events

$$\big\{X_1 = i_1, \ldots, X_n = i_n \,\big\},\ \ i_1, \ldots, i_n \in \mathbb{I}.$$

The σ-algebra $\mathscr{F}^\sigma_{X_1,\ldots,X_n}$ consists of the events

$$\big\{(X_1, \ldots, X_n) \in E \,\big\},\ \ E \subset \mathbb{I}^n.$$

(c) Given two measurable partitions $\mathscr{F}_1, \mathscr{F}_2$ of Ω we denote by $\mathscr{F}_1 \cap \mathscr{F}_2$ the measurable partition of Ω consisting of the events $E_1 \cap E_2$, $E_1 \in \mathscr{F}_1$ and $E_2 \in \mathscr{F}_2$. Observe that if $X_1, X_2 : \Omega \to \mathbb{I}$ are two random objects that

$$\mathscr{F}_{X_1,X_2} = \mathscr{F}_{X_1} \cap \mathscr{F}_{X_2}. \qquad \square$$

Suppose that $\mathscr{F} = (F_n)_{n\in\mathbb{N}} \subset \mathscr{O}$ is a measurable partition of Ω. For any event $E \in \mathscr{O}$ we define *the conditional probability of E given $\mathscr{F}$* to be the *function*

$$\mathbb{P}(E|\mathscr{F}) : \Omega \to \mathbb{R},$$

whose restriction to $F_n \in \mathscr{F}$ is equal to the constant $\mathbb{P}(E|F_n)$.

Definition 2.7. Suppose that we are given three measurable partitions of $\mathscr{O}$, $\mathscr{F}, \mathscr{F}_0, \mathscr{F}_1$. We say that $\mathscr{F}_1$ *is independent of $\mathscr{F}_0$ given $\mathscr{F}$*, and we write this $\mathscr{F}_1 \perp\!\!\!\perp_{\mathscr{F}} \mathscr{F}_0$ if

$$\mathbb{P}(A_0 \cap A_1|\mathscr{F}) = \mathbb{P}(A_0|\mathscr{F})\mathbb{P}(A_1|\mathscr{F}),\ \ \forall A_0 \in \mathscr{F}_0,\ \ A_1 \in \mathscr{F}_1. \tag{2.4}$$

It is not hard to verify that the above condition is equivalent to

$$\mathbb{P}(E_0 \cap E_1|\mathscr{F}) = \mathbb{P}(E_0|\mathscr{F})\mathbb{P}(E_1|\mathscr{F}),\ \ \forall E_0 \in \mathscr{F}_0^\sigma,\ \ E_1 \in \mathscr{F}_1^\sigma. \tag{2.5}$$

In the special case when $\mathscr{F}$ is the partition defined by a random quantity $X : \Omega \to \mathbb{I}$, $\mathbb{I}$ countable, then we say that $\mathscr{F}_0$ is independent of $\mathscr{F}_1$ given X. If additionally, $\mathscr{F}_1$ is defined by a random quantity $(X_1, \ldots, X_n) : \Omega \to \mathbb{I}^n$, then we say that $X_1, \ldots, X_n$ are independent of $\mathscr{F}_0$ given X and we write this $(X_1, \ldots, X_) \perp\!\!\!\perp_X \mathscr{F}_0$. $\square$

In the special case when $\mathscr{F}$ is the partition defined by a random quantity $X : \Omega \to \mathbb{I}$, $\mathbb{I}$ countable, then we say that $\mathscr{F}_0$ is independent of $\mathscr{F}_1$ given X. If additionally, $\mathscr{F}_1$ is defined by a random quantity $(X_1, \ldots, X_n) : \Omega \to \mathbb{I}^n$, then we say that $X_1, \ldots, X_n$ are independent of $\mathscr{F}_0$ given X.

We have the following elementary but very useful alternate characterization of the conditional independence of measurable partitions. More precisely given measurable partitions $\mathscr{F}_0, \mathscr{F}, \mathscr{F}_1$, then $\mathscr{F}_1 \perp\!\!\!\perp_{\mathscr{F}} \mathscr{F}_0$ if and only if these partitions satisfy the *abstract Markov property*

$$\mathbb{P}(E_1|\,\mathscr{F} \cap \mathscr{F}_0) = \mathbb{P}(E_1|\mathscr{F}),\ \ \forall E_1 \in \mathscr{F}_1. \tag{2.6}$$

3. Markov Chain Theory

3.1. **Basic facts and examples.**

Definition 3.1. Fix a finite or countable set $\mathbb{I}$. A *discrete stochastic process* consists of a probability space $(\Omega, \mathscr{O}, \mathbb{P})$ and a sequence of $\mathbb{I}$-valued random variables

$$X_n : \Omega \to \mathbb{I},\ \ n = 0, 1, 2, \ldots,$$

The set $\mathbb{I}$ is called the *state space* of the process.

We denote by $\mathscr{F}_n$ the measurable partition of Ω associated to $X_0, X_1, \ldots, X_n$ (see Example 2.6), i.e.,

$$\mathscr{F}_n := \mathscr{F}_{X_0,\ldots,X_n}.$$

We call X_0 the *initial state* of the process. We define the *initial distribution* to be the probability distribution of X_0, i.e., the function $\mu : \mathbb{I} \to [0,1]$ given by

$$\mu(i) := \mathbb{P}(X_0 = i), \quad \forall i \in \mathbb{I}.$$

In the sequel we will think of probability distributions such as μ as a *row* vectors. □

A Markov chain is a special type of discrete stochastic process, one that has certain properties that describe how our process transitions from one state to another.

Definition 3.2. Given a probability space $(\Omega, \mathscr{O}, \mathbb{P})$, a countable set $\mathbb{I}$ and a probability distribution μ on $\mathbb{I}$ we say that a discrete stochastic process $\left\{ X_n : (\Omega, \mathscr{O}, \mathbb{P}) \to \mathbb{I} \right\}_{n\geq 0}$ is a *Markov chain* with initial distribution μ if the following hold:

(i) The distribution of X_0 is μ.
(ii) For all $n \geq 1$, the random variable X_{n+1} is independent of $\mathscr{F}_n$ given X_n,

$$X_{n=1} \perp\!\!\!\perp_{X_n} \mathscr{F}_n.$$

□

Using the characterization (2.4) of conditional independence we see that the Markov property is equivalent with the requirement that for any $n \in \mathbb{N}$ and any $i_0, i_1, \ldots, i_{n-1}, i, j \in \mathbb{I}$, we have

$$\mathbb{P}(X_{n+1} = j, X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0 | X_n = i)$$
$$= \mathbb{P}(X_{n+1} = j | X_n = i)\mathbb{P}(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0 | X_n = i).$$

Assuming that $\mathbb{P}(X_n = i) \neq 0$, we deduce

$$\frac{\mathbb{P}(X_{n+1} = j, X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)}{\mathbb{P}(X_n = i)}$$
$$= \mathbb{P}(X_{n+1} = j | X_n = i)\frac{\mathbb{P}(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)}{\mathbb{P}(X_n = i)}.$$

This implies

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j | X_n = i), \tag{3.1}$$

or, equivalently,

$$\begin{aligned} &\mathbb{P}(X_{n+1} = j, X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) \\ &= \mathbb{P}(X_{n+1} = j | X_n = i)\mathbb{P}(X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0). \end{aligned} \tag{3.2}$$

The above equality implies inductively that for any $m, n \in \mathbb{N}$ and any $i_0, i_1, \ldots, i_{m+n} \in \mathbb{I}$ we have

$$\begin{aligned} &\mathbb{P}\big(X_{m+n} = i_{m+n}, \ldots, X_{m+1} = i_{m+1}, X_m = i_m, X_{m-1} = i_{m-1}, \ldots, X_0 = i_0\big) \\ &= \mathbb{P}(X_{m+n} = i_{m+n}, \ldots, X_{m+1} = i_{m+1} | X_m = i_m)\mathbb{P}(X_m = i_m, X_{m-1} = i_{m-1}, \ldots, X_0 = i_0). \end{aligned} \tag{3.3}$$

In other words, $\forall m, n \in \mathbb{N}$ we have

$$\mathscr{F}_{X_{m+1},\ldots,X_{m+n}} \perp\!\!\!\perp_{X_m} \mathscr{F}_{X_0,\ldots,X_m}. \tag{3.4}$$

Remark 3.3. For any $n \geq 0$ the σ-algebra $\mathscr{F}_n^\sigma$ consists precisely of the events that we can detect only from the evolution of the process up to time n, the present. The conditional independence (3.4) above is called the *Markov property* and it expresses the fact that *the future is independent of the past given the present.*

Using the abstract Markov property (2.6) we see that (3.4) is equivalent with the condition

$$\mathbb{P}(E|\mathscr{F}_{X_0,X_1,\dots,X_m}) = \mathbb{P}(E|\mathscr{F}_{X_m}), \quad \forall m,n \in \mathbb{N}, \quad E \in \mathscr{F}_{X_{m+1},\dots,X_{m+n}}. \tag{3.5}$$

□

Let us point out a useful consequence of the Markov property. We have

$$\begin{aligned}
&\mathbb{P}(X_n = i_n, \dots, X_1 = i_1, X_0 = i_0)\\
&= \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}, \dots, X_0 = i_0)\mathbb{P}(X_{n-1} = i_{n-1}, \dots, X_0 = i_0)\\
&\overset{(3.1)}{=} \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1})\mathbb{P}(X_{n-1} = i_{n-1}, \dots, X_0 = i_0).
\end{aligned}$$

Arguing inductively we deduce

$$\mathbb{P}(X_n = i_n, \dots, X_1 = i_1, X_0 = i_0) = \prod_{k=1}^{n} \mathbb{P}(X_k = i_k | X_{k-1} = i_{k-1}). \tag{3.6}$$

Using the last inequality in (3.3) we deduce that for any $m, n \in \mathbb{N}$, and any $i_m, i_{m+1}, \dots, i_{m+n} \in \mathbb{I}$ we have

$$\mathbb{P}(X_{m+n} = i_{m+n}, \dots, X_{m+1} = i_{m+1} | X_m = i_m) = \prod_{k=1}^{n} \mathbb{P}(X_{m+k} = i_k | X_{m+k-1} = i_{m+k-1}). \tag{3.7}$$

Definition 3.4. A *homogeneous Markov chain* (or HMC for brevity) is a Markov chain

$$\big\{ X_n : (\Omega, \mathscr{O}, \mathbb{P}) \to \mathbb{I} \big\}_{n \geq 0}$$

such that

$$\mathbb{P}(X_n = j | X_{n-1} = i) = \mathbb{P}(X_{n+1} = j | X_n = i), \quad \forall n \in \mathbb{N}. \tag{3.8}$$

To a homogeneous Markov chain we associate the $\mathbb{I} \times \mathbb{I}$ matrix

$$\mathbf{P} = \{p_{ij}\}_{i,j \in \mathbb{I}},$$

where

$$p_{ij} := \mathbb{P}(X_1 = j | X_0 = i) = \cdots = \mathbb{P}(X_n = j | X_{n-1} = i) = \cdots .$$

We call $\mathbf{P}$ the *transition matrix* of the HMC. Its entries are called *transition probabilities.* □

Remark 3.5. The homogeneity signifies that the probability of moving between states does not depend on what time we happen to be moving through them, and this allows us to define our transition matrix $\mathbf{P}$ to be the transition matrix of states at *any time* of the Markov chain. Thus the value p_{ij} only depends on which states the chain is moving between, which shows why we represent the probability with such a notation. The transition matrix $\mathbf{P}$ inherits nice properties.

The entries in $\mathbf{P}$ are probabilities, so $p_{ij} \in [0,1]$ for all $i, j \in \mathbb{I}$. If at a moment n the system is in a state i, then at the next moment the system will be, with probability one, in some state $k \in \mathbb{I}$. This means that the sum of all entries in each row of the transition matrix is equal to 1, i.e.,

$$\sum_{k \in E} p_{ik} = 1, \quad \forall i \in \mathbb{I}.$$

A matrix of this type is called a *stochastic matrix.* □

Example 3.6. A Health Insurance company can assign probabilities to transitions between states of health, and treat these as values in a Markov Chain. Let H mean that a person is healthy, S sick, and D dead, and the unit of time be in months. Thus, in this case

$$\mathbb{I} = \{H, S, D\},$$

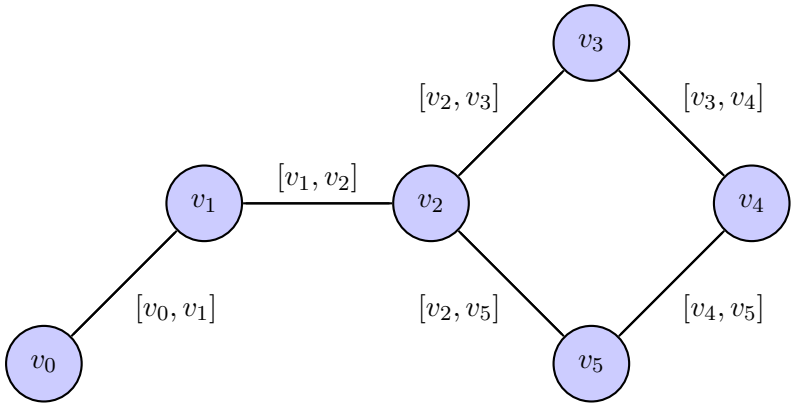
p_{HH} is the probability that in one month a healthy person stays healthy, p_{HS} that a healthy person becomes sick, p_{HD} that a healthy person dies, etc. Obviously $p_{DH} = p_{DS} = 0$ and $p_{DD} = 1$. Further, we require that $p_{HH} + p_{HS} + p_{HD} = 1 = p_{SH} + p_{SS} + p_{SD}$. Therefore we have defined a Markov chain, and this gives us the transition matrix

$$\begin{pmatrix} p_{HH} & p_{HS} & p_{HD} \\ p_{SH} & p_{SS} & p_{SD} \\ 0 & 0 & 1 \end{pmatrix}.$$

□

Definition 3.7. A *graph* is a pair (V, E) consisting of an at most countable set of *vertices* V, and a set of *edges* E, i.e., unordered pairs $[v_1, v_2]$ of distinct vertices v_1, v_2. The vertices v_1 and v_2 are called the *endpoints* of the edge. □

The following is an example of a graph:



Definition 3.8. Let $G = (V, E)$ be a graph.

(i) We say a vertex $y \in V$ is a *neighbor* to $x \in V$ if $[x, y] \in E$. We write this $x \sim y$. We define the *degree* of a vertex v to be the number of neighbors of v.
(ii) The graph is called *locally finite* if each vertex has finite degree, i.e., each vertex has only finitely many neighbors.
(iii) A *weighted graph* is a triplet (V, E, w) where (V, E) is a graph and w is a function $w : E \to (0, \infty)$ that associates a positive number (weight) to each edge of the graph.

□

Example 3.9. Given a locally finite weighted graph (V, E, w) we define

$$Z_w : V \to (0, \infty), \quad P : V \times V \to [0, 1],$$

by setting

$$Z_w(x) = \sum_{x \sim y} w([x, y]),$$

$$p_{xy} = \begin{cases} 0, & x \not\sim y, \\ \frac{w([x,y])}{Z_w(x)}, & x \sim y. \end{cases}$$

The *random walk on the graph* (V, E) weighted by w, is the Markov chain with state space V and transition matrix $P = \{p_{xy}\}_{x,y\in V}$. A very important case of this construction is when the weight w is constant. This walk is called the *canonical random walk* on the graph, and the transition matrix is

$$p_{xy} = \begin{cases} \frac{1}{\deg(x)}, & \text{if } x \sim y, \\ 0, & \text{otherwise.} \end{cases}$$

For the picture of the graph above, if we set each edge with even weights and define a Markov chain on the graph, we get the transition matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}. \qquad \square$$

We have the following important existence result, [7, Thm. 6.1.1].

Theorem 3.10 (Existence of HMCs)**.** *Suppose we are given a countable set $\mathbb{I}$, and a stochastic $\mathbb{I}\times\mathbb{I}$-matrix $\mathbf{P}$. We set $\Omega := \mathbb{I}^{\mathbb{N}}$ so the elements of Ω are sequences*

$$\omega = (i_0, i_1, i_2, \dots), \quad i_k \in \mathbb{I}.$$

We denote by $\mathscr{O}$ the σ-algebra of all the subset of Ω. For each $k = 0, 1, 2, \dots$ we define the measurable map

$$X_k : \Omega \to \mathbb{I}, \quad X_k(i_0, i_1, \dots) = i_k.$$

Then, for any probability distribution μ on $\mathbb{I}$, there exists a probability measure $\mathbb{P}_\mu$ on $\mathscr{O}$ such that the discrete stochastic process $(X_n)_{n\geq 0}$ is a HMC with initial distribution μ and transition matrix $\mathbf{P}$. For $i \in \mathbb{I}$ we denote by $\mathbb{P}_i$ the probability measure $\mathbb{P}_{\delta_i}$, where δ_i is the Dirac measure on $\mathbb{I}$ concentrated at i defined in (2.1). $\square$

Definition 3.11. Let $(X_n)_{n\geq 0}$ be a HMC with state space $\mathbb{I}$. Its *distribution* is the sequence $(\mathbb{P}_n)_{n\geq 0}$ of probability distributions on $\mathbb{I}^{n+1}$ defined by

$$\mathbb{P}_n(\{i_0, \dots, i_n\}) := \mathbb{P}(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n), \quad i_0, i_1, \dots, i_n \in \mathbb{I}.$$

$\square$

Remark 3.12. From now on we say that $\{X_n\}_{n\geq 0}$ is HMC $(\mathbf{P}, \mu)_{\mathbb{I}}$ if $\{X_n\}_{n\geq 0}$ is a HMC with transition matrix $\mathbf{P}$ and initial distribution μ on a state space $\mathbb{I}$, which are objects that exist in this probability space we have just constructed. $\square$

Before we look at more examples of Markov chains, let us look at some immediate consequences of the definition of a HMC.

Theorem 3.13. *The distribution of a HMC is determined by its initial distribution μ and its transition matrix $\mathbf{P}$.*

Proof. Using the Markov condition we deduce

$$\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mu(i_0)p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.$$

□

Definition 3.14. The product $\mathbf{P}^m$ is called the *m-step transition matrix*. Its entries are characterized by the equalities

$$p_{ij}(m) := \mathbb{P}(X_{n+m} = j | X_n = i), \quad i, j \in \mathbb{I}, \quad \forall n \geq 0. \tag{3.9}$$

For Markov chains, we assume that the future is only dependent upon the present. However, given this iterative relationship with current steps and next steps, when we know the set of probabilities of changing between certain states ($\mathbf{P}$), we can predict things far into the future just by knowing our initial state and powers of this matrix $\mathbf{P}$. Our next theorem gives us a clever way to construct more Markov chains.

Theorem 3.15. *Let $(\mathbb{S}, \mathscr{A}_S)$ be a state space and $\mathbb{I}$ a finite or countable set. Let $\{Z_n\}_{n\geq 1}$ be an i.i.d. sequence of $\mathbb{S}$-valued random variables. Let $f : \mathbb{I} \times \mathbb{S} \to \mathbb{I}$ be a measurable function, i.e., for any $i, j \in \mathbb{I}$ the set*

$$\big\{ s \in \mathbb{S}; \quad f(i, s) = j \big\},$$

is a measurable subset of the state space $(\mathbb{S}, \mathscr{A}_S)$. Fix an $\mathbb{I}$-valued random variable X_0, independent of $\{Z_n\}_{n\geq 1}$. Then the sequence of $\mathbb{I}$-valued random variables defined recurrently by

$$X_{n+1} = f(X_n, Z_{n+1}), \quad n \geq 0,$$

defines a HMC with state space $\mathbb{I}$.

Proof. Applying the equation $X_{n+1} = f(X_n, Z_{n+1})$ multiple times shows that there is a measurable function

$$g_n : \mathbb{I} \times \underbrace{\mathbb{S} \times \cdots \times \mathbb{S}}_{n} \to \mathbb{I},$$

such that $X_n = g_n(X_0, Z_1, \ldots, Z_n)$. Thus, the event $\{X_0 = i_0, \ldots, X_{n-1} = i_{n-1}, X_n = i\}$ can be expressed in terms of $X_0, Z_1, \ldots, Z_n$, so it is independent of Z_{n+1}. Therefore

$$\mathbb{P}(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0)$$
$$= \mathbb{P}(f(i, Z_{n+1}) = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(f(i, Z_{n+1}) = j).$$

This proves the Markov Property because

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(f(i, Z_{n+1}) = j).$$

Moreover, the probability $\mathbb{P}(f(i, Z_{n+1}) = j)$ is independent of n. Therefore we have a homogeneous Markov chain with transition matrix entries $p_{ij} = P(f(i, Z_1) = j)$. □

Example 3.16. Let X_0 be an integer valued random variable, and $\{Z_n\}_{n\geq 1}$ a sequence of i.i.d. random variables taking values ± 1, where

$$\mathbb{P}(Z_n = +1) = p \in (0, 1). \tag{3.10}$$

Define a HMC with the relation

$$X_{n+1} = X_n + Z_{n+1}. \tag{3.11}$$

This yields a transition matrix

$$\begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & 1-p & 0 & p & 0 & 0 & \cdots \\ \cdots & 0 & 1-p & 0 & p & 0 & \cdots \\ \cdots & 0 & 0 & 1-p & 0 & p & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We call this Markov chain a *random walk* with probability p. □

3.2. **Classifying the states of HMCs.** In this section we will start to classify different types of HMCs. We can then analyze which types of Markov chains behave better than others. Fix a Markov chain $\{X_n\}_{n\geq 0}$ that is HMC $(\mathbf{P}, \mu)_{\mathbb{I}}$.

Definition 3.17. (a) We say that the state j is *accessible* from the state i if there exists an integer $m \geq 0$ such that

$$\mathbb{P}(X_m = j | X_0 = i) > 0.$$

Using the notation (3.9) we can rewrite this condition

$$p_{ij}(m) > 0.$$

Note that any state i is always accessible to itself since we allow $m = 0$.

(b) We say that states i and j *communicate* if i is accessible from j and j is accessible from i. We denote this by $i \leftrightarrow j$.

(c) We say that a state i is *closed* if $p_{ii} = 1$. A set C of states is *closed* if for any $i \in C$ we have that $\sum_{j\in C} p_{ij} = 1$. □

Remark 3.18. If we define a relation $i \sim j$ if and only if $i \leftrightarrow j$, it becomes clear that communication is an equivalence relation that partitions our state space $\mathbb{I}$ into equivalence classes. We call these classes *communication classes*. □

Example 3.19. In Example 3.6 on health insurance, we see that D is a closed state since $p_{DD} = 1$. Therefore H and S are not accessible from D, and D does not communicate with states H or S. □

Definition 3.20. A HMC, along with its transition matrix, is said to be *irreducible* if there exists only one communication class. □

Example 3.21. The random walk in Example 3.16 is irreducible. For any integers $a_0, a_n \in \mathbb{Z}$, we can find a path of integers $a_0, a_1, \ldots, a_n$ such that $p_{a_0a_1}p_{a_1a_2}\cdots p_{a_{n-1}a_n} > 0$, that is, we have a number $n \geq 0$ such that $p_{a_0a_n}(n) > 0$. Hence a_0, a_n communicate. Since we picked these integers arbitrarily, all states in $\mathbb{Z}$ communicate, hence there is one communication class and the random walk is irreducible. □

Example 3.22. Let us consider the Example 3.6 on health insurance. Since D is a closed state, the chain is not irreducible. □

Definition 3.23. We define the period d_i of a state $i \in \mathbb{I}$ to be given by

$$d_i := \gcd\{n \geq 1 : p_{ii}(n) > 0\}.$$

We define $d_i = \infty$ if there is no $n \geq 1$ such that $p_{ii}(n) > 0$. If $d_i = 1$, we say that state i is *aperiodic*. □

Theorem 3.24. *If the states i and j communicate, they have the same period.*

Proof. If $i = j$, then clearly $d_i = d_j$. Assume $i \neq j$. Since i and j communicate, there exist n, m such that $p_{ij}(n), p_{ji}(m) > 0$. We have that for any $k, \nu \geq 1$,

$$p_{ii}(m + \nu k + n) \geq p_{ij}(m)(p_{jj}(k))^{\nu} p_{ji}(n)$$

Thus for any $k \geq 1$ such that $p_{jj}(k) > 0$ we have that $p_{ii}(m + \nu k + n) > 0$ for any $n \geq 1$. This implies that $d_i | (m + \nu k + n)$ for any $n \geq 1$. Since $d_i | (m + n)$, we have that $d_i | \nu k$ for any $n \geq 1$, and that $d_i | k$. And since d_i divides any k such that $p_{jj}(k) > 0$, we have that $d_i | d_j$. Reversing roles of i and j, we also get that $d_j | d_i$, thus $d_i = d_j$. □

Remark 3.25. If a HMC, along with its transition matrix, has one communication class, i.e. is irreducible, then we define the *period* of the HMC to be the period of all the states in the state space (which is the same by the last theorem). Therefore if an irreducible HMC has period one, the HMC is referred to as *aperiodic*. □

Example 3.26. We know that the random walk in Example 3.16 is irreducible, therefore the period of one state is the period of the entire chain. Consider the state $0 \in \mathbb{Z}$. To start at 0 and return to 0 takes a minimum of two steps. In fact, we can only return to 0 in an even amount of steps. Therefore the set $\{n \geq 1 : p_{00}(n) > 0\} = \{2, 4, 6, 8, \ldots\}$ Hence $d_0 = \gcd\{2, 4, 6, 8, \ldots\} = 2$, and thus the random walk has period 2. □

Theorem 3.27. *Consider an irreducible HMC with state space $\mathbb{I}$, period d and transition matrix $\mathbf{P}$. Then for any states $i, j \in \mathbb{I}$ there exist integers $m, n_0 \geq 0$ such that*

$$p_{ij}(m + nd) > 0, \quad \forall n \geq n_0.$$

Lemma 3.28. *Let d be the gcd of $A = \{a_n : n \geq 1\}$, a set of positive integers closed under addition. Then A contains all but a finite number of positive multiples of d.*

Proof of Theorem 3.27. Consider the set

$$A := \{\, k \geq 1 : p_{jj}(k) > 0 \,\}.$$

This set is closed under addition. Indeed, if $k_1 k_2 \in A$, then

$$p_{jj}(k_1 + k_2) \geq p_{jj}(k_1) p_{jj}(k_2) > 0.$$

Since j is accessible from i, there exists m such that $p_{ij}(m) > 0$. Since A is closed under addition and $d = \gcd A$ we deduce from Lemma 3.28 that

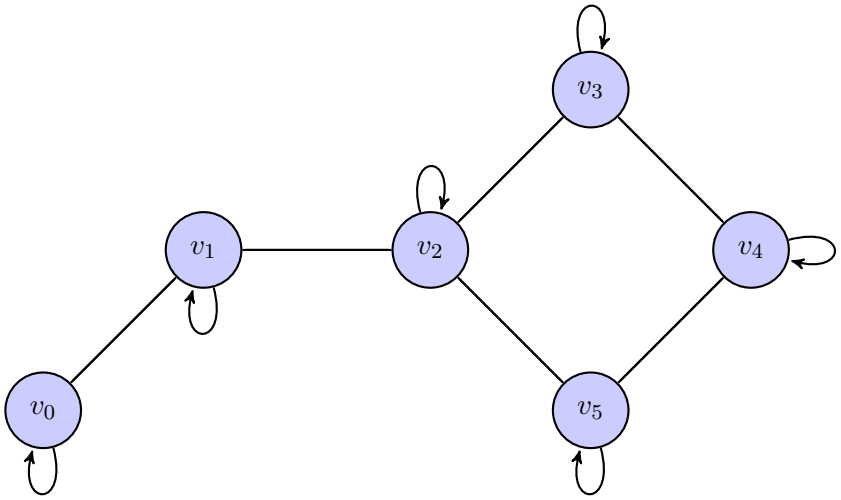
$$\exists n_0 > 0 : \quad \forall n \geq n_0, \quad nd \in A \text{ i.e., } p_{jj}(nd) > 0, \quad \forall n \geq n_0.$$

Hence, for any $n \geq n_0$ we have

$$p_{ij}(m + nd) \geq p_{ij}(m) p_{jj}(n) > 0.$$

□

Example 3.29. Reconsider the random walk on the graph (V, E) in Example 3.9, where we change the graph to be the following:



The canonical random walk on the graph is irreducible and aperiodic. For any v_i, we can get back to v_i in one step with the edge connecting it to itself. Therefore v_i has period one. Since we can get to any other vertex v_j from v_i through a series of walks along edges connecting v_i to v_j, the chain is irreducible. Therefore the chain is irreducible and aperiodic. □

3.3. **Stationary Distributions.** We have established two special features of Markov chains: irreducibility and aperiodicity. These will be important later for our convergence properties. Now we establish the concept of a stationary distribution. Again, fix a Markov chain $\{X_n\}_{n\geq 0}$ that is HMC $(\mathbf{P}, \mu)_{\mathbb{I}}$.

Definition 3.30. A probability distribution π satisfying

$$\pi = \pi\mathbf{P} \tag{3.12}$$

is called a *stationary distribution* of a HMC, along with its transition matrix (here π is a row vector). □

Remark 3.31. This definition is equivalent to saying that $\forall i \in \mathbb{I}$

$$\pi(i) = \sum_{j\in E} \pi(j)p_{ji}. \tag{3.13}$$

Note also that when we iterate equation (3.12), we get for any $n \geq 0$,

$$\pi = \pi\mathbf{P}^n. \tag{3.14}$$

If a chain is started with a stationary distribution, it stays stationary, i.e.

$$\mathbb{P}(X_n = i_0, X_{n+1} = i_1, \ldots, X_{n+k} = i_k) = \pi(i_0)p_{i_0 i_1} \cdots p_{i_{k-1} i_k}$$

does not depend on n. □

A stationary distribution is an initial distribution that remains the distribution after any amount of steps in the Markov chain. While this seems like a very useful distribution to have, one may ask how we find this distribution, and what guarantees its existence.

Definition 3.32. Let $\{X_n\}_{n\geq 0}$ be a HMC with transition matrix $\mathbf{P}$ and stationary distribution π such that $\pi(i) > 0$ for all $i \in \mathbb{I}$. We define the *time reversal matrix* $\mathbf{Q}$, indexed by the same

state space $\mathbb{I}$ as $\mathbf{P}$, by the equation

$$\pi(i)q_{ij} = \pi(j)p_{ji}$$

□

Remark 3.33. $\mathbf{Q}$ is stochastic because $q_{ij} \in [0,1]$ for all $i, j \in \mathbb{I}$, and

$$\sum_{j\in E} q_{ij} = \sum_{j\in E} \frac{\pi(j)}{\pi(i)} p_{ji} = \frac{1}{\pi(i)} \sum_{j\in E} \pi(j)p_{ji} = \frac{\pi(i)}{\pi(i)} = 1.$$

Suppose the initial distribution of the HMC is the stationary distribution ($\mu = \pi$). Then $\mathbb{P}(X_n = i) = \pi(i)$. Thus we have that

$$\mathbb{P}(X_n = j | X_{n+1} = i) = \frac{\mathbb{P}(X_{n+1} = i | X_n = j)\mathbb{P}(X_n = j)}{\mathbb{P}(X_{n+1} = i)} = \frac{p_{ji}\pi(j)}{\pi(i)} = q_{ij}.$$

Thus the i, j entry of $\mathbf{Q}$ is the probability of moving backwards in time from state i to state j. □

Theorem 3.34. *Let $\mathbf{P}$ be a stochastic matrix indexed by a countable set $\mathbb{I}$, and π a probability distribution on $\mathbb{I}$. Let $\mathbf{Q}$ be a stochastic matrix indexed by $\mathbb{I}$ such that $\forall i, j \in \mathbb{I}$,*

$$\pi(i)q_{ij} = \pi(j)p_{ji}. \tag{3.15}$$

Then π is a stationary distribution of $\mathbf{P}$.

Proof. Fix i in $\mathbb{I}$ and sum over j in equation (3.15).

$$\begin{aligned} \sum_{j\in E} \pi(i)q_{ij} &= \sum_{j\in E} \pi(j)p_{ji} \\ \pi(i) &= \sum_{j\in E} \pi(j)p_{ji} \end{aligned}$$

Thus π is a stationary distribution. □

Definition 3.35. A HMC is *reversible* if there is a probability distribution π such that for all $i, j \in \mathbb{I}$ we have

$$\pi(i)p_{ij} = \pi(j)p_{ji}. \tag{3.16}$$

Equation (3.16) is a condition known as *detailed balance.* □

Corollary 3.36. *Let $\mathbf{P}$ by a transition matrix over $\mathbb{I}$, and π a probability distribution on $\mathbb{I}$. If for all $i, j \in \mathbb{I}$ we have detailed balance, then π is a stationary distribution.*

Proof. If $\mathbf{P}$ and π are such that detailed balance holds, then it follows from Theorem (3.34) that π is a stationary distribution. □

Example 3.37. Consider the random walk on the graph that is irreducible and aperiodic (3.29), and consider the probability distribution $\pi = (\frac{1}{9}, \frac{1}{6}, \frac{2}{9}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. The graph in (3.29) has the transition matrix

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \end{pmatrix}.$$

It is easy to check that π satisfies detailed balance, and to check that $\pi\mathbf{P} = \pi$. □

If we can find a probability distribution for our HMC that satisfies detailed balance, then it is the stationary distribution of our HMC. It still remains to show how we can find such a probability distribution. Further, we must determine whether such a distribution is unique, and whether or not this is the object that the chain will converge towards. First, we establish the notions of stopping times in order to create another classification of Markov chains.

3.4. **Stopping Times and the Strong Markov Property.**

Definition 3.38. A *stopping time* for a discrete stochastic process $\{X_n : \Omega \to \mathbb{I}\}_{n\geq 0}$ is a random variable $T : \Omega \to \mathbb{N} \cup \{\infty\}$ such that, for any $m \geq 0$, the event $\{T = m\}$ belongs to the σ-algebra $\mathscr{F}_m$. □

Remark 3.39. The condition $\{T = m\} \subset \mathscr{F}_m^\sigma$ signifies that at a given time m we can decide to stop the process, i.e., decide that $T = m$ using information about the evolution of the system up to and including the present moment m. □

The most important example of stopping time is a *return time*, which we now define.

Definition 3.40. The return time to a set $A \subset \mathbb{I}$ is defined as

$$T_A = \inf\{n \geq 1 : X_n \in A\},$$

where $T_A = \infty$ if $X_n \notin A$ for any $n \geq 1$. For $i \in \mathbb{I}$ we set $T_i := T_{\{i\}}$. □

Remark 3.41. Clearly the event $\{T_A = m\}$ can be decided from the knowledge of $X_0, X_1, \ldots, X_m$ so T_A is a stopping time. We will usually consider stopping times T_i for states $i \in \mathbb{I}$. Further we can define the r^{th} return time to a state $i \in \mathbb{I}$ as

$$T_i^{(r)} = \inf\{n > T_i^{(r-1)} : X_n = i\},$$

where $T_i^{(0)} = 0$, $T_i^{(1)} = T_i$. □

Theorem 3.42. *Denote by $T_i^{(r)}$ the r^{th} return time to $i \in \mathbb{I}$ and set*

$$a_0 := 0, \;\; a_1 := T_i^{(1)}, \;\; a_2 = T_i^{(2)} - T_i^{(1)}, \cdots .$$

Define

$$E := \left(\bigcup_{k=1}^{\infty}(\{k\} \times \mathbb{I}^k\right) \cup \Big(\{\infty\} \times \mathbb{I}^{\mathbb{N}} \Big),$$

Let

$$\xi_1 = (a_1, X_1, \ldots, X_{T_i^{(1)}}) \in E,$$

$$\xi_2 = \big(a_2, X_{T_i^{(1)}+1}, \ldots, X_{T_i^{(2)}}\big) \in E,$$

and, in general,

$$\xi_{n+1} = (a_{n+1}, X_{T_i^{(n)}+1}, \ldots, X_{T_i^{(n+1)}}), \in E, \;\; n = 1, 2, \ldots .$$

Fix $k \in \mathbb{N}$, $k \geq 2$. Denote by S_k the event

$$S_k := \big\{\, T_i^{(1)} < \infty, \ldots, T_i^{(k)} < \infty \,\big\}.$$

Note that S_k depends on the choice of the starting point i. Then the E-valued random variables $\xi_1, \xi_2, \ldots, \xi_k$ are i.i.d. with respect to the probability measure $\mathbb{P}_i^{S_k}(\cdot) := \mathbb{P}_i(\cdot|S_k)$ (on the event S_k).

Proof. We will prove for $k = 2$. We have that

$$\mathbb{P}_i(\xi_1 = (k, i_1, \ldots, i_k), \xi_2 = (l, j_1, \ldots, j_l), T_i^{(1)} < \infty, T_i^{(2)} < \infty)$$
$$= \mathbb{P}_i(a_1 = k, X_1 = i_1, \ldots, X_k = i_k, a_2 = l, X_{k+1} = j_1, \ldots, X_{k+l} = j_l, a_1 < \infty, a_2 < \infty).$$

Consider the case when

$$i \neq i_1, \ldots, i_{k-1}, j_1, \ldots, j_{l-1},$$
$$i_k = i = j_l.$$

because otherwise, the probability is 0. When we have this we get

$$= \mathbb{P}_i(X_1 = i_1, \ldots, X_{k-1} = i_{k-1}, X_k = i, X_{k+1} = j_1, \ldots X_{k+l-1} = j_{l-1}, X_{k+l} = i)$$
$$= \mathbb{P}_i(X_1 = i_1, \ldots, X_k = i)\mathbb{P}_i(X_{k+1} = j_1, \ldots, X_{k+l} = i | X_1 = i_1, \ldots, X_k = i).$$

And by the Markov Property we get that

$$= \mathbb{P}_i(X_1 = i_1, \ldots, X_k = 1)\mathbb{P}_i(X_1 = j_1, \ldots, X_l = i)$$
$$= \mathbb{P}_i(X_1 = i_1, \ldots, X_k = i, T_i^{(1)} = k)\mathbb{P}_i(X_1 = j_1, \ldots, X_l = i, T_i^{(1)} = l).$$

Summing over all $i_1, \ldots, i_k, j_1, \ldots, j_l$, we get that

$$\mathbb{P}_i(a_1 = k, a_2 = l) = \mathbb{P}_i(a_1 = k)\mathbb{P}_i(a_1 = l).$$

Thus we get that

$$\mathbb{P}_i(a_1 < \infty, a_2 < \infty) = \mathbb{P}_i(T_i^{(1)} < \infty, T_i^{(2)} < \infty) = \mathbb{P}_i(a_1 < \infty)^2.$$

Dividing this last equality through the first string of equalities, we get that

$$\mathbb{P}_i^{S(k)}(\xi_1 = (k, i_1, \ldots, i_k), \xi_2 = (l, j_1, \ldots, j_l))$$
$$= \mathbb{P}_i(\xi_1 = (k, i_1, \ldots, i_k)|a_1 < \infty)\mathbb{P}_i(\xi_1 = (l, j_1, \ldots, j_l)|a_1 < \infty)$$
$$= \mathbb{P}_i^{S(k)}(\xi_1 = (k, i_1, \ldots, i_k))\mathbb{P}_i^{S(k)}(\xi_1 = (l, j_1, \ldots, j_l)).$$

Thus ξ_1 and ξ_2 are i.i.d. □

Corollary 3.43. *For $i, j \in \mathbb{I}$, $i \neq j$, we have that, with respect to $\mathbb{P}_j^{S_k}(\cdot) = \mathbb{P}_j\big(\,\cdot\,|S_k(i)\,\big)$, that $\xi_2, \ldots, \xi_k$ are i.i.d. and also independent of ξ_1.* □

These results demonstrate evidence of the *Strong Markov Property*. A thorough proof of the Strong Markov Property is very involved, and beyond the scope of this paper. Instead of providing an unsatisfactory proof of it, we have included the previous results as building blocks suggesting the property, and state its final form in the following theorem.

Theorem 3.44 (Strong Markov Property)**.** *Suppose $\{X_n\}_{n\geq 0}$ is HMC $(\mathbf{P}, \mu)_{\mathbb{I}}$ and T is the first return time to $i \in \mathbb{I}$. Conditional on $T < \infty$, we have that $\{X_{T+n}\}_{n\geq 0}$ is Markov $(\mathbf{P}, \delta_i)_{|\mathbb{I}}$ independent of $X_0, X_1, \ldots, X_T$.* □

Remark 3.45. Note that the Strong Markov Property works for any stopping time, however we will only apply it for return times. □

Definition 3.46. We also denote the *time since the rth return time* to a state i in $\mathbb{I}$ as

$$S_i^{(r)} = \{T_i^{(r)} - T_i^{(r-1)} : T_i^{(r-1)} < \infty\}.$$

$S_i^{(r)} = 0$ when $T_i^{(r-1)} = \infty$. □

Lemma 3.47. *For $r = 2, 3, \ldots$, and conditional on $T_i^{(r-1)} < \infty$ we have that $S_i^{(r)}$ is independent of $\{X_m : m \leq T_i^{(r-1)}\}$ and*

$$\mathbb{P}_i(S_i^{(r)} = n | T_i^{(r-1)} < \infty) = \mathbb{P}_i(T_i = n).$$

Proof. Let $\{X_n\}_{n\geq 0}$ be HMC $(\mathbf{P}, \mu)_{\mathbb{I}}$. Apply the Strong Markov Property to $T = T_i^{(r-1)}$. Conditional on the fact that $T < \infty$, we have that $X_T = i$ and that $\{X_{T+n}\}_{n\geq 0}$ is Markov $(\mathbf{P}, \delta_i)$ and independent of $X_0, X_1, \ldots, X_T$. But

$$S_i^{(r)} = \inf\{n \geq 1 : X_{T+n} = i\}.$$

So we have that $S_i^{(r)}$ is the first return time to i of $\{X_{T+n}\}_{n\geq 0}$, which is HMC $(\mathbf{P}, \delta_i)_{\mathbb{I}}$. Therefore $S_i^{(r)}$ are stopping times, independent of the process before $T_i^{(r-1)}$, and have the same distribution as a stopping time starting from the initial distribution δ_i. □

Definition 3.48. Given a Markov chain $\{X_n\}_{n\geq 0}$ on a state space $\mathbb{I}$, we define the *number of visits* to a state $i \in \mathbb{I}$ as

$$N_i = \sum_{n\geq 1} \mathbb{1}_{\{X_n = i\}}. \tag{3.17}$$

Further, we define the *number of visits until time n* to a state $i \in \mathbb{I}$ as

$$N_i(n) = \sum_{k=0}^{n-1} \mathbb{1}_{\{X_n = i\}}.$$

□

Theorem 3.49. *Let $\{X_n\}_{n\geq 0}$ be a Markov chain. Given $X_0 = j$ for a state $j \in \mathbb{I}$, the distribution of N_i is*

$$\mathbb{P}_j(N_i = r) = \begin{cases} f_{ji} f_{ii}^{r-1}(1 - f_{ii}) & r \geq 1, \\ 1 - f_{ji} & r = 0. \end{cases}$$

where $f_{ji} = \mathbb{P}_j(T_i < \infty)$.

Proof. For $r = 0$, we have that the HMC never visits state i, which is equal to the probability $1 - f_{ji}$. We prove $r \geq 1$ by induction. Assume the equation is true for $k \in [1, r]$. We have that

$$\mathbb{P}_j(N_i > r) = 1 - \sum_{k=0}^{r} \mathbb{P}_j(N_i = k) = f_{ji} f_{ii}^r. \tag{3.18}$$

Define $T_r = T_i^{(r)}$, the r^{th} return time to state i.

$$\begin{aligned} \mathbb{P}_j(N_i = r+1) &= \mathbb{P}_j(N_i = r+1, X_{T_{r-1}} = i) \\ &= \mathbb{P}_j(T_{r+2} - T_{r+1} = \infty, X_{T_{r-1}} = i) \\ &= \mathbb{P}_j(T_{r+2} - T_{r+1} = \infty | X_{T_{r+1}} = i) \mathbb{P}_j(X_{T_{r+1}} = i) \end{aligned}$$

But, by the Strong Markov Property,

$$\begin{aligned}\mathbb{P}_j(T_{r+2} - T_{r+1} = \infty | X_{T_{r+1}} = i) &= \mathbb{P}(T_{r+2} - T_{r+1} = \infty | X_{T_{r+1}} = i, X_0 = j) \\ &= \mathbb{P}(T_{r+2} - T_{r+1} = \infty | X_{T_{r+1}} = i) \\ &= \mathbb{P}(T_i = \infty | X_0 = i)\end{aligned}$$

And since $\mathbb{P}_j(X_{T_{r+1}} = i) = \mathbb{P}_j(N_i > r)$, we get that

$$\begin{aligned}\mathbb{P}_j(N_i = r+1) &= \mathbb{P}_j(T_{r+2} - T_{r+1} = \infty | X_{T_{r+1}} = i)\mathbb{P}_j(X_{T_{r+1}} = i) \\ &= \mathbb{P}_i(T_i = \infty)\mathbb{P}_j(N_i > r) \\ &= (1 - f_{ii}) f_{ji} f_{ii}^r.\end{aligned}$$

Thus we have proved the equation by induction. □

Theorem 3.50. *For any $i \in \mathbb{I}$, we have that*

(i) $\mathbb{P}_i(T_i < \infty) = 1 \Leftrightarrow \mathbb{P}_i(N_i = \infty) = 1$,
(ii) $\mathbb{P}_i(T_i < \infty) < 1 \Leftrightarrow \mathbb{E}_i[N_i] < \infty \Leftrightarrow \mathbb{P}_i(N_i = \infty) = 0$.

Therefore we deduce that the probability $\mathbb{P}_i(N_i = \infty)$ has only two possible outcomes, 0 and 1.

Proof. $\mathbb{P}_i(T_i < \infty) = 1$ if and only if $f_{ii} = 1$. Therefore by equation (3.18) we have that

$$\lim_{r\to\infty} \mathbb{P}_i(N_i > r) = \lim_{r\to\infty} f_{ii}^{r+1} = 1.$$

Thus $\mathbb{P}_i(N_i = \infty) = 1$. Note that

$$\mathbb{E}_i[N_i] = \sum_{r=1}^{\infty} r\mathbb{P}_i(N_i = r) = \sum_{r=1}^{\infty} r f_{ii}^r (1 - f_{ii}) = \frac{f_{ii}}{1 - f_{ii}}.$$

Thus we have that

$$\mathbb{P}_i(T_i < \infty) < 1 \Leftrightarrow \frac{f_{ii}}{1 - f_{ii}} < \infty \Leftrightarrow \mathbb{E}_i[N_i] < \infty.$$

And if the expected value of N_i starting from i is less than ∞, we have that $\mathbb{P}_i(N_i = \infty) = 0$. Thus we have proven both items in the theorem, and see that the only possible values for $\mathbb{P}_i(N_i = \infty)$ are 0 and 1. □

3.5. **Recurrence.** The notion of stopping times allow us to define another class of Markov chains, those that are recurrent.

Definition 3.51. If $\mathbb{P}_i(T_i < \infty) = 1$, then we call the state $i \in \mathbb{I}$ *recurrent*. Otherwise we call it *transient*. If a state $i \in \mathbb{I}$ is recurrent and $\mathbb{E}_i[T_i] < \infty$ then we call it *positive recurrent*. Otherwise we call it *null recurrent*. □

A state i is recurrent if the return time to i when starting at i is almost surely finite. Further, i is positive recurrent if the expected return time is also finite. Recurrence and positive recurrence give guidelines as to how reasonable it is to get to a state in the chain.

Theorem 3.52. *The state $i \in \mathbb{I}$ is recurrent if and only if*

$$\sum_{n=0}^{\infty} p_{ii}(n) = \infty.$$

Proof. We have the following succesion of equivalent statements:

$$\text{The state } i \in \mathbb{I} \text{ is recurrent } \iff \mathbb{P}_i(T_i < \infty) = 1$$
$$\iff \mathbb{E}_i[N_i] = \infty \iff \mathbb{E}_i\left[\sum_{n\geq 1} \mathbb{1}_{\{X_n=i\}}\right] = \infty \iff \mathbb{E}_i\left[\sum_{n\geq 0} \mathbb{1}_{\{X_n=i\}}\right] = \infty$$
$$\iff \sum_{n=0}^{\infty} p_{ii}(n) = \infty.$$

□

Remark 3.53. Suppose $i \in \mathbb{I}$ is recurrent and accessible from a state $j \in \mathbb{I}$. This happens if and only if there exists an $m \geq 0$ such that $p_{ji}(m) > 0$ and $\sum_{n=0}^{\infty} p_{ii}(n) = \infty$, which happens if and only if $\sum_{n=0}^{\infty} p_{ji}(n) = \infty$. □

Corollary 3.54. *Suppose* $\mathbf{P}$ *is a transition matrix for an irreducible recurrent HMC* $\{X_n\}_{n\geq 0}$. *Then for any* $j \in \mathbb{I}$ *we have that*

$$\mathbb{P}(T_j < \infty) = 1.$$

Proof. If we consider all possible starting points $i \in \mathbb{I}$, we have that

$$\mathbb{P}(T_j < \infty) = \sum_{i\in\mathbb{I}} \mathbb{P}(X_0 = i)\mathbb{P}_i(T_j < \infty)$$

Therefore we must show that $\mathbb{P}_i(T_j < \infty) = 1$ for all $i \in \mathbb{I}$. Choose an m such that $p_{ji}(m) > 0$. Theorem (3.52) implies that

$$\begin{aligned} 1 &= \mathbb{P}_j(X_n = j \text{ for infinitely many } n) \\ &= \mathbb{P}_j(X_n = j \text{ for some } n \geq m+1) \\ &= \sum_{k\in\mathbb{I}} \mathbb{P}_j(X_n = j \text{ for some } n \geq m+1 | X_m = k)\mathbb{P}_j(X_m = k) \\ &= \sum_{k\in\mathbb{I}} \mathbb{P}_k(T_j < \infty)p_{jk}(m). \end{aligned}$$

Since $\sum_{i\in\mathbb{I}} p_{ji}(m) = 1$ we deduce that $\mathbb{P}_i(T_j < \infty) = 1$ for all $i \in \mathbb{I}$. □

Theorem 3.55. *If* i, j *communicate, then they are both recurrent or both transient.*

Proof. i and j communicating implies that there exist $M, N \geq 0$ such that

$$p_{ij}(M) > 0,$$
$$p_{ji}(N) > 0.$$

We can go from i to j in M steps, return to j in n steps, and go from j to i in N steps. Thus we have that

$$p_{ii}(M+n+N) \geq p_{ij}(M)p_{jj}(n)p_{ji}(N).$$

Similarly we derive

$$p_{jj}(N+n+M) \geq p_{ji}(N)p_{ii}(n)p_{ij}(M).$$

Let $\alpha = p_{ij}(M)p_{ji}(N) > 0$. We get that

$$p_{ii}(M+n+N) \geq \alpha p_{jj}(n),$$
$$p_{jj}(N+n+M) \geq \alpha p_{ii}(n).$$

Thus $\sum_{n=0}^{\infty} p_{ii}(n)$ is divergent $\Leftrightarrow \sum_{n=0}^{\infty} p_{jj}(n)$ is divergent, and i is recurrent $\Leftrightarrow j$ is recurrent. Similarly, i is transient $\Leftrightarrow j$ is transient. □

Remark 3.56. If a HMC is irreducible all states communicate, and therefore all states are either recurrent or transient. Therefore we can define an irreducible HMC as *recurrent* if the states are recurrent. □

3.6. **Invariant Measures.** We now establish the idea of invariant measures, and demonstrate their relationship to stationary distributions.

Definition 3.57. An *invariant measure* of a HMC with state space $\mathbb{I}$ and transition matrix $\mathbf{P}$ is a function $x : \mathbb{I} \to [0, \infty)$ such that, when viewed as a row vector it satisfies the equality

$$x\mathbf{P} = x, \text{ i.e., } x_i = \sum_{j \in \mathbb{I}} x_j p_{ji}, \quad \forall i \in \mathbb{I}.$$

□

Theorem 3.58. *Let* $\mathbf{P}$ *be the transition matrix of an irreducible HMC* $\{X_n\}_{n \geq 0}$*. Let* 0 *be an arbitrary state in* $\mathbb{I}$*, and* T_0 *the return time to* 0*. Define for all* $i \in \mathbb{I}$

$$x_i = \mathbb{E}_0 \left[\sum_{n \geq 1} \mathbb{1}_{\{X_n = i\}} \mathbb{1}_{\{n \leq T_0\}} \right].$$

In other words, x_i *is the expected number of visits to state* i *before returning to* 0*. Then, for any* $i \in \mathbb{I}$*, we have that* x *is an invariant measure of* $\mathbf{P}$ *and* $x_i \in (0, \infty)$ *for all* $i \in \mathbb{I}$*.*

Proof. We define

$$p_{0i}^0(n) := \mathbb{E}_0[\mathbb{1}_{\{X_n = i\}} \mathbb{1}_{\{n \leq T_0\}}] = \mathbb{P}_0(X_1 \neq 0, \ldots, X_{n-1} \neq 0, X_n = i).$$

Therefore $x_i = \sum_{n \geq 1} p_{0i}^0(n)$. For any number of steps $n \in [1, T_0]$, $X_n = 0 \Longleftrightarrow n = T_0$. Thus

$$x_0 = \sum_{n \geq 1} p_{00}^0(n) = 1.$$

We have that $p_{0i}^0(1) = p_{0i}$. For any $n \geq 2$ we can go from i to j in $(n-1)$ steps and then from j to i in one step, i.e.,

$$p_{0i}^0(n) = \sum_{j \neq 0} p_{0j}^0(n-1) p_{ji}.$$

Thus

$$x_i = \sum_{n \geq 1} p_{0i}^0(n) = p_{0i}^0(1) + \sum_{n \geq 2} p_{0i}^0(n)$$

$$= p_{0i} + \sum_{n \geq 2} \sum_{j \neq 0} p_{0i}^0(n-1) p_{ij} = p_{0i} \sum_{j \neq 0} \left[\sum_{n \geq 2} p_{0j}^0(n-1) \right] p_{ji}$$

$$= p_{0i} + \sum_{j \neq 0} x_j p_{ji} = x_0 p_{0i} + \sum_{j \neq 0} x_j p_{ji} = \sum_{j \in \mathbb{I}} x_j p_{ji},$$

where the second-last equality comes from the fact that $x_0 = 1$. This is true for any $i \in \mathbb{I}$, thus we have that $x = x\mathbf{P}$. Iterating this we get $x = x\mathbf{P}^n$. Therefore

$$x_i = \sum_{j \in \mathbb{I}} x_j p_{ji}(n) = p_{0i}(n) + \sum_{j \neq 0} x_j p_{ji}(n).$$

We have that $x_0 = 1 > 0$. Suppose that $x_i = 0$ for some $i \in \mathbb{I}$ such that $i \neq 0$. Then the above equality would imply that $p_{0i}(n) = 0$ for any $n \geq 0$. This would imply that 0 and i do not communicate, which is a contradiction since our HMC is irreducible. Hence $x_i > 0$ for all $i \in \mathbb{I}$. We also have that $1 = x_0 = \sum_{j\in\mathbb{I}} x_j p_{j0}(n)$ for any $n \geq 1$. Therefore, if there is some i such that $x_i = \infty$, then $p_{i0}(n) = 0$ for any $n \geq 1$. This contradicts the irreducibility. Hence x is an invariant measure and $x_i \in (0, \infty)$ for all $i \in \mathbb{I}$. □

Remark 3.59. Note that

$$\begin{aligned}\sum_{i\in\mathbb{I}}\sum_{n\geq 1} \mathbb{1}_{\{X_n=i\}}\mathbb{1}_{\{n\leq T_0\}} &= \sum_{n\geq 1}\left[\sum_{i\in\mathbb{I}} \mathbb{1}_{\{X_n=i\}}\right]\mathbb{1}_{\{n\leq T_0\}} \\ &= \sum_{n\geq 1} \mathbb{1}_{\{n\leq T_0\}} = T_0\end{aligned}$$

Therefore

$$\sum_{i\in\mathbb{I}} x_i = \mathbb{E}_0[T_0]. \tag{3.19}$$

□

Theorem 3.60. *Invariant measures of irreducible recurrent stochastic matrices are unique up to multiplication.*

Proof. We define a matrix $\mathbf{Q}$ by

$$q_{ji} = \frac{y_i}{y_j} p_{ij} \tag{3.20}$$

where y is an invariant measure of an irreducible recurrent stochastic matrix $\mathbf{P}$. $\mathbf{Q}$ makes sense since we know that $y_j > 0$ for any $j \in \mathbb{I}$ from the previous theorem. Also,

$$\sum_{i\in\mathbb{I}} q_{ji} = \frac{1}{y_j}\sum_{i\in\mathbb{I}} y_i p_{ij} = \frac{y_j}{y_j} = 1.$$

Suppose that $q_{ji}(n) = \frac{y_i}{y_j} p_{ij}(n)$. Then

$$\begin{aligned} q_{ji}(n+1) &= \sum_{k\in\mathbb{I}} q_{jk} q_{ki}(n) \\ &= \sum_{k\in\mathbb{I}} \frac{y_k}{y_j} p_{kj} \frac{y_i}{y_k} p_{ik}(n) \\ &= \frac{y_i}{y_j}\sum_{k\in\mathbb{I}} p_{ik}(n) p_{kj} \\ &= \frac{y_i}{y_j} p_{ij}(n+1).\end{aligned}$$

Thus by induction the matrix $\mathbf{Q}^n$ has the general term

$$q_{ji}(n) = \frac{y_i}{y_j} p_{ij}(n), \tag{3.21}$$

for all $n \geq 1$. Since $\mathbf{P}$ is irreducible, for any $i, j \in \mathbb{I}$ there exist $n \geq 0$ such that $p_{ij}(n) > 0$. But equation (3.21) implies that

$$p_{ij}(n) > 0 \leftrightarrow q_{ji}(n) > 0.$$

Thus $\mathbf{Q}$ is irreducible. Since $\mathbf{P}$ is recurrent, we have that $\sum_{n\geq 0} p_{ii}(n) = \infty$. But $q_{ii}(n) = p_{ii}(n)$ implies that $\sum_{n\geq 0} q_{ii}(n) = \sum_{n\geq 0} p_{ii}(n) = \infty$. Thus $\mathbf{Q}$ is recurrent as well. Define

$$g_{ji}(n) = \mathbb{P}(Y_0 = j, Y_1 \neq i, \ldots, Y_{n-1} \neq i, Y_n = i),$$

where $\{Y_n\}_{n\geq 0}$ is a HMC with respect to $\mathbf{Q}$. We get that

$$g_{i0}(n+1) = \sum_{j\neq 0} q_{ij} g_{j0}(n),$$

which implies that

$$y_i g_{i0}(n+1) = \sum_{j\neq 0} y_j g_{j0}(n) p_{ji}.$$

We recall that

$$p^0_{0i}(n+1) = \sum_{j\neq 0} p^0_{0i}(n) p_{ji},$$

which implies that

$$y_0 p^0_{0i}(n++1) = \sum_{j\neq 0} y_0 p^0_{0j} p_{ji}.$$

We have that

$$y_0 p^0_{0i}(1) = y_0 p_{0i} = y_i q_{i0} = y_i g_{i0}(1).$$

Thus it follows that

$$p^0_{0i}(n) = \frac{y_i}{y_0} g_{i0}(n),$$

for any $n \geq 1$. Summing across n on both sides we deduce that

$$\begin{aligned} \sum_{n\geq 1} p^0_{0i}(n) &= \sum_{n\geq 1} \frac{y_i}{y_0} g_{i0}(n) \\ x_i &= \frac{y_i}{y_0}. \end{aligned}$$

Thus x is a multiple of y. □

Theorem 3.61. *Positive recurrence and null recurrence are also class properties.*

Proof. We will see in the first part of Theorem (3.67) that for a recurrent irreducible Markov chain we have that $\lim_{n\to\infty} \frac{N_i}{n} = \frac{1}{\mathbb{E}_i[T_i]}$. Suppose that i is null recurrent, i.e. $\mathbb{E}_i[T_i] = \infty$. Since the chain is irreducible, we have that there exists $r, m \geq 0$ such that $p_{ij}(r) > 0, p_{ji}(m) > 0$. This implies that

$$\begin{aligned} 0 = \lim_{k\to\infty} \frac{\sum_{n=0}^{k} p_{ii}(n)}{k} &\geq \lim_{k\to\infty} \frac{\sum_{n=0}^{k-m-n} p_{jj}(n)}{k} p_{ij}(r) p_{ji}(m) && (3.22) \\ &= \lim_{k\to\infty} \frac{k-m-n}{k} \frac{\sum_{n=0}^{k-m-n} p_{jj}(n)}{k-m-n} p_{ij}(r) p_{ji}(m) && (3.23) \\ &= \lim_{l\to\infty} \frac{\sum_{n=0}^{l} p_{jj}(n)}{l} p_{ij}(r) p_{ji}(m) && (3.24) \\ &= \frac{p_{ij}(r) p_{ji}(m)}{\mathbb{E}_j[T_j]}. && (3.25) \end{aligned}$$

Which implies that $\mathbb{E}_j[T_j] = \infty$, i.e. that j is null recurrent. □

Theorem 3.62. *An irreducible recurrent HMC is positive recurrent if and only if its invariant measure x satisfies*

$$\sum_{i\in\mathbb{I}} x_i < \infty.$$

Proof. From equation (3.19) we have that $\sum_{i\in\mathbb{I}} x_i = \mathbb{E}_0[T_0]$. State 0 is positive recurrent if and only if $\mathbb{E}_0[T_0] < \infty$, that is, $\sum_{i\in\mathbb{I}} x_i < \infty$. Thus state 0 is positive recurrent if the inequality holds. Since the HMC is irreducible, the HMC is positive recurrent. □

If we have an irreducible HMC, then we have an invariant measure, which is unique up to multiplication by a constant. From this, we are able to determine whether or not the HMC is positive recurrent.

Lemma 3.63. *(Dominated Convergence Theorem) Let $\{a_{nk}\}_{n\geq 1,k\geq 1}$ be an array of real numbers such that for some sequence of non-negative real numbers $\{b_k\}_{k\geq 1}$ satisfying $\sum_{k=1}^{\infty} b_k < \infty$, we have that*

$$|a_{nk}| \leq b_k.$$

Moreover, if we have

$$\lim_{n\to\infty} a_{nk} = a_k,$$

then

$$\lim_{n\to\infty} \sum_{k=1}^{\infty} a_{nk} = \sum_{k=1}^{\infty} \lim_{n\to\infty} a_{nk} = \sum_{k=1}^{\infty} a_k.$$

□

Theorem 3.64. *An irreducible HMC is positive recurrent if and only if there exists a stationary distribution π. When the stationary distribution π exists, it is unique and $\pi > 0$.*

Proof. If a HMC is irreducible and positive recurrent, we have the existence of a stationary distribution by taking a multiple of the invariant measure. Conversely, assume there exists a stationary distribution π. Iterating the equation $\pi = \pi\mathbf{P}$, we get $\pi = \pi\mathbf{P}^n$. That is, for all $i \in \mathbb{I}$,

$$\pi(i) = \sum_{j\in\mathbb{I}} \pi(j)p_{ji}(n).$$

If the HMC were transient, we have that for any $i, j \in \mathbb{I}$,

$$\lim_{n\to\infty} p_{ji}(n) = 0.$$

Since $p_{ji}(n)$ is bounded by 1, the dominated convergence theorem (3.63) implies that

$$\pi(i) = \lim_{n\to\infty} \sum_{j\in\mathbb{I}} \pi(j)p_{ji}(n) = \sum_{j\in\mathbb{I}} \pi(j)(\lim_{n\to\infty} p_{ji}(n)) = 0.$$

Thus π is not a stationary distribution, since we must have $\sum_{i\in\mathbb{I}} \pi(i) = 1$. Therefore the chain is recurrent. And since the sum equals 1, i.e. is finite, we have positive recurrence. π is unique when we take the invariant measure constant to be 1. Further, viewing π as an invariant measure, we see that $\pi(i) > 0$ for all $i \in \mathbb{I}$. □

If we have an irreducible HMC, we can form the invariant measure and test whether or not the chain is positive recurrent. If it is, we have a stationary distribution, which is unique and strictly positive. Also, if we have an irreducible HMC and a stationary distribution (perhaps by detailed balance), we have that the chain is positive recurrent, and again that the stationary distribution is unique and strictly positive. The important thing is that for irreducible chains,

positive recurrence and unique stationary distributions go hand in hand, giving us a criterion for the uniqueness of stationary distributions. We will now see that with such assumptions we get a nice relation between stationary distributions and the expected value of a return time.

Theorem 3.65. *Let π be the unique stationary distribution of an irreducible positive recurrent HMC. Let T_i be the return time to a state $i \in \mathbb{I}$. Then*

$$\pi(i)\mathbb{E}_i[T_i] = 1.$$

Proof. We obtain π by normalization of the invariant measure x.

$$\pi(i) = \frac{x_i}{\sum_{j\in\mathbb{I}} x_j}.$$

For $i = 0$

$$\pi(0) = \frac{x_0}{\sum_{j\in\mathbb{I}} x_j} = \frac{1}{\mathbb{E}_0[T_0]}.$$

Recall that we picked $0 \in \mathbb{I}$ arbitrarily, therefore we get that

$$\pi(i)\mathbb{E}_i[T_i] = 1,$$

for any $i \in \mathbb{I}$. □

Theorem 3.66. *An irreducible HMC with finite state space is positive recurrent.*

Proof. The chain is recurrent because if it were transient then for any $i, j \in \mathbb{I}$ we would have that

$$\sum_{n\geq 0} p_{ij}(n) < \infty.$$

Since the state space is finite, we also have that

$$\sum_{j\in\mathbb{I}}\sum_{n\geq 0} p_{ij}(n) < \infty.$$

However

$$\sum_{n\geq 0}\sum_{j\in\mathbb{I}} p_{ij}(n) = \sum_{n\geq 0} 1 = \infty,$$

which is a contradiction. Therefore the chain is recurrent. Since we have an irreducible chain, we have an invariant measure. And since the state space $\mathbb{I}$ is finite, we have that $\sum_{i\in\mathbb{I}} x_i < \infty$, which implies positive recurrence. □

A lot of applications of Markov chains deal with finite state spaces. Given this theorem, we need only check that the chain is irreducible to establish the existence of a unique stationary distribution. We now establish one of our most important theorems thus far, the Ergodic Theorem.

3.7. **Ergodic Theorem.**

Theorem 3.67 (Ergodic Theorem)**.** *Let $\{X_n\}_{n\geq 0}$ be an irreducible HMC $(\mathbf{P}, \mu)_{\mathbb{I}}$. Define $m_i = \mathbb{E}_i[T_i]$. Then, a.s.*

$$\lim_{n\to\infty} \frac{N_i}{n} = \frac{1}{m_i}.$$

Further, when $\{X_n\}_{n\geq 0}$ is positive recurrent, with stationary distribution π, then for any bounded function $f : \mathbb{I} \to \mathbb{R}$, a.s.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) = \bar{f},$$

where $\bar{f} = \sum_{i\in\mathbb{I}} \pi(i)f(i)$.

Proof. If $\{X_n\}_{n\geq 0}$ is transient, then a.s. $N_i < \infty$, so

$$\frac{N_i(n)}{n} \leq \frac{N_i}{n} \to 0 \leftarrow \frac{1}{m_i},$$

since $\mathbb{E}_i[T_i] = \infty$.

Suppose $\{X_n\}_{n\geq 0}$ is recurrent. Fix a state $i \in \mathbb{I}$. Recurrence implies that $\mathbb{P}(T_i < \infty) = 1$. Theorem(3.42) shows that the long run proportion $\frac{N_i}{n}$ is the same for $\{X_n\}_{n\geq 0}$ and $\{X_{T+n}\}_{n\geq 0}$, so we can consider $\mu = \delta_i$. By Lemma (3.47), we have that

$$S_i^{(1)}, S_i^{(2)}, \dots$$

are i.i.d. with $\mathbb{E}_i[T_i] = m_i$. We have the relationship that

$$S_i^{(1)} + \cdots + S_i^{(N_i-1)} \leq n - 1,$$

since the length of time spent away from i up until the $(N_i(n) - 1)$th visit cannot exceed $n - 1$. By similar reasoning we have that

$$S_i^{(1)} + \cdots + S_i^{(N_i(n))} \geq n.$$

Therefore, dividing by $N_i(n)$, we get

$$\frac{S_i^{(1)} + \cdots + S_i^{(N_i-1)}}{N_i(n)} \leq \frac{n}{N_i(n)} \leq \frac{S_i^{(1)} + \cdots + S_i^{(N_i)}}{N_i(n)}. \tag{3.26}$$

By Theorem (2.4) we have that

$$\frac{S_i^{(1)} + \cdots + S_i^{(n)}}{n} \to m_i.$$

Since $\{X_n\}_{n\geq 0}$ is recurrent, $N_i(n) \to \infty$ as $n \to \infty$. Therefore by the squeezing principle

$$\frac{n}{N_i(n)} \to m_i.$$

Suppose $\{X_n\}_{n\geq 0}$ is positive recurrent as well. This implies that $m_i = \mathbb{E}_i[T_i] = \frac{1}{\pi(i)}$. Therefore we have that

$$\frac{N_i(n)}{n} \to \pi(i).$$

Let $f : \mathbb{I} \to \mathbb{R}$ be bounded. Without a loss of generality assume that $|f| \leq 1$. If f were bounded by some M, then we will have a factor of M in our inequality and would pick an ϵ accordingly.

For any $J \subset \mathbb{I}$, we have that

$$\begin{aligned}
\left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \bar{f}\right| &= \left|\sum_{i\in\mathbb{I}}\left(\frac{N_i(n)}{n} - \pi(i)\right) f(i)\right| \\
&\leq \sum_{i\in\mathbb{I}}\left|\frac{N_i(n)}{n} - \pi(i)\right| \\
&\leq \sum_{i\in J}\left|\frac{N_i(n)}{n} - \pi(i)\right| + \sum_{i\in J^C}\left|\frac{N_i(n)}{n} - \pi(i)\right| \\
&\leq \sum_{i\in J}\left|\frac{N_i(n)}{n} - \pi(i)\right| + \sum_{i\in J^C}\left(\frac{N_i(n)}{n} + \pi(i)\right) \\
&\leq \sum_{i\in J}\left|\frac{N_i(n)}{n} - \pi(i)\right| + \sum_{i\in J^C}\frac{N_i(n)}{n} + \sum_{i\in J^C}\pi(i).
\end{aligned}$$

Denote $N_n(J) = \sum_{i\in J}\frac{N_i(n)}{n}$, $\pi(J) = \sum_{i\in J}\pi(i)$.

- We can pick $J \subset \mathbb{I}$ such that $\pi(J^C) < \frac{\epsilon}{4}$.
- We can choose N such that for any $n \geq N$ we have that$\sum_{i\in J}|\frac{N_i(n)}{n} - \pi(i)| = N_n(J) - \pi(J) < \frac{\epsilon}{4}$.
- $|N_n(J^C) - \pi(J^C)| = |N_n(J) - \pi(J)|$, which implies that $N_n(J^C) \leq \pi(J^C) + |N_n(J) - \pi(J)| \leq \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}$.

Thus,

$$\left|\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) - \bar{f}\right| \leq \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{2} = \epsilon.$$

□

For an irreducible positive recurrent HMC (or an irreducible HMC over a finite state space), and a well behaved function f, we have a convergence relation. We have proven Nekrosov wrong! However we cannot stop here, as we will deduce yet another convergence property of Markov chains.

4. Convergence To Stationary Distributions

4.1. **Distance in Variation.** First, we establish a way to measure the distance between probability distributions, and derive some important consequences.

Definition 4.1. Let $\mathbb{I}$ be a countable set, α, β probability distributions on $\mathbb{I}$. The *distance in variation* between α and β is

$$d(\alpha, \beta) = \frac{1}{2}|\alpha - \beta| = \frac{1}{2}\sum_{i\in\mathbb{I}}|\alpha(i) - \beta(i)|,$$

The distance in variation between two random variables X, Y with values in $\mathbb{I}$ and distributions $\mathscr{L}(X), \mathscr{L}(Y)$ is

$$d(\mathscr{L}(X), \mathscr{L}(Y)) =: d(X, Y).$$

□

Lemma 4.2. *Let X, Y be two random variables with values in a countable space $\mathbb{I}$. Then*

$$\sup_{A\subset\mathbb{I}} \big|\, \mathbb{P}(X\in A) - \mathbb{P}(Y\in A)\,\big| = \sup_{A\subset\mathbb{I}} \big\{\, \mathbb{P}(X\in A) - \mathbb{P}(Y\in A)\,\big\} = d(X,Y).$$

Proof. If $\mathbb{P}(X\in A) - \mathbb{P}(Y\in A) < 0$, then let $B = A^C$, and we get that

$$-\big(\,\mathbb{P}(X\in A) - \mathbb{P}(Y\in A)\big) = \mathbb{P}(X\in B) - \mathbb{P}(Y\in B) > 0.$$

Thus we have the first equality. We have that

$$\mathbb{P}(X\in A) - \mathbb{P}(Y\in A) = \sum_{i\in\mathbb{I}} \mathbb{1}_A(i)\big(\,\mathbb{P}(X=i) - \mathbb{P}(Y=i)\,\big).$$

The right hand side is maximal for the set

$$\widetilde{A} = \big\{ i\in\mathbb{I};\ \ \mathbb{P}(X=i) > \mathbb{P}(Y+i)\,\big\}.$$

Since

$$\sum_{i\in\mathbb{I}} \left(\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right) = 0, \tag{4.1}$$

we deduce that, for any $A\subset\mathbb{I}$, we have

$$\sum_{i\in\mathbb{I}} \mathbb{1}_A(i)\left(\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right) + \sum_{i\in\mathbb{I}} \mathbb{1}_{A^C}(i)\left(\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right) = 0.$$

On $\widetilde{A}$ we have that

$$\mathbb{P}(X=i) - \mathbb{P}(Y=i) = \big|\,\mathbb{P}(X=i) - \mathbb{P}(Y=i)\,\big|,$$

and on $\widetilde{A}^C$ we have that

$$\mathbb{P}(X=i) - \mathbb{P}(Y=i) = -\big|\,\mathbb{P}(X=i) - \mathbb{P}(Y=i)\,\big|.$$

Therefore for $\widetilde{A}$, we have that

$$\begin{aligned}
\sum_{i\in\mathbb{I}} \mathbb{1}_{\widetilde{A}}(i)\left(\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right) &= \sum_{i\in\mathbb{I}} \mathbb{1}_{\widetilde{A}}(i)\left|\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right| \\
&= \sum_{i\in\mathbb{I}} \mathbb{1}_{\widetilde{A}^C}(i)\left|\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right| \\
&= \frac{1}{2}\sum_{i\in\mathbb{I}} \left|\mathbb{P}(X=i) - \mathbb{P}(Y=i)\right|,
\end{aligned}$$

where both equalities come from equation (4.1). □

Definition 4.3. For probability distributions α, β on a countable set $\mathbb{I}$, let $\mathscr{D}(\alpha,\beta)$ be the collection of *couplings* of α with β, i.e., random vectors (X,Y) taking values in $\mathbb{I}\times\mathbb{I}$, such that the marginal distribution of X is α and the marginal distribution of Y is β. □

Theorem 4.4. *For any $(X,Y)\in\mathscr{D}(\alpha,\beta)$, we have that*

$$\mathbb{P}(X\neq Y) \geq d(\alpha,\beta).$$

Proof. For any $A\subset\mathbb{I}$, we have that

$$\begin{aligned}
\mathbb{P}(X\neq Y) &\geq \mathbb{P}(X\in A, Y\in A^C) = \mathbb{P}(X\in A) - \mathbb{P}(X\in A^C, Y\in A) \\
&\geq \mathbb{P}(X\in A) - \mathbb{P}(Y\in A).
\end{aligned}$$

This implies that

$$\mathbb{P}(X \neq Y) \geq \sup_{A \subset \mathbb{I}} \{\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\} = d(\alpha, \beta).$$

□

4.2. **Convergence.**

Definition 4.5. Let $\{\alpha_n\}_{n\geq 0}, \beta$ be probability distributions on a countable set $\mathbb{I}$. If

$$\lim_{n\to\infty} d(\alpha_n, \beta) = 0,$$

we say that $\{\alpha_n\}_{n\geq 0}$ *converges in variation* to β. Let $\{X_n\}_{n\geq 0}$ be an $\mathbb{I}$-valued stochastic process. If π is some probability distribution on $\mathbb{I}$, and if the distribution $\mathscr{L}(X_n)$ of the random variable X_n converges in variation to π, that is, if

$$\lim_{n\to\infty} \sum_{i\in\mathbb{I}} |\mathbb{P}(X_n = i) - \pi(i)| = 0,$$

then $\{X_n\}_{n\geq 0}$ *converges in variation* to π. □

Remark 4.6. Recall that

$$\bar{f} = \sum_{i\in\mathbb{I}} \pi(i) f(i),$$

for some bounded $f : \mathbb{I} \to \mathbb{R}$ and M is an upper bound of $|f|$. Then

$$\left| \mathbb{E}[f(X_n)] - \bar{f} \right| = \left| \sum_{i\in\mathbb{I}} f(i)\big(\mathbb{P}(X_n = i) - \pi(i)\big) \right| \leq M \sum_{i\in\mathbb{I}} |\mathbb{P}(X_n = i) - \pi(i)|.$$

Thus $\{X_n\}_{n\geq 0}$ converges in variation to $\pi \Leftrightarrow \lim_{n\to\infty} \mathbb{E}[f(X_n)] = \bar{f}$. □

Definition 4.7. Two stochastic processes $\{X'_n\}_{n\geq 0}, \{X''_n\}_{n\geq 0}$ taking values in $\mathbb{I}$ are said to *couple* if there exists almost surely a finite random time T such that

$$n \geq T \Rightarrow X'_n = X''_n.$$

T is called a *coupling time* of $\{X'_n\}_{n\geq 0}$ and $\{X''_n\}_{n\geq 0}$. □

Theorem 4.8. *For any coupling time T of $\{X'_n\}_{n\geq 0}$ and $\{X''_n\}_{n\geq 0}$, we have that*

$$d\big(X'_n, X''_n\big) \leq \mathbb{P}(T > n), \quad \forall n \geq 0.$$

Proof. For all $A \subset \mathbb{I}$

$$\begin{aligned} &\mathbb{P}(X'_n \in A) - \mathbb{P}(X''_n \in A) \\ &= \mathbb{P}(X'_n \in A, T > n) - \mathbb{P}(X''_n \in A, T > n) + \mathbb{P}(X'_n \in A, T \leq n) - \mathbb{P}(X''_n \in A, T \leq n) \\ &= \mathbb{P}(X'_n \in A, T > n) - \mathbb{P}(X''_n \in A, T > n) \leq \mathbb{P}(X'_n \in A, T > n) \leq \mathbb{P}(T > n), \end{aligned}$$

where the first equality comes from T being a coupling time. □

We so far have three ideal characteristics of Markov chains. Aperiodicity, irreducibility, and positive recurrence. We summarize chains that meet all three conditions in the following definition.

Definition 4.9. A HMC, along with its transition matrix, that is irreducible, positive recurrent, and aperiodic is called *ergodic*. □

Remark 4.10. Note that a HMC need not be ergodic for the Ergodic Theorem (does not need aperiodicity). This definition may be confusing, but is nevertheless standard in various books. □

Theorem 4.11 (Convergence to Stationary Distribution)**.** *Let* $\mathbf{P}$ *be the transition matrix for an ergodic Markov chain. For all probability distributions* μ, v *on* $\mathbb{I}$*, we have that*

$$\lim_{n\to\infty} d(\mu\mathbf{P}^n, v\mathbf{P}^n) = 0.$$

Further since $\mathbf{P}$ *is ergodic, there exists a stationary distribution* π *with the consequence that for any initial distribution* μ *we have that*

$$\lim_{n\to\infty} |\mu\mathbf{P}^n - \pi| = 0.$$

Proof. Let $\{X_n^{(1)}\}_{n\geq 0}$, $\{X_n^{(2)}\}_{n\geq 0}$ be HMC $(\mathbf{P}, \mu)_{\mathbb{I}}, (\mathbf{P}, v)_{\mathbb{I}}$, respectively, and independent of each other. For some $b \in \mathbb{I}$, consider the stopping time

$$T = \inf\{n \geq 0 : X_n^{(1)} = X_n^{(2)} = b\}.$$

We refer to T as *coupling time*.

Step 1. Show $\mathbb{P}(T < \infty) = 1$.

Consider the chain $\{Z_n\}_{n\geq 0}$ given by $Z_n = (X_n^{(1)}, X_n^{(2)})$, which we will show is a HMC on $\mathbb{I} \times \mathbb{I}$. Z_n has a transition matrix $\mathbf{P}'$ with entries

$$p_{(i,k)(j,l)} = p_{ij}p_{kl},$$

and initial distribution $\lambda_{(i,k)} = \mu(i)v(k)$. Therefore Z_n is HMC $(\mathbf{P}', \lambda)$ where λ is a probability distribution on $\mathbb{I} \times \mathbb{I}$. Since $X_n^{(1)}, X_n^{(2)}$ are both irreducible and aperiodic, we have by Theorem (3.27) that there exists an $n_0 \geq 0$ such that for any $n \geq n_0$

$$p_{ik}(n) > 0,$$

$$p_{jl}(n) > 0,$$

which implies that $p_{(i,k)(jl)}(n) = p_{ik}(n)p_{jl}(n) > 0$, and therefore Z_n is irreducible. Since $X_n^{(1)}, X_n^{(2)}$ are aperiodic, Z_n is also aperiodic. Z_n inherits a stationary distribution $\{\pi(i)\pi(j)\}_{(i,j)\in\mathbb{I}\times\mathbb{I}}$. Therefore Z_n is positive recurrent. Since T is a return time to (b, b) for Z_n, which is positive recurrent, we have that $\mathbb{P}(T < \infty) = 1$.

Step 2. Create a coupling HMC.

Consider the process $\{Y_n\}_{n\geq 0}$ given by

$$Y_n = \begin{cases} X_n^{(1)} & n \leq T, \\ X_n^{(2)} & n > T. \end{cases}$$

Apply the Strong Markov Property to Z_n and we see that $\{X_{n+T}^{(1)}, X_{n+T}^{(2)}\}_{n\geq 0}$ is HMC $(\mathbf{P}', \delta_{(b,b)})$ and independent of $(X_0^{(1)}, X_0^{(2)}), \ldots, (X_T^{(1)}, X_T^{(2)})$. Similarly, by symmetry, $\{X_{n+T}^{(2)}, X_{n+T}^{(1)}\}_{n\geq 0}$ is HMC $(\mathbf{P}', \delta_{(b,b)})$ independent of $(X_0^{(2)}, X_0^{(1)}), \ldots, (X_T^{(2)}, X_T^{(1)})$. Hence if

$$Y_n' = \begin{cases} X_n^{(2)} & n \leq T, \\ X_n^{(1)} & n > T, \end{cases}$$

then (Y_n, Y_n') is HMC $(\mathbf{P}', \lambda)$, which implies that Y_n is HMC $(\mathbf{P}, \mu)$.

Step 3. Y_n and $X_n^{(2)}$ couple, and by Theorem (4.8), we get that

$$d(\mu\mathbf{P}^n, v\mathbf{P}^n) = d(Y_n, X_n^{(2)}) \leq \mathbb{P}(T > n) \to 0,$$

as $n \to \infty$. Further, since $\mathbf{P}$ is ergodic, we have a stationary distribution π, and if we let $v = \pi$, then $\pi\mathbf{P}^n = \pi$, and we get the convergence to the stationary distribution. □

4.3. **Rate of Convergence.** Now that we have established a notion of convergence for ergodic Markov chains, one wonders how quickly this convergence may occur. We establish a rate with the following theorem.

Definition 4.12 (E. Landau)**.** The notation $f(n) = o(g(n))$ means that

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

The notation $f(n) = O\big(\, g(n)\,\big)$ means that there exists $c > 0$ such that

$$|f(n)| \leq c|g(n)|, \quad \forall n.$$ □

Theorem 4.13. *Suppose that the coupling time T defined in (4.2) satisfies*

$$\mathbb{E}[\phi(T)] < \infty,$$

for some non-decreasing function $\phi : \mathbb{N} \to \mathbb{R}^+$ such that

$$\lim_{n\to\infty} \phi(n) = \infty.$$

Then, for any initial distributions μ, ν, we have that

$$|\mu\mathbf{P}^n - \nu\mathbf{P}^n| = o\left(\frac{1}{\phi(n)}\right).$$

Proof. Since ϕ is non-decreasing, we have that $\phi(T)\mathbb{1}_{\{T>n\}} \geq \phi(n)\mathbb{1}_{\{T>n\}}$. Thus

$$\mathbb{P}(T > n)\phi(n) \leq \mathbb{E}[\phi(T)\mathbb{1}_{\{T>n\}}].$$

Since T is finite we have that $\lim_{n\to\infty} \phi(T)\mathbb{1}_{\{T>n\}} = 0$ And since $\phi(T)\mathbb{1}_{\{T>n\}}$ is bounded by $\phi(T)$, which is integrable, Lemma (3.63) implies that

$$\lim_{n\to\infty} \mathbb{E}[\phi(T)\mathbb{1}_{\{T>n\}}] = 0.$$

Hence

$$\phi(n)|\mu\mathbf{P}^n - \nu\mathbf{P}^n| \leq \phi\mathbb{P}(T > n) \to 0.$$

Hence for any $c > 0$ we have $n_0 > 0$ such that for any $n \geq n_0$,

$$\phi(n)|\mu\mathbf{P}^n - \nu\mathbf{P}^n| \leq c \Rightarrow |\mu\mathbf{P}^n - \nu\mathbf{P}^n| \leq c\frac{1}{\phi(n)}.$$

That is, $|\mu\mathbf{P}^n - \nu\mathbf{P}^n| = o\Big(\frac{1}{\phi(n)}\Big)$. □

4.4. **Eigenvalues of the Transition Matrix.** Up until now, we have been thinking of all vectors as *row* vectors. For the entirety of this section, since we will be dealing with both column vectors and row vectors, denote a vector written v as a column vector and a vector written v^T as a row vector.

Definition 4.14. A square matrix A is called *positive/non-negative*, and we indicate this using the notation $A > 0$ (respectively$A \geq 0$) if all its entries are positive/non-negative. A non-negative square matrix A is called *primitive* if there exists a natural number $k > 0$ such that $A^k > 0$. □

Theorem 4.15 (Peron-Frobenius)**.** *Let A be a non-negative primitive $r \times r$ matrix. We denote by* $\operatorname{spec}(A)$ *its spectrum (the set of all its eigenvalues). The multiplicity of an eigenvalue is its multiplicity as a root of the characteristic polynomial. Then the following hold:*

(i) *The spectral radius of A defined by*
$$\max_{\lambda \in \operatorname{spec}(A)} |\lambda| \in (0, \infty),$$
is a simple eigenvalue of A. We denote it by λ_1.

(ii) *If $\lambda \in \operatorname{spec}(A) \setminus \{\lambda_1\}$, then $|\lambda| < \lambda_1$.*

(iii) *The left eigenvector u_1 and the right eigenvector v_1 of A corresponding to the eigenvalue λ_1 can be chosen such that they are positive and*
$$u_1^T v_1 = 1.$$

(iv) *Order* $\operatorname{spec}(A) \setminus \{\lambda_1\}$ *as $\lambda_2, \dots \lambda_r$ so that*
$$\lambda_1 > |\lambda_2| \geq \ldots \geq |\lambda_r|.$$
and if $|\lambda_2| = |\lambda_j|$ for $j \geq 3$, then the multiplicity of λ_2 is greater or equal that the multiplicity of λ_j. Then
$$A^n = \lambda_1^n v_1 u_1^T + O(n^{m_2 - 1} |\lambda_2|^n),$$
where m_2 is the multiplicity of λ_2.

□

Remark 4.16. If we consider a transition matrix $\mathbf{P}$ on $\mathbb{I} = \{1, \dots, r\}$, and $\mathbf{P}$ is irreducible and aperiodic, then there is some n such that $\mathbf{P}^n > 0$. For we know that if it is irreducible and aperiodic, then there is an n such that we can travel between states with positive probability. $\mathbf{P}$ has a unique stationary distribution π, and $u_1 = \pi$, $v_1 = \mathbb{1}$, the vector consisting of all 1's. Therefore
$$\mathbf{P}^n = \mathbb{1} \cdot \pi^T + O\big(n^{m_2 - 1} |\lambda_2|^n \big).$$
This gives us an estimate of how long convergence to a stationary distribution may take, however this estimate requires a lot of knowledge about the structure of $\mathbf{P}$. For reversible transition matrices, we will be able to develop better bounds, and in order to do that we must delve into some vector space theory. □

Definition 4.17. Let $\mathbf{P}$ be an irreducible transition matrix on the finite[1] state space $\mathbb{I} = \{1, \dots, r\}$. Let π be a strictly positive probability distribution on $\mathbb{I}$. We denote by $\mathscr{L}^2(\pi)$ to be the Hilbert space of functions $x : \mathbb{I} \to \mathbb{R}$ equipped with the scalar product
$$\langle x, y \rangle_\pi := \sum_{i \in \mathbb{I}} x(i) y(i) \pi(i),$$

[1]The finiteness of $\mathbb{I}$ implies positive recurrence.

and norm

$$||x||_\pi := \left(\sum_{i \in \mathbb{I}} x(i)^2 \pi(i) \right)^{\frac{1}{2}}.$$

The π-*mean* of a function $x : \mathbb{I} \to \mathbb{R}$ is the real number

$$\langle x \rangle_\pi := \sum_{i \in \mathbb{I}} x(i)\pi(i) = \langle x, 1 \rangle_\pi,$$

while its π-*variance* is

$$Var_\pi(x) := ||x||_\pi^2 - \langle x \rangle_\pi^2.$$

We denote by $\mathscr{L}^2(\frac{1}{\pi})$ to be the dual of $\mathbb{R}^r \cong \mathbb{R}^{\mathbb{I}}$ equipped with scalar product

$$\langle x^T, y^T \rangle_{\frac{1}{\pi}} := \sum_{i \in \mathbb{I}} x(i)y(i)\frac{1}{\pi(i)}.$$

Note that $\langle \cdot, \cdot \rangle_\pi$ takes *column* vectors and $\langle \cdot, \cdot \rangle_{\frac{1}{\pi}}$ takes *row* vectors. □

Theorem 4.18. *The transition matrix* $\mathbf{P}$ *is reversible (see Definition 3.35) with stationary probability distribution* π *if and only if* $\mathbf{P}$ *is self adjoint in* $\mathscr{L}^2(\pi)$*, i.e.,*

$$\langle \mathbf{P}x, y \rangle_\pi = \langle x, \mathbf{P}y \rangle_\pi, \quad \forall x, y \in \mathscr{L}^2(\pi).$$

Proof. Suppose $\mathbf{P}$ is reversible with invariant probability distribution π. Then

$$\begin{aligned}
\langle \mathbf{P}x, y \rangle_\pi &= \sum_{i \in \mathbb{I}} \left(\sum_{j \in \mathbb{I}} p_{ij} x(j) \right) y(i)\pi(i) \\
&= \sum_{i,j \in \mathbb{I}} \pi(i) p_{ij} x(j) y(i) \\
&= \sum_{i,j \in \mathbb{I}} \pi(j) p_{ji} y(i) x(j) \\
&= \sum_{j \in \mathbb{I}} x(j) \left(\sum_{i \in \mathbb{I}} p_{ji} y(i) \right) \pi(j) \\
&= \langle x, \mathbf{P}y \rangle_\pi.
\end{aligned}$$

Hence $\mathbf{P}$ is self adjoint. Now suppose that $\mathbf{P}$ is self adjoint in $\mathscr{L}^2(\pi)$. Pick $x = \delta_i$, $y = \delta_j$. Then we get that

$$\pi(i)p_{ij} = \langle \mathbf{P}\delta_i, \delta_j \rangle_\pi = \langle \delta_i, \mathbf{P}\delta_j \rangle_\pi = \pi(j)p_{ji}.$$

Hence $\mathbf{P}$ is reversible with distribution π. □

Remark 4.19. We can also deduce that $\mathbf{P}$ is reversible with probability distribution π if and only if the matrix

$$\mathscr{P}^* := D^{\frac{1}{2}} \mathbf{P} D^{-\frac{1}{2}}$$

is symmetric, where $D = \text{diag}\{\pi(1), \ldots, \pi(r)\}$. Indeed $\mathscr{P}^*$ is symmetric if and only if

$$\frac{\sqrt{\pi(i)}p_{ij}}{\sqrt{\pi(j)}} = \frac{\sqrt{\pi(j)}p_{ji}}{\sqrt{\pi(i)}} \iff \pi(i)p_{ij} = \pi(j)p_{ji}.$$

We also have that $\langle x, y \rangle_\pi = x^T D y$.

Since $\mathscr{P}^*$ is symmetric, it has real eigenvalues, it is diagonalizable, and its right eigenvectors and the same as its left eigenvectors. Let $\{w_1, \ldots, w_r\}$ be the set of orthonormal eigenvectors with corresponding eigenvalues $\lambda_1, \ldots, \lambda_r$. Define u_i and v_i by

$$w_i = D^{-\frac{1}{2}} u_i,$$

$$w_i = D^{\frac{1}{2}} v_i.$$

We get that $\frac{1}{\sqrt{\pi(i)}} u_i = \sqrt{\pi(i)} v_i$, and therefore have that $u = Dv$. The matrices $\mathbf{P}$ and $\mathscr{P}^*$ have the same eigenvalues, and for a eigenvalue λ_i, v_i is a right (column) eigenvector and u_i^T is a left (row) eigenvector.

The column vectors v_i are orthonormal in $\mathscr{L}^2(\pi)$, because

$$\langle v_i, v_j \rangle_\pi = \sum_{k \in \mathbb{I}} v_i(k) v_j(k) \pi(k) = \sum_{k \in \mathbb{I}} w_i(k) w_j(k) = \delta_{ij},$$

$$||v_i||^2 = \sum_{k \in \mathbb{I}} v_i(k)^2 \pi(k) = \sum_{k \in \mathbb{I}} w_i(k)^2 = 1.$$

Where $\delta_{ij} = 1$ if $i = j$ and 0 if $i \neq j$.

Similarly, the row eigenvectors u_i^T's are orthonormal in $\mathscr{L}^2(\frac{1}{\pi})$. Recall that for $\mathbf{P}$, $u_1 = \pi$, $v_1 = \mathbb{1}$. Since $\{v_1, \ldots, v_r\}$ are r orthonormal vectors in $\mathbb{R}^r$, they are a basis of $\mathbb{R}^r$. This implies that for any $x \in \mathbb{R}^r$ we can write

$$x = \sum_{i \in \mathbb{I}} \alpha_i v_i,$$

for some α_i's in $\mathbb{R}$. Since $\langle x, v_j \rangle_\pi = \sum_{k \in \mathbb{I}} \alpha_k \langle v_k, v_j \rangle_\pi = \alpha_j$, we can write

$$x = \sum_{j=1}^{r} \langle x, v_j \rangle_\pi v_j.$$

Similarly, we deduce that

$$x^T = \sum_{j=1}^{r} \langle x^T, u_j^T \rangle_{\frac{1}{\pi}} u_j^T.$$

For any $j \in \mathbb{I}$, $n \in \mathbb{N}$, we have that $\mathbf{P}^n v_j = \lambda_j^n v_j$. Therefore we deduce that

$$\mathbf{P}^n x = \sum_{j=1}^{r} \lambda_j^n \langle x, v_j \rangle_\pi v_j,$$

$$x^T \mathbf{P}^n = \sum_{j=1}^{r} \lambda_j^n \langle x^T, u_j^T \rangle_{\frac{1}{\pi}} u_j^T.$$

□

Definition 4.20. We define the χ^2 *contrast of* a probability distribution α *with respect to* a probability distribution β as

$$\chi^2(\alpha; \beta) = \sum_{i \in \mathbb{I}} \frac{(\alpha(i) - \beta(i))^2}{\beta(i)}.$$

Note that $\chi^2(\alpha; \pi) = ||\alpha - \pi||^2_{\frac{1}{\pi}}$. □

Theorem 4.21. *For probability distributions α and β, we have that*

$$4d(\alpha,\beta)^2 \leq \chi^2(\alpha;\beta).$$

Proof.

$$\begin{aligned}\left(\sum_{i\in\mathbb{I}}\big|\alpha(i)-\beta(i)\big|\right)^2 &= \left(\sum_{i\in\mathbb{I}}\left|\frac{\alpha(i)}{\beta(i)}-1\right|\beta(i)^{\frac{1}{2}}\beta(i)^{\frac{1}{2}}\right)^2\\ &\leq \sum_{i\in\mathbb{I}}\left(\frac{\alpha(i)}{\beta(i)}-1\right)^2\beta(i)\\ &= \sum_{i\in\mathbb{I}}\frac{1}{\beta(i)}\big(\,\alpha(i)-\beta(i)\,\big)^2,\end{aligned}$$

where the second line is from Cauchy-Schwarz. □

Theorem 4.22. *Let $\mathbf{P}$ be an irreducible transition matrix on a finite state set $\mathbb{I}=\{1,\ldots,r\}$ and reversible on its stationary distribution π. Then for any initial probability distribution μ on $\mathbb{I}$, and for all $n\geq 1$,*

$$||\mu^T\mathbf{P}^n-\pi^T||_{\frac{1}{\pi}} \leq \rho^n||\mu^T-\pi^T||_{\frac{1}{\pi}},$$

where $\rho=\sup(\lambda_2,|\lambda_r|)$. Further, for any $i\in\mathbb{I}$, for any $n\geq 1$, and for any $A\subset\mathbb{I}$, we have that

$$|\delta_i^TP^n(A)-\pi^T(A)| \leq \left(\frac{1-\pi(i)}{\pi(i)}\right)^{\frac{1}{2}}\min\left(\pi(A)^{\frac{1}{2}},\frac{1}{2}\right)\rho^n.$$

From this, we can deduce that

$$4d_v(\delta_i^TP^n,\pi)^2 \leq \frac{1-\pi(i)}{\pi(i)}\rho^{2n} \leq \frac{\rho^{2n}}{\pi(i)}.$$

Proof. Recall that $u_1=\pi$, $v_1=\mathbb{1}$. This implies that

$$\langle\mu^T-\pi^T,u_1^T\rangle_{\frac{1}{\pi}} = \sum_{i\in\mathbb{I}}(\mu(i)-\pi(i))\pi(i)\frac{1}{\pi(i)} = \sum_{i\in\mathbb{I}}(\mu(i)-\pi(i)) = 0.$$

Let $\alpha_j=\langle\mu^T-\pi^T,u_j\rangle_{\frac{1}{\pi}}$, and recall that $x^T\mathbf{P}^n=\sum_{j=1}^r\lambda_j^n\langle x^T,u_j^T\rangle_{\frac{1}{\pi}}u_j^T$. This yields

$$\begin{aligned}||(\mu-\pi)^T\mathbf{P}^n||_{\frac{1}{\pi}}^2 &= \sum_{j=2}^r\alpha_j^2\lambda_j^{2n}||u_j^T||_{\frac{1}{\pi}}^2\\ &= \sum_{j=2}^r\alpha_j^2\lambda_j^{2n}\\ &\leq \rho^{2n}\sum_{j=2}^r\alpha_j^2 = \rho^{2n}||\mu^T-\pi^T||_{\frac{1}{\pi}}^2,\end{aligned}$$

where the second line is from the fact that u_j's are othornormal, and the third line is from the definition of ρ.

For the second part, define $\delta_i^T \mathbf{P}^n = \mu_n^T$.

$$\begin{aligned} |\mu_n^T(A) - \pi^T(A)|^2 &= \left| \sum_{i \in A} \left(\frac{\mu_n(i)}{\pi(i)} - 1 \right) \pi(i) \right|^2 \\ &\leq \left(\sum_{i \in A} \left(\frac{\mu_n(i)}{\pi(i)} - 1 \right)^2 \pi(i) \right) \pi(A) \\ &\leq \left(\sum_{i \in \mathbb{I}} \left(\frac{\mu_n(i)}{\pi(i)} - 1 \right)^2 \pi(i) \right) \pi(A) \\ &= ||\delta_i^T \mathbf{P}^n - \pi^T||_{\frac{1}{\pi}}^2 \pi(A) \leq \rho^{2n} ||\delta_i^T - \pi^T||_{\frac{1}{\pi}}^2 \pi(A), \end{aligned}$$

where the second line is by Cauchy-Schwarz, and the last line is from the first part. It is easy to check that $||\delta_i^T - \pi^T||_{\frac{1}{\pi}}^2 = \frac{1-\pi(i)}{\pi(i)}$, which implies that

$$|\delta_i^T \mathbf{P}^n(A) - \pi^T(A)| \leq \left(\frac{1 - \pi(i)}{\pi(i)} \right)^{\frac{1}{2}} \pi(A)^{\frac{1}{2}} \rho^n.$$

We also recall that

$$\left| \mu_n(A) - \pi(A) \right|^2 \leq d_v(\mu_n, \pi)^2 \leq \frac{1}{4} \chi^2(\mu_n; \pi),$$

$$\chi^2(\mu_n; \pi) = ||\mu_n^T - \pi^T||_{\frac{1}{\pi}}^2 \leq \rho^{2n} ||\delta_i^T - \pi^T||_{\frac{1}{\pi}}^2 = \rho^{2n} \frac{1 - \pi(i)}{\pi(i)}.$$

Thus we get that

$$|\delta_i^T P^n(A) - \pi^T(A)| \leq \left(\frac{1 - \pi(i)}{\pi(i)} \right)^{\frac{1}{2}} \frac{1}{2} \rho^n.$$

Thus we deduce that

$$|\delta_i^T P^n(A) - \pi^T(A)| \leq \left(\frac{1 - \pi(i)}{\pi(i)} \right)^{\frac{1}{2}} \min \left(\pi(A)^{\frac{1}{2}}, \frac{1}{2} \right) \rho^n.$$

By Lemma (4.2), we get the third part of the theorem. □

Fact 4.23. Let $A : V \to V$ be a linear map on an arbitrary vector space V of finite dimension r with an arbitrary inner product $\langle \cdot, \cdot \rangle$. Suppose that this map is self adjoint, i.e. for any $x \in V$

$$\langle Ax, y \rangle = \langle x, Ay \rangle.$$

Then there is an orthonormal basis $v_1, \ldots, v_r$ of eigenvectors of A, corresponding to eigenvalues $\lambda_1, \ldots, \lambda_r$, all of which are real. Suppose also that these eigenvalues are ordered in such a way that

$$\lambda_1 \leq \cdots \leq \lambda_r$$

Then we have that

$$\inf_{||x||=1} \langle Ax, x \rangle = \lambda_1, \quad \inf_{\substack{x \perp v_1, \\ \|x\|=1}} \langle Ax, x \rangle = \lambda_2,$$

etc. □

Definition 4.24. If $\mathbf{P}$ is a reversible transition matrix with invariant probability distribution π, we define the *Dirichlet form*

$$\mathcal{E} : \mathscr{L}^2(\pi) \times \mathscr{L}^2(\pi) \to \mathbb{R},$$
$$\mathcal{E}_\pi(x, x) = \left\langle (I - \mathbf{P})x, x \right\rangle_\pi.$$

□

Theorem 4.25.

$$\mathcal{E}_\pi = \frac{1}{2} \sum_{i,j \in \mathbb{I}} \pi(i) p_{ij} (x(j) - x(i))^2$$

Proof.

$$\begin{aligned}
\langle (I - \mathbf{P})x, x \rangle_\pi &= \sum_{i,j \in \mathbb{I}} \pi(i) p_{ij} x(i)(x(i) - x(j)) \\
&= \sum_{i,j \in \mathbb{I}} \pi(j) p_{ji} x(j)(x(j) - x(i)) \\
&= \sum_{i,j \in \mathbb{I}} \pi(i) p_{ij} x(j)(x(j) - x(i)) \\
&= \frac{1}{2} \sum_{i,j \in \mathbb{I}} \pi(i) p_{ij} (x(j) - x(i))^2,
\end{aligned}$$

where the second line is a change of index, the third line is from reversiblity, and the fourth line is the sum of the first and third lines, halved to keep equality. □

Remark 4.26. Note that for any $c \in \mathbb{R}$, we have that

$$\begin{aligned}
\mathcal{E}_\pi(x - c \cdot \mathbb{1}, x - c \cdot \mathbb{1}) &= \frac{1}{2} \sum_{i,j \in \mathbb{I}} \pi(i) p_{ij} ((x(i) - c) - (x(j) - c))^2 \\
&= \mathcal{E}_\pi(x, x).
\end{aligned}$$

□

Remark 4.27. If $\mathbf{P}$ is irreducible, we have that $\lambda_1 = 1$ and that $v_1 = \mathbb{1}$ is a right eigenvector of $\mathbf{P}$. Further we can order the eigenvalues

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_r \geq -1.$$

Here, λ_2 is the second largest eigenvalue of $\mathbf{P}$ (SLE). Recall $\rho = \sup(\lambda_2, |\lambda_r|)$ was the second largest eigenvalue modulus (SLEM).
Consider the matrix $I - \mathbf{P}$. This has eigenvalues

$$\beta_i = 1 - \lambda_i,$$

for each $i \in \mathbb{I}$. This gives the ordering

$$0 = \beta_1 < \beta_2 \leq \cdots \leq \beta_r \leq 2.$$

And the right eigenvectors for $I - \mathbf{P}$ are the same as $\mathbf{P}$, the v_i's. Note that

$$\frac{\mathcal{E}_\pi(x, x)}{\mathrm{Var}_\pi(x)} = \frac{\langle (I - \mathbf{P})x, x \rangle_\pi}{\langle x, x \rangle_\pi}.$$

Then by Fact (4.23), we get the following theorem. □

Theorem 4.28. *Let* $\mathbf{P}$ *be an irreducible transition matrix on a finite state space* $\mathbb{I} = \{1, \dots r\}$ *with stationary distribution* π*. If* $\mathbf{P}$ *is reversible with* π*, then for all* $j \geq 2$ *we have that*

$$\beta_j = \inf \left\{ \frac{\mathcal{E}_\pi(x,x)}{Var_\pi(x)} : \langle x, v_i \rangle_\pi = 0, i \in [i, \dots j-1] \right\}.$$

Any vector x *in this infimum is an eigenvector of* $\mathbf{P}$ *corresponding to the eigenvalue* $\lambda_j = 1 - \beta_j$*.* □

Remark 4.29. Since $v_1 = \mathbb{1}$, we have that

$$\beta_2 = \inf \left\{ \frac{\mathcal{E}_\pi(x,x)}{Var_\pi(x)} : \langle x \rangle_\pi = 0, x \neq 0 \right\} = \inf \left\{ \frac{\mathcal{E}_\pi(x,x)}{Var_\pi(x)}, x \neq \textit{constant} \right\}.$$

□

Corollary 4.30. *If there exists an* $A > 0$ *such that for any* $x \in \mathbb{R}^r$

$$Var_\pi(x) \leq A\mathcal{E}_\pi(x,x),$$

then

$$\lambda_2 \leq 1 - \frac{1}{A},$$

where λ_2 *is the SLE of* $\mathbf{P}$*.*

Proof. If $\frac{1}{A} \leq \frac{\mathcal{E}_\pi(x,x)}{Var_\pi(x)}$ for all $x \in \mathbb{R}^r$, then

$$\frac{1}{A} \leq \beta_2,$$

which implies that

$$\lambda_2 \leq 1 - \frac{1}{A}.$$

□

4.5. **Summary of Consequences of Ergodic Markov Chains.** We summarize our results in the following theorem.

Theorem 4.31. *For an ergodic HMC* $\{X_n\}_{n \geq 0}$ *with transition matrix* $\mathbf{P}$*, we have that there exists a stationary distribution* $\pi > 0$*. For any bounded function* $f : \mathbb{I} \to \mathbb{R}$ *we have the following:*

- *The temporal averages of* f*,*

$$\frac{1}{n} \sum_{k=1}^{n-1} f(X_k),$$

  converge almost surely to the spatial average of f

$$\bar{f} = \sum_{i \in \mathbb{I}} f(i)\pi(i).$$

- *For any initial state* $x_0 \in \mathbb{I}$*,* $\lim_{n \to \infty} \mathbb{P}(X_n = j | X_0 = x_0) = \pi(j)$ *for any* $j \in \mathbb{I}$*.*

Note that in the Ergodic Theorem (3.67) aperiodicity is not needed, but here we include is as a result of ergodic chains as well. □

Remark 4.32. If our HMC is ergodic, it behaves well. It is aperiodic, meaning that it does not travel between sets of states. It is irreducible, meaning that it can never get stuck. And it is positive recurrent, meaning that it will return to states in a finite amount of time. Given these properties, our HMC has a property similar to the SLLN for i.i.d. random variables, that is, it converges to a probabilistic average. Also, no matter where our HMC starts, if we let it run long enough it will end with a probability distribution according to π. □

5. The Shuffling Problem

We have established the concept of convergence to a stationary distribution. One may wonder how long this convergence takes. One interesting example is the application of Markov chains with card shuffling.

How many shuffles does it take to randomize a deck? In this example, our state space is the symmetric group of 52 elements $\mathscr{X} = S_{52}$, $|S_{52}| = 52!$. We will define *riffle shuffling* as a probability density $\mathbf{Q}$ on S_{52} A riffle shuffle will be defined as cutting the deck into 2 packets of non-negative size, and dropping cards from each of the packets into a new pile with probability proportional to packet size. For $g \in S_{52}$ we have that $\mathbf{Q}(g) \geq 0$ and $\sum_g \mathbf{Q}(g) = 1$. We define a Markov chain by setting X_0 as the identity permutation. From there,

$$\mathbb{P}(X_1 = g) = \mathbf{Q}(g),$$

$$\mathbb{P}(X_2 = g) = \mathbf{Q} * \mathbf{Q}(g) = \sum_{h \in \mathscr{X}} \mathbf{Q}(h)\mathbf{Q}(gh^{-1}),$$

$$\vdots$$

$$\mathbb{P}(X_k = g) = \mathbf{Q}^{k^*}(g) = \mathbf{Q} * \mathbf{Q}^{(k-1)^*}(g) = \sum_h \mathbf{Q}(h)\mathbf{Q}^{(k-1)^*}(gh^{-1}),$$

where $\mathbf{Q}^{(k-1)^*}(g)$ is the $(k-1)$ repeated convolution.

The Markov chain $\{X_n\}_{n \geq 0}$ is irreducible, since one can get to any permutation by a series of riffle-shuffles, and aperiodic, because it can get back to a state in one step by the identity, which is a riffle shuffle with one packet of size 0 and the other of size 52. Therefore (since the state space is finite) the chain is ergodic. It turns out that its stationary distribution is the uniform distribution, since repeated convolutions go towards the uniform distribution as the number of convolutions goes to infinity, and that

$$\mathbf{Q}^{k^*}(g) \to U(g) = \frac{1}{52!},$$

as $k \to \infty$. However, the question remains as to how many shuffles it takes to get reasonably close to the uniform distribution. We let $d_Q(k) = |\mathbf{Q}^{k^*} - U|$.

Computing $d_Q(k)$ is possible, but is beyond the scope of this paper. Its formula involves rising sequences. A rising sequence of a permutation is a maximal consecutively increasing subsequence. $\left\langle {n \atop r} \right\rangle$ is defined as the number of permutations of n elements with r rising sequences, and there are recursive formulas to easily calculate these. With all of this in mind, it can be deduced that

$$d_Q(k) = |\mathbf{Q}^{k^*} - U| = \frac{1}{2} \sum_{r=1}^{52} \left\langle {52 \atop r} \right\rangle \left| \frac{\binom{2^k + 52 - r}{52}}{2^{52k}} - \frac{1}{52!} \right|.$$

For 9 shuffles, we have a graph of the distances in Figure 1. It shows that after 7 shuffles, we have reasonable enough distance to assume that all possible orderings of the deck are close to being equally likely. Further information can be found in [1] and [9].

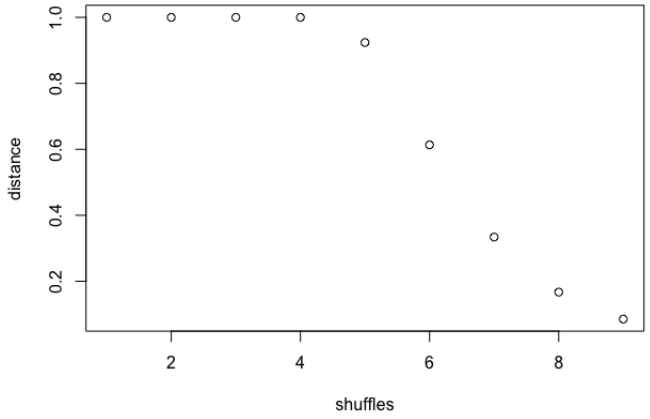


FIGURE 1. Shuffle Distances

6. Metropolis-Hastings Algorithm

Suppose we want to know information about a probability distribution π on a finite state space $\mathscr{X}$. We will use a process called Markov Chain Monte Carlo, or MCMC. Our goal will be to find a Markov chain with a stationary distribution equal to the probability distribution π. Then, if we run this chain for a long time (and if it is ergodic), we will arrive at π. If we run the Markov chain for N steps, and do this M times, we can find the proportion of samples that end in a state $i \in \mathscr{X}$, which we can call $\mu_N(i)$, and do this for each possible $i \in \mathscr{X}$. We can then measure how far our approximation is by calculating the distance in variation, that is

$$|\mu_N - \pi| = \frac{1}{2} \sum_{i \in \mathscr{X}} |\mu_N(i) - \pi(i)|.$$

We will not get 0 unless we could make $N, M = \infty$, however we can get pretty close. It remains to discuss how to construct a HMC with stationary distribution π.

6.1. Hard Disks in a Box: Motivating Example.

Our motivating example will be the problem for which the Metropolis algorithm was originally created.

Suppose we have N disks of radius $\delta > 0$ contained in the unit square where $\delta \ll 1$ for large N. Disks are contained in the unit square with a periodic boundary, meaning that if a disk goes outside the square to the right it comes in on the left. In other words, we think of hard disks on a torus. The disks in the configuration are not allowed to overlap.

Further, the centers of the disk can only be at the centers of a grid generated by horizontal and vertical lines. These lines are determined by their intersections with the axes. If we pick m points on the horizontal axis and n points on the vertical axis, we have mn possible centers for our disks.

Let $\mathbb{H}_x$ be the set of m points in the interval $[0, 1]$, and $\mathbb{H}_y$ be the set of n points in the interval $[0, 1]$. Therefore the set of all possible centers is $\mathbb{H} = \mathbb{H}_x \times \mathbb{H}_y$. We denote by $\mathscr{X}(N, \delta, \mathbb{H})$ the set of all possible arrangements of the N disks of radius δ with centers at points in the array $\mathbb{H}$. Thus $\mathscr{X}(N, \delta, \mathbb{H})$ is a subset of the family $\binom{\mathbb{H}}{N}$ of subsets of cardinality N of $\mathbb{H}$.

Suppose we had a function $f : \mathscr{X}(N, \delta, \mathbb{H}) \to \mathbb{R}$ and want to approximate the average

$$\frac{1}{|\mathscr{X}(N, \delta, \mathbb{H})|} \sum_{x \in \mathscr{X}(N,\delta,\mathbb{H})} f(x). \tag{6.1}$$

It is easy to see why this is difficult: we do not have a good idea what the set $\mathscr{X}(N, \delta, \mathbb{H})$ is. In particular, finding its cardinality could be a daunting task.

Observe that the map

$$\pi : \mathscr{X}(N, \delta, \mathbb{H}) \to [0, 1], \quad \pi(x) = \frac{1}{|\mathscr{X}(N, \delta, \mathbb{H})|}.$$

is the uniform probability distribution on $\mathscr{X}(N, \delta, \mathbb{H})$.

We will evaluate the average (6.1) by constructing an irreducible ergodic Markov chain $(X_k)_{k\geq 0}$ with state space $\mathscr{X}(N, \delta, \mathbb{H})$ and stationary measure π. The Ergodic Theorem implies that the temporal average

$$\frac{1}{k} \sum_{i=1}^{k} f(X_i),$$

converges almost surele to the spatial average (6.1) as $k \to \infty$. The construction of the Markov chain $(X_k)_{k\geq 0}$ proceeds as follows.

Start with a possible configuration $X_0 = x_0 \in \mathscr{X}(N, \delta, \mathbb{H})$. Than for $k \geq 0$, we proceed by:

- We are at a configuration $X_k = x \in \mathscr{X}(N, \delta, \mathbb{H})$.
- Pick uniformly at random a disk from the configuration of N disks x.
- With the selected disk, create a ball of radius $h > 0$ from the center of selected disk, picking h uniformly from the interval $(0, 1)$, and pick a random point uniformly from the intersection of the ball and the grid $\mathbb{H}$.
- Move the center of the disk to this new point, and denote this new configuration by y. If $y \in \mathscr{X}(N, \delta, \mathbb{H})$, set $X_{k+1} = y$. Otherwise, set $X_{k+1} = X_k = x$.

This forms a chain $X_0, X_1, \ldots X_k, \ldots$ of configurations of disks in the unit square. This can be thought of as a Markov chain if we were to define a transition matrix based upon the possibilities of moving between arrangements. This chain is irreducible because any configuration in $\mathscr{X}(N, \delta, \mathbb{H})$ can be achieved by a finite set of moves described above. The chain is aperiodic because you can return to a state in one step (when the proposed configuration is rejected). This chain is positive recurrent since it is irreducible over a finite number of configurations (due to $\mathbb{H}$ being finite).

Therefore the chain is ergodic (for certain values of δ and N) and we can compute averages that approximate our desired information. I will not delve into why this specific algorithm works or what kind of functions we may wish to approximate. I use this example to show that this algorithm can be used to study very complicated objects, such as $\mathscr{X}(N, \delta, \mathbb{H})$. I also use this example to illustrate the concepts of acceptance and rejection that are used in the Metropolis-Hastings Algorithm. Further reading on Hard Disks in a Box can be found in [5].

6.2. **Proposal and Acceptance.** In general, we have a probability distribution $\pi(x) > 0$, for all $x \in \mathscr{X}$, and a Markov chain $\{K_n\}_{n\geq 0}$ with transition matrix $\mathbf{Q}$ on a finite state space $\mathscr{X}$. We wish to change $\{K_n\}_{n\geq 0}$ so that it has a stationary distribution π, and we proceed in a way similar to how the hard disks in a box worked.

Given that we are at a state $i \in \mathscr{X}$, we propose a state $j \in \mathscr{X}$ according to $\{K_n\}_{n\geq 0}$ (specifically, with probability q_{ij}). Then we accept this proposed state j with a probability α_{ij}.

Clearly, $0 \leq \alpha_{ij} \leq 1$. This creates a transition probability for our new process given by

$$q_{ij}\alpha_{ij},$$

for $i \neq j$ (for $i = j$, we have $1 - \sum_{i \neq j} q_{ij}\alpha_{ij}$). In order to ensure that π will become the stationary distribution of our new process, we require reversibility with π, that is,

$$\pi(i)q_{ij}\alpha_{ij} = \pi(j)q_{ji}\alpha_{ji}.$$

Setting $r(i,j) = \frac{\pi(j)q_{ji}}{\pi(i)q_{ij}}$, we get that

$$\alpha_{ji} = \frac{1}{r(i,j)}\alpha_{ij}.$$

Since $0 \leq \alpha_{ij}, \alpha_{ji} \leq 1$, we require that $0 \leq \alpha_{ij} \leq 1$ and $0 \leq \alpha_{ij} \leq r(i,j)$. Thus we have that

$$0 \leq \alpha_{ij} \leq \min\{1, r(i,j)\}.$$

In order to maximize our acceptance rates, we will take the maximum value for α_{ij}. Therefore

$$\alpha_{ij} = \min\{1, r(i,j)\}.$$

6.3. **Defining the Algorithm.** Given an arbitrary transition matrix $\mathbf{Q}$ and a probability distribution $\pi > 0$ over a state space $\mathscr{X}$, given an initial state $x_0 \in \mathscr{X}$, and given that our process is at $X_n = i \in \mathscr{X}$, we pick X_{n+1} by:

- Choose $Y_{n+1} = j$ according to the transition matrix Q, probability q_{ij} (*proposal*).
- Set $\alpha_{ij} = \min\{1, \frac{\pi(j)q_{ji}}{\pi(i)q_{ij}}\}$ (*acceptance probability*).
- Set $X_{n+1} = Y_{n+1} = j$ with probability α_{ij} (*acceptance*). Otherwise, set $X_{n+1} = X_n = i$ (*rejection*).

We have defined a Markov chain $\{X_n\}_{n \geq 0}$ with probabilities

- $p_{ij} = q_{ij}\alpha_{ij}$ for $i \neq j$,
- $p_{ii} = 1 - \sum_{j \in \mathscr{X}} q_{ij}\alpha_{ij}$.

Proposition 6.1. *This HMC has a stationary distribution equal to π.*

Proof. It suffices to show that we have reversibility with π, that is

$$\pi(i)p_{ij} = \pi(j)p_{ji}.$$

For $i = j$, this is trivial. For $i \neq j$, we have

$$\begin{aligned} \pi(i)p_{ij} &= \pi(i)q_{ij}\alpha_{ij} \\ &= \pi(i)q_{ij}\min\{1, \frac{\pi(j)q_{ji}}{\pi(i)q_{ij}}\} \\ &= \pi(i)\min\{q_{ij}, \frac{\pi(j)q_{ji}}{\pi(i)}\} \\ &= \min\{\pi(i)q_{ij}, \pi(j)q_{ji}\}. \end{aligned}$$

Similarly, $\pi(j)p_{ji} = \min\{\pi(j)q_{ji}, \pi(i)q_{ij}\}$. Hence we have reversibility. □

Thus we have formed a new Markov chain with stationary distribution π. If we can show that the chain formed from the algorithm is ergodic, then we obtain the desired convergence to the stationary distribution π. Since $\mathscr{X}$ is finite, we just need to check that the chain is irreducible and aperiodic.

6.4. **Cryptography.** The following example is influenced by Diaconis. The R-code can be found here.[2] Suppose we have a piece of text that has been encrypted with a substitution cipher. For example, our original text is

> THE PROBABILITY THAT WE MAY FAIL IN THE STRUGGLE OUGHT NOT TO DETER US FROM THE SUPPORT OF A CAUSE WE BELIEVE TO BE JUST

It has been given a substitution cipher, giving a coded text

> OVB CTEAJADKDOM OVJO SB RJM HJDK DN OVB WOTYXXKB EYXVO NEO OE ZBOBT YW HTER OVB WYCCETO EH J QJYWB SB ABKDBLB OE AB UYWO

Our problem is to find the bijective function $f : \{A, B, \ldots, Z\} \to \{A, B, \ldots, Z\}$ that decodes the quote.

Let $\mathscr{X}$ be the set of all bijective functions described above. This is a set consisting of 26! permutations. Finding the correct permutation several orders of magnitude harder than finding the needle in a hay stack.

To understand how large 26! is note that $\log_{10}(26!) \approx 26.6$, so $26! \approx 4 \cdot 10^{26}$. One cubic meter contains roughly $4 \cdot 10^9$ grains of sand so there are roughly $4 \cdot 10^{18}$ grains sand per km^3. The surface area of the United States is roughly 10^7 km^2. Imagine that we cover the entire surface of the United States with a 10 km (6 miles) tall layer of sand. The volume of such a layer would be roughly 10^8 km^3 and it would contain roughly 26! grains of sand, give or take a few. Our task is to find the magic grain in this immensity!

We assign a weight $\mathcal{L}(f)$ to each permutation f as follows. Take a long English text. E.g., in my example I used *Moby Dick*. Use this text and create a matrix M of relative frequencies of letter transitions in the text in an effort to approximate the true frequencies of letter transitions for the English language. We can visualize this matrix with Figure 2. This shows us patterns of what letters tend to follow each other in the English language. From this, we can create a function that measures the likelihood of a cipher function $f \in \mathscr{X}$, given by

$$\mathcal{L}(f) = \prod_i M(f(s_i), f(s_{i+1})).$$

Here, s_i, s_{i+1} are consecutive letters in our ciphered text. The idea behind this weight is simple. If a permutation f is not the one we seek, then the string of successive letters

$$\big(f(s_1), f(s_2)\big), \quad \big(f(s_2), f(s_3)\big), \quad \big(f(s_3), f(s_4)\big), \ldots$$

is unlikely to occur in an English text so we expect the product of the frequencies $M\big(f(s_i), f(s_{i+1})\big)$ to be small so the weight $L(f)$ is small. Thus, we expect the weight of the true code f_{true} to be a lot higher. In our specific example, the original text is 2.6×10^{115} times more likely than the coded text!!!

Using the above weight, we create a probability distribution given by

$$\pi(f) = \frac{1}{Z} \prod_i M(f(s_i), f(s_{i+1})), \tag{6.2}$$

[2] http://www.r-bloggers.com/text-decryption-using-mcmc/

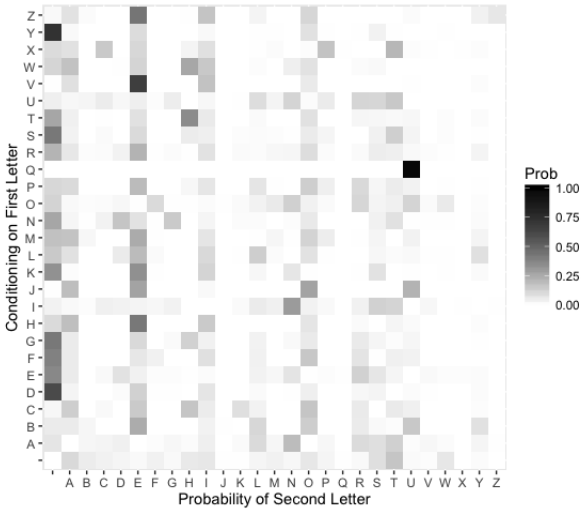


FIGURE 2. Moby Dick Transition Matrix

where

$$Z = \sum_{f \in \mathscr{X}} \prod_i M(f(s_i), f(s_{i+1})).$$

We wish to find out more information about $\pi(f)$, and which f is likely to be our cipher function. As we have pointed out, the set $\mathscr{X}$ is huge and, for all intents and purposes the normalizing constant Z is unknownable.

To overcome these obstacles we employ the Metropolis-Hastings algorithm. We define a proposal distribution $\mathbf{Q} = \{q_{ff'}\}_{f,f' \in \mathscr{X}}$ where

$$q_{ff'} = \begin{cases} \frac{1}{\binom{26}{2}} & \text{if } f, f' \text{ differ in at most two places,} \\ 0 & \text{otherwise.} \end{cases}$$

The above formula is justified by the fact that there are $\binom{26}{2}$ transpositions of the English alphabet, and the fact that f' differs from f at at most two places signifies that either $f' = f$, or f' is the the product between f and a transposition.

Note that $q_{ff'} = q_{f'f}$. We start with a guess $f_0 \in \mathscr{X}$. Then given $X_n = f$, we pick X_{n+1} using the Metropolis-Hastings proposal-acceptance/rejection protocol.

- Choose $Y_{n+1} = f'$ by $\mathbf{Q}$. (This is the proposal step.)
- Set
$$\alpha_{ff'} := \min\left\{1, \frac{\pi(f')q_{f'f}}{\pi(f)q_{ff'}}\right\} = \min\left\{1, \frac{\pi(f')}{\pi(f)}\right\} = \min\left\{1, \frac{\mathcal{L}(f')}{\mathcal{L}(f)}\right\}.$$
- Set $X_{n+1} = Y_{n+1} = f'$ with probability $\alpha_{ff'}$. (This is the acceptance part.)
- Otherwise, set $X_{n+1} = X_n = f$. (This is the rejection part.)

The chain is irreducible because any permutation $f \in \mathscr{X}$ is a product of transpositions. The chain is aperiodic because you can return to a state in one step, when the proposal is rejected. Therefore the chain is ergodic. Its stationary distribution is our mysterious distribution (6.2).

The state X_n of this Markov chain is a random point of $\mathscr{X}$. For n sufficiently large, the distribution of X_n is very close to the mysterious π. By construction, $\pi(f)$ will be very high for any f that is close to the actual code to something that resembles real English, and low for

something that does not. We can see that in this example. After $3,000$ accepted functions in our Metropolis algorithm, we get close to our original text:

> THE PROLALINITY THAT WE MAY FAIN ID THE STRUGGNE OUGHT
> DOT TO KETER US FROM THE SUPPORT OF A JAUSE WE LENIEVE
> TO LE BUST

Accuracy can be due to the number of iterations ran, the length of text we are decoding, as well as the text we use for our matrix. The dependence on these factors is explored in [17]. In fact, this text is actually more likely than our real original text. Even if the process does not converge to the original text, it will converge to something that the human brain can still make sense of, because it will be something that closely resembles English. Longer texts converge more accurately and quickly. For instance, when I decode the Gettysburg Address, it is decoded with complete accuracy with only 91 accepted functions.

6.5. **Hyperlinks.** One of the Metropolis-Hastings Algorithm's largest strengths is its ability to determine global information locally. For instance, say we wanted to estimate the average amount of hyperlinks among web pages on the internet. Let this true average be denoted by μ, and suppose there are M web pages on the internet. Let $\mathscr{X}$ be the set of all web pages on the internet ($|\mathscr{X}| = M$), and let $h : \mathscr{X} \to \mathbb{R}$ be a function that calculates the number of hyperlinks on a website $i \in \mathscr{X}$. h is easy to calculate on a given web page, and if we were able to visit all of them we would obtain

$$\mu = \frac{1}{M} \sum_{i \in \mathscr{X}} h(i).$$

However, this would be nearly impossible due to the size of M. In fact, it is nearly impossible to even know M. If we could construct an ergodic Markov chain that allowed us to approximate μ, we would be able to accomplish this nearly impossible task. We know by the ergodic theorem that such a Markov chain would obey the property

$$\frac{1}{n} \sum_{k=1}^{n-1} h(X_k) \to \sum_{i \in \mathscr{X}} \pi(i) h(i).$$

If we can construct a Markov chain with stationary distribution $\pi(i) = \frac{1}{M}$ for all $i \in \mathscr{X}$, then we get that

$$\frac{1}{n} \sum_{k=1}^{n-1} h(X_k) \to \sum_{i \in \mathscr{X}} \pi(i) h(i) = \frac{1}{M} \sum_{i \in \mathscr{X}} h(i) = \mu.$$

So we need to construct an ergodic Markov chain with stationary distribution $\pi(i) = \frac{1}{M}$ for all $i \in \mathscr{X}$. A strength of the Metropolis algorithm is that we can form such a chain without needing to know M. Consider the following algorithm:

Start with an arbitrary web page $X_0 = x_0 \in \mathscr{X}$. Then given $X_n = i$, pick X_{n+1} by the following algorithm:

- Pick $Y_{n+1} = j$ uniformly from all the hyperlinks on website i (with probability $\frac{1}{deg(i)}$) .
- Set $\alpha_{ij} = \min\left\{1, \frac{h(i)}{h(j)}\right\}$.
- Set $X_{n+1} = Y_{n+1} = j$ with probability α_{ij}. Otherwise set $X_{n+1} = X_n = i$.

We assume that the chain is irreducible by assuming that we can get to any website by a series of hyperlink clicks. It is aperiodic because we can return to a state in one step, when the proposal is rejected. Therefore the chain is ergodic.

Hence we have formed our desired Markov chain, and can approximate μ by calculating

$$\frac{1}{n}\sum_{k=1}^{n-1} h(X_k)$$

for large n. We may not know M, but we can approximate μ just by knowing $h(i)$ for a given website in which we visit.

6.6. **Rates of Convergence for Metropolis-Hastings.** We have see that the Metropolis-Hastings Algorithm can be applied in a variety of areas, and it seems for certain examples such as the coding problem to converge rather quickly. However, the true rate of convergence is not known for many of these problems. In this section, we will show that we can approximate rate of convergence for a certain type of stationary distribution and candidate generating matrix.

Let $\mathbb{I} = \{1, \dots, r\}$ be a finite state space, and $\pi(i) = z(a)a^{h(i)}$ be a probability distribution where $a \in (0,1)$, $z(a)$ a constant, and h a function such that for $i \in [1, r-1]$, $h(i+1) - h(1) \geq c \geq 1$. Let $\mathbf{Q} = \{q_{ij}\}_{i,j\in\mathbb{I}}$ be the symmetric random walk on $\mathbb{I}$ with holding probability at states 1 and r equal to $\frac{1}{2}$. That is, for $1 < i, j < r$,

$$q_{1,1} = \frac{1}{2} = q_{r,r},$$

$$q_{ij} = q_{ji} = \frac{1}{2},$$

$$q_{ii} = 0.$$

With $\alpha_{ij} = \min\left\{1, \frac{\pi(j)q_{ji}}{\pi(i)q_{ij}}\right\}$, we get a Markov chain with transition probabilities

$$p_{1,2} = \frac{1}{2}\frac{\pi(2)}{\pi(1)} = \frac{1}{2}a^{h(2)-h(1)},$$

$$p_{1,1} = 1 - \frac{1}{2}a^{h(2)-h(1)},$$

$$p_{r,r-1} = p_{r,r} = \frac{1}{2},$$

and for $i \in [2, r-1]$,

$$p_{i,i-1} = \frac{1}{2},$$

$$p_{i,i+1} = \frac{1}{2}a^{h(i+1)-h(i)},$$

$$p_{ii} = 1 - p_{i,i-1} - p_{i,i+1}.$$

We will show that $\lambda_2 \leq 1 - \frac{(1-a^{\frac{c}{2}})^2}{2}$. By Corollary (4.30), we must show that $\mathrm{Var}_\pi(x) \leq A\mathcal{E}_\pi(x,x)$ for $A \leq \frac{2}{(1-a^{\frac{c}{2}})^2}$. Denote an *oriented edge* from $i \to j$ in $\mathbf{P}$ as e where $e^- = i$, $e^+ = j$. Define $Q(e) = \pi(i)p_{ij}$. For any two $i, j \in \mathbb{I}$, we can select a path $i, i_1, i_2, \dots, i_m, j \in \mathbb{I}$ of distinct edges such that $p_{i,i_1} \cdots p_{i_m,j} > 0$. Let Γ be the collection of all such paths and for a path $\gamma_{ij} \in \Gamma$ we define

$$|\gamma_{ij}|_\theta = \sum_{e\in\gamma_{ij}} \frac{1}{Q(e)^{2\theta}}.$$

With this notation, we deduce that

$$\begin{aligned}
2\operatorname{Var}_\pi(x) &= \sum_{i,j\in\mathbb{I}}\left[\sum_{e\in\gamma_{ij}}\frac{1}{Q(e)^\theta}Q(e)^\theta(x(e^-)-x(e^+))\right]^2\pi(i)\pi(j)\\
&\leq \sum_{i,j\in\mathbb{I}}\left[\sum_{e\in\gamma_{ij}}Q(e)^{2\theta}(x(e^-)-x(e^+))^2\right]\left[\sum_{e\in\gamma_{ij}}\frac{1}{Q(e)^{2\theta}}\right]\pi(i)\pi(j)\\
&= \sum_{i,j\in\mathbb{I}}|\gamma_{ij}|_\theta\sum_{e\in\gamma_{ij}}Q(e)^{2\theta}(x(e^-)-x(e^+))^2\pi(i)\pi(j)\\
&= \sum_e(x(e^-)-x(e^+))^2Q(e)Q(e)^{2\theta-1}\sum_{e\in\gamma_{ij}}\pi(i)\pi(j)|\gamma_{ij}|_\theta\\
&\leq A\mathcal{E}_\pi(x,x),
\end{aligned}$$

where $A=\max_e\left\{Q(e)^{2\theta-1}\sum_{e\in\gamma_{ij}}\pi(i)\pi(j)|\gamma_{ij}|_\theta\right\}$, and the second line is from Cauchy-Shwarz. So it remains to bound A. Take a path $\gamma_{ij}=(i,i+1,\ldots,j-1,j)$ for any $i,j\in\mathbb{I}$ such that $i\leq j$. Note that

$$\begin{aligned}
Q(i,i+1) &= \pi(i)p_{i,i+1}=z(a)a^{h(i)}\frac{a^{h(i+1)-h(i)}}{2}\\
&= z(a)\frac{a^{h(i+1)}}{2}=\frac{\pi(i+1)}{2}.
\end{aligned}$$

And by reversibility, $Q(i,i+1)=Q(i+1,i)=\frac{\pi(i+1)}{2}$. Now we have that

$$\begin{aligned}
|\gamma_{ij}|_\theta &= \sum_{e\in\gamma_{ij}}\frac{1}{Q(e)^{2\theta}}\\
&= \left(\frac{\pi(i+1)}{2}\right)^{-2\theta}+\cdots+\left(\frac{\pi(j)}{2}\right)^{-2\theta}\\
&= \left(\left(\frac{\pi(i+1)}{\pi(j)}\right)^{-2\theta}+\cdots+\left(\frac{\pi(j)}{\pi(j)}\right)^{-2\theta}\right)\left(\frac{\pi(j)}{2}\right)^{-2\theta}\\
&\leq \frac{\pi(j)^{-2\theta}}{1-a^{2c\theta}}.
\end{aligned}$$

Fix an edge $e=(k,k+1)$. We must bound over i,j the quantity

$$\frac{Q(e)^{2\theta-1}}{1-a^{2c\theta}}\sum_{0\leq i\leq k,k+1\leq j\leq r}\pi(i)\pi(j)^{1-2\theta}.$$

The sum in i is bounded by 1, and the sum in j is bounded by $\frac{\pi(k+1)^{1-2\theta}}{1-a^{c(1-2\theta}}$. Thus $A\leq\frac{2}{(1-a^{c(1-2\theta)})(1-a^{2c\theta})}$. For $\theta=\frac{1}{4}$, this gives us

$$A\leq\frac{2}{(1-a^{\frac{c}{2}})^2}.$$

Lemma 6.2 (Gershgorin Bound)**.** *If B is a finite $r \times r$ matrix with complex elements, then for any eigenvalue λ of B, and any $k \in [1, r]$, we have that*

$$|\lambda - a_{kk}| \leq \min(r_k, s_k),$$

where $r_k = \sum_{j=1, j\neq k}^{r} |a_{kj}|$, $s_k = \sum_{j=1, j\neq k}^{r} |a_{jk}|$. □

By Lemma (6.2), we get that the smallest eigenvalue λ_s is

$$\lambda_s \geq -1 + 2\min\{p_{ii}\} \geq -1 + 2\left(\frac{1}{2} - \frac{a^c}{2}\right) = a^c.$$

This implies that

$$\rho = \min\{\lambda_2, |\lambda_s|\} \leq \min\left\{1 - \frac{(1 - a^{\frac{c}{2}})^2}{2}, a^c\right\}.$$

7. Conclusion

Metropolis-Hastings allows us to estimate a value that requires a global knowledge of a structure (such as the size of the state space) by only using local knowledge. It tends to converge quickly and accurately .It is an extremely important algorithm involving Markov chains, which are very powerful ways to model stochastic processes. They emerged due to a religious debate between Markov and Nekrosov, but evolved to have extremely vast applications.

8. Acknowledgments

Special thanks to Liviu Nicolaescu for being not only my thesis advisor but a mentor throughout my career at the University of Notre Dame. He has been the most influential teacher I have had during my time here. Thank you to my parents for the opportunity to go to such a wonderful university. Thank you to Paul Bosley for sparking my interest in mathematics during high school. Thank you to Jeff Diller for creating the opportunity for me to discover the joy behind mathematical research.

References

[1] D. Alous and P. Diaconis: *Shuffling Cards and Stopping Times*, American Mathematical Monthly, vol. 93, no. 5, 1986.
[2] P Billingsley: *Probability and Measure*, 3rd Edition, John Wiley, 1995.
[3] P. Bremaud: *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Corrected Edition, Springer, 2008.
[4] D. Clement: *Convergence Rates of Markov Chain Monte Carlo Methods*, http://probability.ca/jeff/ftpdir/clement.pdf
[5] P. Diaconis: *The Markov Monte Carlo Revolution*, American Mathematical Society, vol. 46, no. 2, 2009.
[6] R. Dobrow: *Probability with Applications and R*, John Wiley & Sons, 2014.
[7] R. Durret: *Probability. Theory and Examples*, 4th Edition, Cambridge University Press,2010
[8] W. Feller: *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd Edition, John Wiley & Sons, 1968.
[9] H. Hammarstrom: *Card-Shuffling Analysis with Markov Chains*, 2005, http://www.math.chalmers.se/~olleh/Markov_Hammarstrom.pdf
[10] B. Hayes: *First Links in the Markov Chain*, American Scientist, http://www.americanscientist.org/issues/pub/first-links-in-the-markov-chain/1
[11] D.A. Levin, Y. Peres, E.L. Wilmer: *Markov Chains and Mixing Times*. Amer. Math. Soc., 2009. http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf
[12] J.R. Norris: *Markov Chains*, Cambridge University Press, 1997.

[13] *Text Decryption Using MCMC*, R-Bloggers, http://www.r-bloggers.com/text-decryption-using-mcmc/
[14] Lothar Breuer, *Introduction to Stochastic Processes*, https://www.kent.ac.uk/smsas/personal/lb209/files/sp07.pdf
[15] *Markov Chains and Coupling*, https://www.cs.duke.edu/courses/fall15/compsci590.4/slides/lec5.pdf
[16] *Little-o Notation*, https://xlinux.nist.gov/dads/HTML/littleOnotation.html
[17] *Decrypting Classical Cipher Text Using Markov Chain Monte Carlo*, http://probability.ca/jeff/ftpdir/decipherart.pdf

Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556-4618.
E-mail address: Michael.M.McCaffrey.13@nd.edu