MARKOV CHAINS: STATIONARY DISTRIBUTIONS AND FUNCTIONS ON STATE SPACES

JAMES READY

Abstract. In this paper, we first introduce the concepts of Markov Chains and their stationary distributions. We then discuss total variation distance and mixing times to bound the rate of convergence of a Markov Chain to its stationary distribution. Functions on state spaces are then considered, including a discussion of which properties of Markov Chains are preserved over functions, and we will show that the mixing time of a Markov chain is greater than or equal to the mixing time of its image.

Contents

1. Introduction

Definition 1.1. A Markov Chain with countable state space Ω is a sequence of Ω - valued random variables, $X_1, X_2, X_3, \ldots$, such that for any states x_i, and any time $n \geq 1$,

$$P\{X_n = x_n || X_{n-1} = x_{n-1}, X_{n-2} = x_{i-2}, \ldots X_0 = x_0\} = P\{X_n = x_n || X_{n-1} = x_{n-1}\}$$

This definition says that the state of a Markov Chain depends only on the state immediately preceding it, and is independent of any behavior of the chain before that. We will often denote the probability of going from one state to another, $P\{X_n = x_n | X_{n-1} = x_{n-1}\}$ as $p(x_{n-1}, x_n)$. To refer to the probability of going from one state to another in j steps, $P\{X_{n+j} = y | X_n = x\}$, we will use the notation of $p^j(x, y)$.

We will also often refer to the transition probability matrix P of a Markov Chain. This means the $|\Omega| \times |\Omega|$ matrix, where for all i, j in Ω, $P_{ij} = p(i, j)$.

Definition 1.2. A Markov Chain X_n with transition probability matrix P is irreducible if for any two states x, y there exists an integer t (possibly depending on x and y) such that $p^t(x, y) > 0$

By this definition, if a Markov chain X_n is irreducible, there is positive probability that, starting at any state, the Markov chain will reach any other state in finite time.

Definition 1.3. Let $T(x) = \{t \geq 1 : p^t(x, x) > 0\}$ be the set of times t for which there is a positive probability for the chain to return to starting position x at time t. The period of state x is defined to be the greatest common divisor of $T(x)$.

Date: August 31, 2010.

Definition 1.4. A Markov Chain X_n is said to be aperiodic if the period of all of its states is 1.

We will often discuss Markov Chains that are both irreducible and aperiodic. In this case, the following theorem is useful:

Theorem 1.5. *If X_n is an aperiodic and irreducible Markov Chain with transition probability matrix P on some finite state space Ω, then there exists an integer M such that for all $m > M$, $p^m(x, y) > 0$ for all $x, y \in \Omega$.*

Proof. For any $x \in \Omega$, recall from the definition of periodicity that $T(x) = \{t \geq 1 \mid P^t(x, x) > 0\}$, and since X_n is aperiodic, we know $\gcd(T(x)) = 1$.
Consider any s and t in $T(x)$. Since

$$p^{s+t}(x, x) \geq p^s(x, x)p^t(x, x) > 0$$

we know that $s + t$ is in $T(x)$ as well, so $T(x)$ is closed under addition.
We will now use the fact from number theory that any set of non-negative integers which is closed under addition and which has greatest common divisor 1 must contain all but finitely many of the non-negative integers. This implies that we can find some t_x such that, for all $t \geq t_x$, $t \in T(x)$.
Fix any x, y in Ω. Because X_n is irreducible, there exists some $k_{(x,y)}$ such that $p^{k_{(x,y)}}(x, y) > 0$.
Thus, for $t \geq t_x + k_{(x,y)}$,

$$p^t(x, y) \geq p^{t-k_{(x,y)}}(x, x)p^{k_{(x,y)}}(x, y) > 0$$

Now, for any $x \in \Omega$, let $t'_x = t_x + \max_{y\in\Omega} k_{(x,y)}$. For all $y \in \Omega$, we have $p^{t'_x}(x, y) > 0$. We then know that for all $m \geq \max_{x\in\Omega} t'_x$, we have $p^m(x, y) > 0$ for any $x, y \in \Omega$. □

2. Stationary Distributions

Definition 2.1. For any Markov Chain X_n with transition probability matrix P, a stationary distribution of X_n is any probability distribution π satisfying the condition that

$$\pi = \pi P$$

By matrix multiplication, it is equivalent that, for any $y \in \Omega$,

$$\pi(y) = \sum_{x\in\Omega} \pi(x)p(x, y)$$

I will show in this section that, for any Markov chain X_n on a finite state space Ω, the stationary distribution π exists, and if X_n is irreducible, then π is unique. In the next section, I will show that if X_n is an irreducible and aperiodic Markov chain, there is a notion of convergence to its stationary distribution.
For most of the theorems in this section, the condition of a finite state space is necessary. To show this, consider the simple random walk on $\mathbb{Z}$, which is a Markov Chain represented by the transition probabilities:

$$p(x, y) = \begin{cases} \frac{1}{2} & : \text{if } |x - y| = 1 \\ 0 & : \text{otherwise} \end{cases}$$

I will show that the simple random walk cannot have any stationary distribution. Suppose the simple random walk on has a stationary distribution, that is, there exists some probability distribution π such that

$$\pi(y) = \sum_{x\in\Omega} \pi(x)p(x,y)$$
$$\Rightarrow \pi(y) = \frac{1}{2}\pi(y-1) + \frac{1}{2}\pi(y+1)$$

This means that $\pi : \mathbb{Z} \to [0,1]$ is a harmonic function; that is, it satisfies the condition that, for all x, $\pi(x)$ equals the average of its neighbors. It can be shown that the only harmonic functions on $\mathbb{Z}$ are linear. Because π must always be non-negative, we know that the slope of π must be zero. If $\pi(x) = 0$ for all x, then π is not a probability distribution; similarly, if $\pi(x) \neq 0$, then $\sum_{x\in\mathbb{Z}} \pi(x) = \infty$, so π is not a probability distribution. Thus, the simple random walk cannot have any stationary distribution.

Theorem 2.2. *If X_n is a Markov Chain with transition probability matrix P on finite state space Ω, then π, the stationary distribution of X_n, exists.*

Proof. Let $S = \{\text{probability distributions on} \{1, 2, \ldots, n\}\} \subset \mathbb{R}^n$. Because the sum of all components of a probability distribution must equal one, S is a compact subset of $\mathbb{R}^n$.
Choose any $\mu \in S$. We know that $\mu P, \mu P^2, \mu P^3, \ldots$ are all also in S.
Now, consider the sequence

$$\nu_n = \frac{1}{n}\sum_{k=0}^{n-1} \mu P^k$$

ν_n is also a sequence in S. Because S is compact, it must have some convergent subsequence. Thus, there exists some subsequence ν_{n_j} that converges to some π in S.
I claim that this π is a stationary distribution for P. We want to show that $\pi P = \pi$, or equivalently, $\pi P - \pi = 0$.

$$\begin{aligned}
\pi P &= \lim_{j\to\infty} v_{n_j} P \\
&= \lim_{j\to\infty} \frac{1}{n_j} \sum_{k=0}^{n_j-1} \mu P^{k+1} \\
\Rightarrow \pi P - \pi &= \lim_{j\to\infty} \frac{1}{n_j} \sum_{k=0}^{n_j-1} \mu P^{k+1} - \frac{1}{n_j} \sum_{k=0}^{n_j-1} \mu P^k \\
&= \lim_{j\to\infty} \frac{1}{n_j} \sum_{k=1}^{n_j} \mu P^k - \frac{1}{n_j} \sum_{k=0}^{n_j-1} \mu P^k \\
&= \lim_{j\to\infty} \frac{1}{n_j} (\mu(P^{n_j}) - \mu) \\
&= 0
\end{aligned}$$

since $(\mu(P^{n_j}) - \mu)$ is bounded and $\lim_{j\to\infty} \frac{1}{n_j}$ goes to 0. □

Theorem 2.3. *If X_n is an irreducible Markov Chain with transition probability matrix P on finite state space Ω, then the stationary distribution π is unique.*

Proof. Suppose there exists π_1 and π_2 such that $\pi_1 P = \pi_1$ and $\pi_2 P = \pi_2$. It is sufficient to show that $\frac{\pi_1(x)}{\pi_2(x)}$ is a constant over all x, because if all stationary vectors are scalar multiples of each other, there can be at most one with components that sum to 1.
Because X_n is irreducible, we know that, if $\pi P = \pi$, then $\pi(x) > 0$ for all $x \in \Omega$. This is true because Because $\sum_{x\in\Omega} \pi(x) = 1$, there must exist some $y \in \Omega$ such that $\pi(y) > 0$, so if we choose any x, there must be some m such that $p^m(y, x) > 0$ (or else X_n would be reducible), so it immediately follows that π can have no 0 components.
Since Ω is finite, it follows that there exist some $\epsilon > 0$ such that

$$\min_x \frac{\pi_1(x)}{\pi_2(x)} = \epsilon$$

This implies that, for all $y \in \Omega$,

$$\pi_1(y) - \epsilon\pi_2(y) \geq 0$$
$$\text{and } = 0 \text{ for some } y \in \Omega$$

If $\pi_1 - \epsilon\pi_2$ is not the zero vector, then we get that, for some appropriate scalar α, $\alpha(\pi_1 - \epsilon\pi_2)$ is also a stationary distribution. However, $\alpha(\pi_1 - \epsilon\pi_2) = 0$ at some x, contradicting what we showed above, that no stationary distribution of an irreducible Markov chain can have any components equal to 0. Thus, $\pi_1 - \epsilon\pi_2$ must equal 0 for all x, so $\frac{\pi_1(x)}{\pi_2(x)}$ is a constant, and P can only have one stationary distribution. □

3. Convergence

The goal of this section is to place an upper bound on the time it takes any irreducible, aperiodic Markov chains converge to its stationary distribution. This section is mostly based on Chapter 4 of *Markov Chains and Mixing Times.*

3.1. **Total Variation Distance.**

In order to define Mixing Times, we first need to define a method to measure distance between probability distributions.

Definition 3.1. For some probability distributions μ and ν, the *Total Variation Distance* between μ and ν, denoted $||\mu - \nu||_{TV}$, is

$$||\mu - \nu||_{TV} = \max_{A\subset\Omega} |\mu(A) - \nu(A)|$$

Now, we will introduce some other definitions of Total Variation distance and show that they are equivalent.

Proposition 3.2.

$$||\mu - \nu||_{TV} = \frac{1}{2} \sum_{x\in\Omega} |\mu(x) - \nu(x)|$$

Proof. Let $B = \{x \mid \mu(x) \geq \nu(x)\}$. For any event $A \in \Omega$,

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

Similarly,

$$\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c)$$

Because

$$\mu(B) + \mu(B^c) = 1 = \nu(B) + \nu(B^c)$$

we get that $\nu(B^c) - \mu(B^c) = \mu(B) - \nu(B)$, so this is an upper bound of $||\mu - \nu||_{TV}$, and if we consider the event $A = B$, we get equality.
Thus, we get,

$$\begin{aligned} ||\mu - \nu||_{TV} &= \frac{1}{2}[\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)] \\ &= \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)| \end{aligned}$$

□

Note that the above proof also shows that

$$||\mu - \nu||_{TV} = \sum_{x\in\Omega, \mu(x)>\nu(x)} |\mu(x) - \nu(x)|$$

To understand the next definition of Total Variation Distance, we need to use coupling.

Definition 3.3. A *coupling* of two probability distributions μ and v is a pair of random variables (X, Y) such that the marginal distribution of X is μ and the marginal distribution of Y is v.

Thus, if q is a joint distribution of X, Y on $\Omega \times \Omega$, meaning $q(x, y) = P\{X = x, Y = y\}$, then $\sum_{y\in\Omega} q(x, y) = \mu(x)$ and $\sum_{x\in\Omega} q(x, y) = \nu(y)$.
Clearly, X and Y can always have the same value only if μ and ν are identical.

Proposition 3.4.

$$||\mu - \nu||_{TV} = \inf\{P\{X \neq Y\} \mid (X, Y) \textit{ is a coupling of } \mu \textit{ and } \nu\}$$

Note: Such a coupling is called an *optimal* coupling. In the proof below, we will show an optimal coupling always exists.

Proof. For any event $A \subset \Omega$,

$$\begin{aligned} \mu(A) - \nu(A) &= P(X \in A) - P(Y \in A) \\ &\leq P\{X \in A, Y \notin A\} \\ &\leq P\{X \neq Y\} \end{aligned}$$

$$\Rightarrow ||\mu - \nu||_{TV} \leq \inf\{P\{X \neq Y \ (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}$$

Now, we need to construct a coupling such that equality holds.
Let

$$p = \sum_{x\in\Omega} \min\{\mu(x), \nu(x)\}$$

We show that

$$\begin{aligned}
p &= \sum_{x\in\Omega} \min\{\mu(x),\nu(x)\} \\
&= \sum_{x\in\Omega,\mu(x)<\nu(x)} \mu(x) + \sum_{x\in\Omega,\nu(x)\leq\mu(x)} \nu(x) \\
&= \sum_{x\in\Omega,\mu(x)<\nu(x)} \mu(x) + \sum_{x\in\Omega,\nu(x)\leq\mu(x)} \mu(x) + \sum_{x\in\Omega,\nu(x)\leq\mu(x)} \nu(x) - \sum_{x\in\Omega,\nu(x)\leq\mu(x)} \mu(x) \\
&= 1 - \sum_{x\in\Omega,\nu(x)\leq\mu(x)} (\mu(x)-\nu(x)) \\
&= 1 - ||\mu(x)-\nu(x)||_{TV}
\end{aligned}$$

Now, flip a coin with probability p of heads.
If the coin is heads, choose the value of X by the probability distribution

$$\gamma(x) = \frac{\min\{\mu(x),\nu(x)\}}{p}$$

In this case, let $Y = X$.
If the coin is tails, choose the value of X using the probability distribution

$$\eta(x) = \begin{cases} \frac{\mu(x)-\nu(x)}{1-p} & : \text{if } \mu(x) > \nu(x) \\ 0 & : \text{otherwise} \end{cases}$$

Independently choose the value of Y using the probability distribution

$$\kappa(x) = \begin{cases} \frac{\nu(x)-\mu(x)}{1-p} & : \text{if } \nu(x) > \mu(x) \\ 0 & : \text{otherwise} \end{cases}$$

Using this, we have, for the distribution of X,

$$p\gamma + (1-p)\eta = \mu$$

and for the distribution of Y, we have

$$p\gamma + (1-p)\kappa = \nu$$

This means that we have defined a coupling (X,Y) of μ and ν. Note that $X \neq Y$ if and only if the tail lands tails. Thus,

$$P\{X \neq Y\} = 1-p \quad = ||\mu-\nu||_{TV}$$

□

3.2. **The Convergence Theorem.**

Theorem 3.5. *Suppose P is an irreducible, aperiodic Markov Chain, with stationary distribution π. Then there exists $\alpha \in (0,1)$ and $C > 0$ such that, for all $t \geq 0$,*

$$\max_{x\in\Omega} ||P^t(x,\cdot) - \pi||_{TV} \leq C\alpha^t$$

Proof. Since P is irreducible and aperiodic, there exists $\delta > 0$ and $r \in \mathbb{N}$ such that, for all x, y in Ω,

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

Let $\Theta = 1-\delta$. Let Π be a square matrix with $|\Omega|$ rows, each of which is π.
$P^r = (1-\Theta)\Pi + \Theta Q$ defines a stochastic matrix Q.

Now, we will show $P^{rk} = (1 - \Theta^k)\Pi + \Theta^k Q^k$ by induction.
For $k = 1$, this is true, because it is exactly how we defined Q. Assume this holds for $k = n$.

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

Because $\Pi P^r = \Pi$, and $Q^n \Pi = \Pi$, we get

$$P^{r(n+1)} = (1 - \Theta^{n+1})\Pi + \Theta^{n+1} Q^{n+1}$$

completing the induction.

$$\begin{aligned} \Rightarrow P^{rk}P^j &= (1 - \Theta^k)\Pi P^j + \Theta^k Q^k P^j \\ P^{rk+j} &= \Pi - \Theta^k \Pi + \Theta^k Q^k P^j \\ P^{rk+j} - \Pi &= \Theta^k (Q^k P^j - \Pi) \\ \frac{P^{rk+j}(x_0, \cdot) - \Pi(x_0, \cdot)}{2} &= \frac{\Theta^k Q^k P^j (x_0, \cdot) - \Pi(x_0, \cdot)}{2} \end{aligned}$$

Thus, the left hand side above becomes the Total Variation Distance, and the second term on the right hand side is at most 1, so we get that there exists some C_j such that

$$\max_{x \in \Omega} ||P^{rk+j}(x_0, \cdot) - \pi||_{TV} \leq C_j \Theta^t$$

Now, let $C = max\{C_0, C_1, \ldots, C_{r-1}, C_r\}$, let $k = [\frac{t}{r}]$, and we have

$$\max_{x \in \Omega} ||P^{rk+j}(x_0, \cdot) - \pi||_{TV} \leq C\Theta^t$$

□

3.3. **Standardizing Distance from Stationary.** We will now introduce notation to show the distance of a Markov Chain from its stationary distribution, and the distance between two Markov Chains.

Definition 3.6.

$$\begin{aligned} d(t) &= \max_{x \in \Omega} ||P^t(x, \cdot) - \pi||_{TV} \\ \overline{d}(t) &= \max_{x,y \in \Omega} ||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \end{aligned}$$

Lemma 3.7.

$$d(t) \leq \overline{d}(t) \leq 2d(t)$$

Proof. Because the triangle inequality holds for Total Variation Distance, for all x, y in Ω,

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \leq ||P^t(x, \cdot) - \pi||_{TV} + ||P^t(y, \cdot) - \pi||_{TV}$$

so $\overline{d}(t) \leq 2d(t)$. Since π is stationary,

$$\pi(A) = \sum_{y \in \Omega} \pi(y) P^t(y, A)$$

Thus, we have

$$\begin{aligned}
d(t) &= \max_{x\in\Omega} ||P^t(x,\cdot) - \pi||_{TV} \\
&= \max_{x\in\Omega} \big(\max_{A\subset\Omega} |P^t(x,A) - \pi(A)| \big) \\
&= \max_{x\in\Omega} \big(\max_{A\subset\Omega} | \sum_{y\in\Omega} \pi(y)|P^t(x,A) - P^t(y,A)| \big) \\
&\leq \max_{x\in\Omega} \big(\sum_{y\in\Omega} \pi(y) \max_{A\subset\Omega} |P^t(x,A) - P^t(y,A)| \big) \\
&\leq \max_{x\in\Omega} \big(\max_{y\in\Omega} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \big) \\
&\leq \max_{x,y\in\Omega} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} = \overline{d}(t)
\end{aligned}$$

The second to last inequality holds since $\sum_{y\in\Omega} \pi(y)$ is a convex linear combination. □

Lemma 3.8. *$\overline{d}$ is submuliplicative; that is,*

$$\overline{d}(s+t) \leq \overline{d}(s)\overline{d}(t)$$

Proof. Fix x, y in Ω, and let (X_s, Y_s) be the optimal coupling of $P^s(x,\cdot)$ and $P^s(y,\cdot)$. We early showed this coupling does exist, and that

$$||P^s(x,\cdot) - P^s(y,\cdot)||_{TV} = P\{X_s \neq Y_S\}$$

$$\begin{aligned}
p^{s+t}(x,w) &= \sum_{z\in\Omega} p^s(x,z)p^t(z,w) \\
&= \sum_{z\in\Omega} P^x\{X_s = z\}p^t(z,w) \\
&= E^x(P^t(X_s,w))
\end{aligned}$$

Similarly, $p^{s+t}(y,w) = E(p^t(Y_s,w))$.

$$\begin{aligned}
\Rightarrow p^{s+t}(x,w) - p^{s+t}(y,w) &= E(P^t(X_s,w) - P^t(Y_s,w)) \\
\Rightarrow ||P^{s+t}(x,w) - P^{s+t}(y,w)||_{TV} &= \frac{1}{2}\sum_{w\in\Omega} |E(P^t(X_s,w) - P^t(Y_s,w))| \\
&\leq E(\frac{1}{2}\sum_{w\in\Omega} |P^t(X_s,w) - P^t(Y_s,w)|) \\
&= E(||P^t(X_s,\cdot) - P^t(Y_s,\cdot)||_{TV}) \\
&\leq \overline{d}(t)P\{X_s \neq Y_s\} \\
&\leq \overline{d}(t)\overline{d}(s)
\end{aligned}$$

□

We now get to the following corollary:

Corollary 3.9. *For all $c \in \mathbb{N}$,*

$$d(ct) \leq \overline{d}(ct) \leq \overline{d}(t)^c$$

Proof. From the earlier lemma, $d(ct) \leq \overline{d}(ct)$. From the lemma immediately above,

$$\begin{aligned}\overline{d}(ct) &= \overline{d}(t + t + \ldots + t) \text{ (c times)} \\ &\leq \overline{d}(t)\overline{d}(t)\ldots\overline{d}(t) \text{ (c times)} \\ &= \overline{d}(t)^c\end{aligned}$$

□

3.4. **Mixing Times.**

Definition 3.10. The *mixing time* of a Markov Chain, with respect to ϵ, is

$$t_{mix}(\epsilon) = \min\{t \mid d(t) \leq \epsilon\}$$

We will also let $t_{mix} = t_{mix}(\frac{1}{4})$.

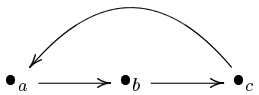
The choice of exactly $\frac{1}{4}$ is somewhat arbitrary, but as we will see below, it simplifies the calculation of a key inequality.
By corollary 3.9, for any $l \in \mathbb{N}$

$$\begin{aligned}d(lt_{mix}(\epsilon)) \leq \overline{d}(lt_{mix}(\epsilon)) \leq \overline{d}(t_{mix}(\epsilon))^l \leq (2\epsilon)^l \\ \Rightarrow d(lt_{mix}) \leq 2^{-l}\end{aligned}$$

This means that a Markov Chain converges to its stationary distribution exponentially fast.

4. Functions

We will now consider the idea of functions on state spaces. We let X_n be a Markov Chain with transition probability matrix P on some state space $\mathcal{X}$, and consider some function $F : \mathcal{X} \to \mathcal{W}$. F will always be surjective, but not necessarily injective. We will let $W_n = F(X_n)$. For all of this section, assume $\mathcal{X}$ is at most countable.
I will show, by counterexample, that W_n is not always a Markov Chain.
Let $\mathcal{X} = \{a, b, c\}$ and $\mathcal{W} = \{d, e\}$. Let X_n be the Markov Chain represented by the below directed graph, where $p(i, j) = 1$ if i has an outgoing edge to j, and 0 otherwise.



Define F such that $F(a) = F(b) = d$, and $F(c) = e$.

Let X_n have the starting distribution $\{p_1, p_2, p_3\}$ over $\{a, b, c\}$. We want to show the Markov condition does not hold for W_n; that is, there exists some n and w_n such that

$$P\{W_n = w_n | W_{n-1} = w_{n-1}\} \neq P\{W_n = w_n | W_{n-1} = w_{n-1}, W_{n-2} = w_{n-2}, \cdots, W_0 = w_0\}$$

This is true for $W_2 = d$, because

$$\begin{aligned}P\{W_2 = d | W_0 = e, W_1 = e\} &= \frac{P\{X_2 = a, X_0 = (b \text{ or } c), X_1 = (b \text{ or } c)\}}{P\{X_0 = (b \text{ or } c), X_1 = (b \text{ or } c)\}} \\ &= \frac{p_2}{p_2} \\ &= 1\end{aligned}$$

but

$$P\{W_2 = d|W_1 = e\} = \frac{P\{X_2 = a, X_1 = (b \text{ or } c)\}}{P\{X_1 = (b \text{ or } c)\}}$$
$$= \frac{p_2}{p_1 + p_2}$$

$P\{W_2 = d|W_0 = e, W_1 = e\} \neq P\{W_2 = d|W_1 = e\}$, so W_n is not a Markov Chain. However, there does exist a condition on X_n that guarantees W_n to be a Markov Chain.

Theorem 4.1. *Let X_n be a Markov Chain with transition probability matrix P on some state space $\mathcal{X}$, and consider some function $F : \mathcal{X} \to \mathcal{W}$. Let $W_n = F(X_n)$. For each $w \in \mathcal{W}$, let $F^{-1}(w) = \{x \in \mathcal{X}|F(x) = w\}$. If for every pair of states w_1, w_2 in $\mathcal{W}$ and every pair of states x_1, x_2 in $F^{-1}(w_1)$,*

$$\sum_{y \in F^{-1}(w_2)} p(x_1, y) = \sum_{y \in F^{-1}(w_2)} p(x_2, y)$$

then W_n is a Markov Chain on $\mathcal{W}$.

I will say that any Markov Chain with a function which satisfies the condition of the above theorem is *function regular*.

Proof. We want to show that

$$P\{W_n = w_n|W_{n-1} = w_{n-1}\} = P\{W_n = w_n|W_{n-1} = w_{n-1}, W_{n-2} = w_{n-2}, \cdots, W_0 = w_0\}$$

We know for every x_1, x_2 in $F^{-1}(w_{n-1})$, there exists some $\omega \in \mathbb{R}$ such that

$$\sum_{y \in F^{-1}(w_n)} p(x_1, y) = \sum_{y \in F^{-1}(w_n)} p(x_2, y) = \omega$$

I claim that

$$P\{W_n = w_n|W_{n-1} = w_{n-1}\} = P\{W_n = w_n|W_{n-1} = w_{n-1}, W_{n-2} = w_{n-2}, \cdots, W_0 = w_0\} = \omega$$

First, I will show this is true for the left hand side.

$$\begin{aligned}
P\{W_n = w_n|W_{n-1} = w_{n-1}\} &= P\{X_n \in F^{-1}(w_n)|X_{n-1} \in F^{-1}(w_{n-1})\} \\
&= \frac{P\{X_n \in F^{-1}(w_n), X_{n-1} \in F^{-1}(w_{n-1})\}}{P\{X_{n-1} \in F^{-1}(w_{n-1})\}} \\
&= \frac{\sum_{x \in F^{-1}(w_{n-1})} (P(X_{n-1} = x) \cdot \sum_{y \in F^{-1}(w_n)} p(x, y))}{P\{X_{n-1} \in F^{-1}(w_{n-1})\}} \\
&= \frac{\sum_{x \in F^{-1}(w_{n-1})} (P(X_{n-1} = x) \cdot \omega)}{P\{X_{n-1} \in F^{-1}(w_{n-1})\}} \\
&= \frac{\omega \cdot \sum_{x \in F^{-1}(w_{n-1})} (P(X_{n-1} = x))}{P\{X_{n-1} \in F^{-1}(w_{n-1})\}} \\
&= \omega
\end{aligned}$$

Now, I will show this is true for the right hand side.

$$P\{W_n = w_n | W_{n-1} = w_{n-1}, W_{n-2} = w_{n-2}, \cdots, W_0 = w_0\}$$
$$= P\{X_n \in F^{-1}(w_n) | X_{n-1} \in F^{-1}(w_{n-1}), X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\}$$
$$= \frac{P\{X_n \in F^{-1}(w_n), X_{n-1} \in F^{-1}(w_{n-1}), X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\}}{P\{X_{n-1} \in F^{-1}(w_{n-1}), X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\}}$$

$$= \left(\sum_{x \in F^{-1}(w_{n-1})} P\{X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\} \cdot P\{X_{n-1} = x | X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\} \cdot \left(\sum_{y \in F^{-1}(w_n)} P(x,y) \right) \right) / \left(P\{X_{n-1} \in F^{-1}(w_{n-1}), X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\} \right)$$

$$= \left(\omega \sum_{x \in F^{-1}(w_{n-1})} P\{X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\} \cdot P\{X_{n-1} = x | X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\} \right) / \left(P\{X_{n-1} \in F^{-1}(w_{n-1}), X_{n-2} \in F^{-1}(w_{n-2}), \cdots, X_0 \in F^{-1}(w_0)\} \right)$$

$$= \omega$$

Thus, W_n is a Markov Chain. □

However, function regularity is only a sufficient condition for the image to be a Markov Chain, not a necessary one. Let X_n be the Markov Chain represented by the transition probability matrix below.

	a	b	c	d
a	0	.5	.5	0
b	1	0	0	0
c	0	0	0	1
d	0	.5	.5	0

Let $F(a) = \alpha, F(b) = F(c) = \beta$, and $F(d) = \gamma$.
Because $p(b,a) \neq p(c,a)$, X_n is not function regular. We then have that W_n is represented by the following transition probability matrix:

	α	β	γ
α	0	1	0
β	.5	0	.5
γ	0	1	0

This counterexample demonstrates an issue with deciding how we determine the starting distribution of the X_n. Given any starting distribution of W_n, can we choose how we want to pull this back onto the X_n? If we know or can fix any pull back of starting probabilities in $\mathcal{X}$, then W_n is a Markov Chain, so this is

a counterexample. However, if we do not know how the starting distribution is pulled back in $\mathcal{X}$, and all we know is the observed probabilities of the W_n, then $P\{W_1 = \alpha | W_0 = \beta\}$ is not well defined, and W_n is not a Markov Chain.
For this rest of this section, unless otherwise stated we will not assume X_n is function regular. However, we will assume that W_n is a Markov Chain.

We will now discuss some properties of Markov Chains that are, or are not, preserved over functions on state spaces.

First, we will prove a lemma needed for the next theorem.

Lemma 4.2. *Let P be a transition probability matrix on a state space $\mathcal{Y}$. Then a probability distribution μ on $\mathcal{Y}$ is stationary for P if, on some probability space, there exist $\mathcal{Y}$-valued random variables Y_0, Y_1 such that (a) each of Y_0 and Y_1 has marginal distribution μ; and (b) the conditional distribution of Y_1 given $Y_0 = y$ is the y^{th} row of P.*

Proof. Suppose we have some random variables Y_0, Y_1 such that $P\{Y_1 = x | Y_0 = y\} = P(y, x)$ and $P\{Y_0 = y\} = P\{Y_1 = y\} = \pi(y)$. We want to show that π is stationary; that is, we want to show $\pi(y_i) = \sum_{y \in \mathcal{Y}} (\pi(y) P(y, y_i))$.

$$\begin{aligned}
\sum_{y \in \mathcal{Y}} (\pi(y) P(y, y_i)) &= \sum_{y \in \mathcal{Y}} (P\{Y_0 = y\} \cdot P\{Y_1 = x | Y_0 = y\}) \\
&= \sum_{y \in \mathcal{Y}} (P\{Y_1 = y_i, Y_0 = y\}) \\
&= \pi(y_i)
\end{aligned}$$

□

Theorem 4.3. *Let X_n, with stationary distribution π, be a function regular Markov Chain. Let $W_n = F(X_n)$, as usual. Then the projection $\pi \circ F^{-1}$ of π to $\mathcal{W}$ is a stationary distribution for W_n.*

Proof. We will first choose two random variables Y_0, Y_1 such that:

$$P\{Y_1 = x \mid Y_0 = y\} = q(y, x) \quad (1)$$

$$\begin{aligned}
P\{Y_0 = y\} = P\{Y_1 = y\} &= \pi(F^{-1}(y)) \quad (2) \\
&= \sum_{z \in F^{-1}(y)} \pi(z)
\end{aligned}$$

By the lemma, we know that if two such random variables exist, than $\pi(F^{-1})$ is a stationary distribution.
Let $P(Y_1 = x, Y_0 = y) = f(x, y)$. Having assumed $P\{Y_1 = x \mid Y_0 = Y\} = q(y, x)$, we want to show that $f(x, y)$ satisfies the condition (2) above.

$$\begin{aligned}
&\frac{P\{Y_1 = x, Y_0 = Y\}}{P\{Y_0 = y\}} = q(y, x) \\
\Rightarrow &\frac{f(x, y)}{\sum_{z \in F^{-1}(y)} \pi(z)} = q(y, x) \\
&\Rightarrow f(x, y) = q(y, x) \sum_{z \in F^{-1}(y)} \pi(z) \\
&\Rightarrow \sum_{x \in \mathcal{W}} f(x, y) = \sum_{z \in F^{-1}(y)} \pi(z)
\end{aligned}$$

The last step follows since $\sum_{x\in\mathcal{W}} q(y,x) = 1$.
Thus, we have just shown that $P(Y_0 = y) = \sum_{z\in F^{-1}(y)} \pi(z)$, the second part of condition 2. Now, all that is left is to show that $P(Y_1 = x) = (\pi \circ F^{-1}(x))$.

$$\begin{aligned} q(y,x) &= P\{W_1 = y \mid W_0 = x\} \\ &= P\{x_1 \in F^{-1}(y) \mid X_0 \in F^{-1}(x)\} \\ &= P\{x_1 \in F^{-1}(y) \mid X_0 = z \in F^{-1}(x)\} \end{aligned}$$

We know that, since X_n is function regular, the last step above does not depend on the choice of z.
Choose any u in $F^{-1}(y)$.

$$\begin{aligned} \sum_{y\in\mathcal{W}} f(x,y) &= \sum_{y\in\mathcal{W}} \Big(\sum_{v\in F^{-1}(x)} P(u,v)\Big)\Big(\sum_{z\in F^{-1}(y)} \pi(z)\Big) \\ &= \sum_{y\in\mathcal{W}} \Big(\sum_{z\in F^{-1}(y)} \Big(\sum_{v\in F^{-1}(x)} P(u,v)\pi(z)\Big)\Big) \\ &= \sum_{y\in\mathcal{W}} \Big(\sum_{z\in F^{-1}(y)} \Big(\sum_{v\in F^{-1}(x)} P(z,v)\pi(z)\Big)\Big) \\ &= \sum_{v\in F^{-1}(x)} \Big(\sum_{z\in\mathcal{X}} p(z,v)\pi(z)\Big) \text{ (Since we were able to combine the } y \text{ sum and the } z \text{ sum)} \\ &= \sum_{v\in F^{-1}(x)} \pi(v) \text{ (Since } \pi \text{ is stationary in } \mathcal{X}) \\ &= (\pi \circ F^{-1})(x) \end{aligned}$$

We just showed that $\sum_y f(x,y) = P\{Y_1 = x\} = (\pi \circ F^{-1})(x)$, so we are done. □

Now, as an example of a Markov Chain with a function on its state space, I will introduce the Ehrenfest Urn model. The state space $\mathcal{X}$ is $\{0,1\}^N$. At each integer time t, one random index $1 \le j \le N$ is chosen, and the j^{th} coordinate from time $t-1$ is switched. Thus, for two vectors x and y, the transition probabilities are:

$$p(x,y) = \begin{cases} \frac{1}{N} & : \text{if } x \text{ and } y \text{ differ in exactly one coordinate} \\ 0 & : \text{otherwise} \end{cases}$$

This can be visualized as two urns of balls. At each time t, one ball is randomly chosen to switch urns.
I claim that the stationary distribution of the Ehrenfest Urn model, π is that all possible vectors are uniformly distributed with probability $\frac{1}{2^N}$. Because there are 2^N vectors, this is a probability distribution. This distribution is stationary because, for any $y \in \mathcal{X}$

$$\begin{aligned} \frac{1}{2^N} &= N\frac{1}{2^N}\frac{1}{N} \\ \frac{1}{2^N} &= \sum_{\{x \mid x \text{ and } y \text{ differ in exactly one coordinate }\}} \frac{1}{2^N}\frac{1}{N} \\ \pi(y) &= \sum_{x\in\Omega} \pi(x)p(x,y) \end{aligned}$$

Because the Ehrenfest Urn model is irreducible and $\mathcal{X}$ is finite, π is unique.

Now, let $\mathcal{W} = \{0, 1, 2, \ldots, N\}$, and define

$$W_n = \sum_{j=1}^{N} X_n(j)$$

We know that W_n is a Markov chain because X_n is function regular (any vector with k 1's at time $t-1$ is equally likely to have l 1's at time t). We can determine the transition probabilities of W_n by looking of the transition probabilities of X_n. If $W_n = w$, then W_{n+1} increases by 1 if a 0 is chosen to be switched in X_n, and decreases by 1 if a 1 is chosen in X_n. the transition probabilities are as follows:

$$\begin{aligned} P(w, w+1) &= \frac{N-w}{N} \\ P(w, w-1) &= \frac{w}{N} \\ P(w, \text{any other value}) &= 0 \end{aligned}$$

Using Theorem 5.3, we can also determine the stationary distribution of W_n by considering the stationary distribution, π, of X_n. Since all vectors are equally likely, it is simply a matter of counting how many vectors in $\mathcal{X}$ go to each value in $\mathcal{Y}$. The number of ways to have w 1's in an N length vector is $\binom{N}{w}$. Because there are 2^N binary vectors of length N, we can see that the stationary distribution of W_n, ϕ, is:

$$\phi(w) = \frac{\binom{N}{w}}{2^N}$$

We will now show that, if X_n converges to its stationary distribution, then W_n converges to its stationary distribution at least as fast.

Theorem 4.4. *If X_n is irreducible and aperiodic, then W_n is irreducible and aperiodic.*

Proof. Because X_n is irreducible and aperiodic, there exists some M such that for all $m > M$ and $x, y \in \mathcal{X}$, $p^m(x, y) > 0$.
We will now show that $q^m(w_1, w_2) > 0$ for any $m > M$ and $w_1, w_2 \in \mathcal{W}$, where $q^m(w_1, w_2)$ represents the transition probabilities of going from w_1 to w_2 in m steps in the chain of the W_n.
Choose any $w_1, w_2 \in \mathcal{W}$.

$$\begin{aligned} q^m(w_1, w_2) &= P\{X_n \in F^{-1}(w_2) \mid X_{n-m} \in F^{-1}(w_1)\} \\ &= \frac{P\{X_n \in F^{-1}(w_2), X_{n-m} \in F^{-1}(w_1)\}}{P\{X_{n-m} \in F^{-1}(w_1)\}} \\ &= \frac{\sum_{x \in F^1(w_1), y \in F^{-1}(w_2)} p^m(x, y)}{P\{X_{n-m} \in F^{-1}(w_1)\}} \\ &> 0 \end{aligned}$$

This condition is stronger than the condition for irreducibility, so W_n is irreducible. If we let $w_2 = w_1$, this condition says that for any $m > M$ and $w_1 \in \mathcal{W}$, $q^m(w_1, w_1) > 0$, so $T(w_1) \supset \{m \mid m > M\}$. This implies that $\gcd(T(w_1)) = 1$ for all $w_1 \in \mathcal{W}$, so W_n is aperiodic. □

Theorem 4.5. $d(t)_{W_n} \leq d(t)_{X_n}$.

Proof.

$$\begin{aligned}
d(t)_{W_n} &= \max_{w\in\mathcal{W}} || \sum_{y\in W} \Big(q^t(w,y) - \pi(y)\Big) ||_{TV} \\
&= \max_{w\in\mathcal{W}} || \sum_{z\in\mathcal{X}} \sum_{x\in F^{-1}(w)} \Big(P\{X_t = x | X_t \in F^{-1}(w)\} p^t(x,z) - \pi(x)\Big) ||_{TV} \quad \text{By Theorem 5.3} \\
&\leq \max_{w\in\mathcal{W}} \left(\max_{x\in F^{-1}(w)} || \sum_{z\in\mathcal{X}} \Big(p^t(x,z) - \pi(x)\Big) ||_{TV} \right) \\
&\leq \max_{x\in\mathcal{X}} || \sum_{z\in\mathcal{X}} \Big(p^t(x,z) - \pi(x)\Big) ||_{TV} \\
&= d(t)_{X_n}
\end{aligned}$$

The third line followed because

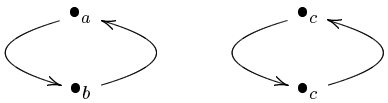
$$\sum_{x\in F^{-1}(w)} P\{X_t = x | X_t \in F^{-1}(w)\} = 1$$

□

Corollary 4.6. *For all ϵ, $t_{mix}(\epsilon)(W_n) \leq t_{mix}(\epsilon)(X_n)$*

Proof. Recall that $t_{mix}(\epsilon) = \min\{t | d(t) \leq \epsilon\}$. This corollary follows immediately from theorem 5. □

I'll now show some counterexamples of properties that are not preserved under functions on state spaces. If X_n is reducible, W_n is not necessarily irreducible. Consider an X_n of two identical, non-communicating classes, for example, the Markov Chain represented by the graph below:



It is clear that there is a function (namely, $F(a) = F(c) = \alpha$ and $F(b) = F(d) = \beta$) that maps the corresponding states in each class to the same point, creating a Markov chain with a single communicating class.
If X_n is periodic, then W_n is not necessarily period. Consider the Markov Chain represented by the transition probability matrix below to be X_n:

	a	b	c	d
a	0	.5	0	.5
b	.5	0	.5	0
c	0	.5	0	.5
d	.5	0	.5	0

Now let $F(a) = F(b) = \alpha$ and $F(c) = F(d) = \beta$. We then have that W_n is represented by the transition probability matrix below:

	α	β
α	.5	.5
β	.5	.5

Here, X_n has period 2 , but its image has period 1, so W_n is aperiodic.
Consider the following lemma:

Lemma 4.7. *Let X_n be a function regular Markov Chain, and $W_n = F(X_n)$. $q(w_1, w_2) = 0$ if and only if for all $x \in F^{-1}(w_1)$ and $y \in F^{-1}(w_2)$, we have $p(x, y) = 0$.*

Proof. First, we will show that the right hand side implies the left hand side. Suppose that for all $x \in F^{-1}(w_1)$, $y \in F^{-1}(w_2)$, $p(x, y) = 0$. We then get that

$$P\{X_n \in F^{-1}(w_2) \mid X_{n-1} \in F^{-1}(w_1)\} = 0$$

which is equivalent to saying $q(w_1, w_2) = 0$, as desired.
Now, we will show the left hand side implies the right hand side by proving the contrapositive. Suppose there exists some $x_0 \in F^{-1}(w_1)$ and $y_0 \in F^{-1}(w_2)$ such that $p(x, y) = 0$. This means that $\sum_{y \in F^{-1}(w_2)} p(x_0, y) = \alpha$, for some $\alpha > 0$. Because X_n is function regular, we know that for all $x \in \mathcal{X}$, $\sum_{y \in F^{-1}(w_2)} p(x, y) = \alpha$. Also, because we are conditioning on the event $X_{n-1} \in F^{-1}(w_1)$, we will assume that $P\{X_{n-1} \in F^{-1}(w_1)\} > 0$. We can now show that

$$\begin{aligned} q(x, y) &= P\{X_n \in F^{-1}(w_2) \mid X_{n-1} \in F^{-1}(w_1)\} \\ &= \frac{P\{X_n \in F^{-1}(w_2), X_{n-1} \in F^{-1}(w_1)\}}{P\{X_{n-1} \in F^{-1}(w_1)\}} \\ &= \frac{\sum_{x \in F^{-1}(w_1)} P\{X_{n-1} = x\} \cdot (\sum_{y \in F^{-1}(w_2)} p(x, y))}{P\{X_{n-1} \in F^{-1}(w_1)\}} \\ &= \frac{\alpha \cdot P\{X_{n-1} \in F^{-1}(w_1)\}}{P\{X_{n-1} \in F^{-1}(w_1)\}} \\ &> 0 \end{aligned}$$

□

Corollary 4.8. *If X_n is irreducible and function regular, then W_n is irreducible.*

Proof. Recall our definition of an irreducible Markov Chain: X_n is irreducible if for any two states x, y there exists an integer t (possibly depending on x and y) such that $p^t(x, y) > 0$. It follows immediately from the lemma that if X_n is irreducible, then W_n is irreducible. □

Corollary 4.9. *If X_n is aperiodic and function regular, then W_n is aperiodic.*

Proof. Recall our set $T(x) = \{t \geq 1 \mid p^t(x, x) > 0\}$ from our definition of the period of a Markov Chain. Let $\overline{T}(x) = \{t \geq 1 \mid q^t(F(x), F(x)) > 0\}$. The above lemma shows that $\overline{T}(x)$ must be always contain $T(x)$. Thus, for any w, the $\gcd(\overline{T}(x)) \leq \gcd(T(x))$. Because X_n is aperiodic, $\gcd(x) = 1$ for all $x \in \mathcal{X}$, so , because F is always onto, $\gcd(w) = 1$ for all $w \in \mathcal{W}$. Thus, W_n is aperiodic. □

Acknowledgments. It is a pleasure to thank my mentor, Yuan Shao, for helping me decide on a topic, assisting me with the difficult proofs and concepts, and looking over all of my work to ensure its correctness. I could not have done this paper without him.

References

[1] David A. Levin, Yuval Peres, and Elizabeth Wilmer. Markov Chains and Mixing Times. http://www.uoregon.edu/ dlevin/MARKOV/markovmixing.pdf.
[2] Steven P. Lalley. Markov Chains: Basic Theory. http://galton.uchicago.edu/ lalley/Courses/312/MarkovChains.pdf.

[3] Steven P. Lalley. Statistics 312: Homework Assignment 5. http://galton.uchicago.edu/ lalley/Courses/312/HW5.pdf