INTRODUCTION TO MARKOV CHAINS AND MARKOV CHAIN MIXING

ERIC SHANG

Abstract. This paper provides an introduction to Markov chains and their basic classifications and interesting properties. After establishing a solid foundation, we then go into Markov chain mixing and mixing time. We finish by briefly introducing coupling and showing how Markov chain coupling can be used to estimate mixing time.

Contents

1. Introduction

Consider a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$, known as a stochastic process. We say that a stochastic process is a Markov chain if for every state in the process, the outcome of the next state depends only on the current state and not on anything previous in the sequence.

Definition 1.1. A stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is known as a **Markov chain** if for all $n \in \mathbb{N}$:

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

where i and j are possible states that the random variable takes.

Example 1.2. One common, simple example of Markov chains are random walks. Imagine we are flipping a fair coin, where the probability of heads is $\mathbf{P}(X = H) = \frac{1}{2}$ and the probability of tails is $\mathbf{P}(X = T) = \frac{1}{2}$. For the purposes of this example, let $H = 1$ and $T = -1$.

Date: August 26, 2018.

Let us define a stochastic process $\{X_n\}_{n\in\mathbb{N}}$ such that $X_0 = 0$ and the difference $(X_{n+1} - X_n)$ has the coin-flip distribution as stated above. So at every step in the process, there is a $\frac{1}{2}$ probability that the next step will be an increment of 1 and a $\frac{1}{2}$ probability that the next step will be a decrement of 1.

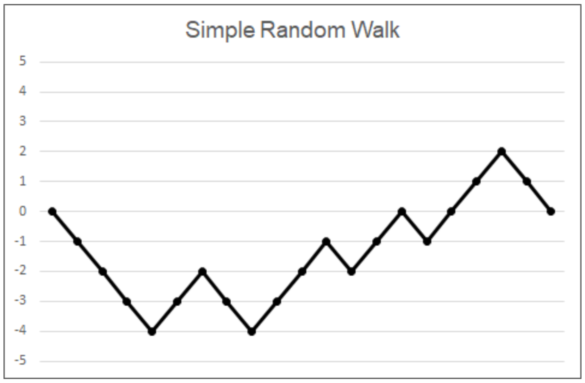


Figure 1. A simulated simple random walk of 20 steps

This figure shows a simulated random walk as defined in the example as a graph with respect to n. The y-axis can be thought of as the current state of the process.

The random walk is a simple example of a Markov chain because at each state, the probabilities of the next state is independent of all previous states. The outcome depends only on the current state because you always have the same probability of incrementing or decrementing by 1 every time.

In this paper, we will only discuss homogeneous Markov chains, meaning that the conditional probabilities of each state is independent of the current time, for all $n \in \mathbb{N}$. That being said, there are two essential pieces of information which describe a Markov chain: the initial probability distribution and the transition probabilities.

Definition 1.3. We say that a Markov chain $\{X_n\}_{n\in\mathbb{N}}$ has **initial probability distribution** λ, a vector of each states' initial probability, if:

$$\mathbf{P}(X_0 = i) = \lambda_i$$

For ease of defining transition probabilities, let

$$p_{ij} = \mathbf{P}_i(X_{n+1} = j) = \mathbf{P}(X_{n+1} = j | X_n = i).$$

Then the transition probabilities can be represented through a matrix as follows.

Definition 1.4. Given a Markov chain with sample space size of finite n, then the **transition matrix** P, of the corresponding Markov chain will be an $n \times n$ matrix represented as:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

If the sample space is infinite, then the Markov chain can similarly be represented by an infinite matrix.

Notice that with this definition, our transition matrix is **stochastic**, meaning that every row is a probability distribution. Consequently, we have that given a $n \times n$ matrix:

$$\sum_{j=1}^{n} p_{ij} = 1 \quad \forall i \in [1, n]$$

In the case of the coin-flip random walk example given above, we can see that the initial probability distribution is simply

$$\lambda_i = \begin{cases} 1 & i = 0 \\ 0 & i \neq 0 \end{cases}$$

because we always start at $X_0 = 0$.

And the transition matrix is given by the following infinite matrix:

$$P = \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \iddots \\ \cdots & 0 & \frac{1}{2} & 0 & 0 & \cdots \\ \cdots & \frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots \\ \cdots & 0 & \frac{1}{2} & 0 & \frac{1}{2} & \cdots \\ \cdots & 0 & 0 & \frac{1}{2} & 0 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

1.1. Multi-step Transitions.

Often times, it is useful to analyze how a Markov chain develops in more than one step at a time. For this, it is useful for us to define an n-th transition probability:

$$p_{ij}^{(n)} = \mathbf{P}(X_n = j | X_0 = i)$$

For our purposes, since we are studying homogeneous Markov chains and probability does not change based on time, any jump of n steps will have the same transition matrix regardless of where exactly the jump takes place in the chain.

In order to calculate the n-th transition matrix, we can simply use matrix multiplication to attain the desired result:

$$p_{ij}^{(n)} = (\mathbf{P}^n)_{ij}$$

2. Classifications and Properties

This section will explore common attributes of Markov chains and properties depending on which categories these chains fall under.

2.1. Communication Classes and Irreducibility.

Definition 2.1. Consider a Markov chain $\{X_n\}_{n \in \mathbb{N}}$ with state space S. For some $i, j \in S$, we say that i leads to j, which we write as $i \to j$, if there exists some $n \in \mathbb{N}$ such that

$$p_{ij}^{(n)} > 0$$

We say that i and j **communicate** with each other, which we write as $i \leftrightarrow j$, if both i leads to j and j leads to i.

In other words, we say that two states in a Markov chain communicate with each other if there is a positive probability to eventually reach one state from the other and vice versa in finite time. This definition of communication also allows us to find interesting groups within a Markov chain.

Proposition 2.2. *The communication between two states, $\leftrightarrow$, defines an equivalence relation over the state space S of a Markov chain $\{X_n\}_{n\in\mathbb{N}}$.*

Proof. To show an equivalence relation, three properties need to be satisfied: reflexivity, symmetry, and transitivity.
1) reflexivity: Since $p_{ii}^{(0)} = 1 > 0$ for all $i \in S$, $i \leftrightarrow i$ holds.
2) symmetry: If $i \leftrightarrow j$, then $j \leftrightarrow i$ for all $i, j \in S$ by definition of communication.
3) transitivity: for all $i, j, k \in S$, if $i \leftrightarrow j$ and $j \leftrightarrow k$, then there exists $m, n \in \mathbb{N}$ such that $p_{ij}^{(m)} > 0$ and $p_{jk}^{(n)} > 0$. Then we have:

$$\begin{aligned}
p_{ik}^{(m+n)} &= \mathbf{P}(X_{m+n} = k | X_0 = i) \\
&\geq \mathbf{P}(X_{m+n} = k \cap X_m = j | X_0 = i) \\
&= \mathbf{P}(X_{m+n} = k | X_m = j)\mathbf{P}(X_m = j | X_0 = i) \\
&= \mathbf{P}(X_n = k | X_0 = j)\mathbf{P}(X_m = j | X_0 = i) \\
&= p_{jk}^{(n)} p_{ij}^{(m)} > 0
\end{aligned}$$

Since all three conditions are satisfied, $\leftrightarrow$ is an equivalence relation. □

Since the communication relation acts as an equivalence relation on the state space of a Markov chain we can group the different states of a Markov chain into classes based on which states communicate with which other states, called communication classes.

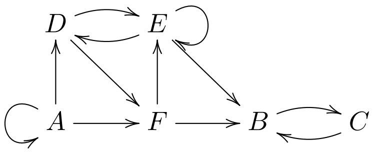
Definition 2.3. For a Markov Chain $\{X_n\}_{n\in\mathbb{N}}$ with state space S, we have a **communication class** $C \subseteq S$ if for all states i and j in C, we have that i communicates with j.

Now let us see an example of partitioning a finite state space into communication classes based on the transition matrix:

Example 2.4. Suppose we have a Markov chain with the following transition matrix:

$$P = \begin{array}{c@{}c}
\begin{array}{cccccc} A & B & C & D & E & F \end{array} & \\
\begin{bmatrix}
\frac{1}{2} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\
0 & \frac{2}{3} & 0 & 0 & \frac{1}{3} & 0
\end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \end{array}
\end{array}$$

We label the states from A to F in order to more clearly see the relation between states. Let's take state A for example. From the matrix, we can see that from A, we can reach itself, D, and F. Continuing down each row of P in this fashion, we can see which states can directly reach which other states. We can then draw the following diagram to represent the transition matrix (without any probabilities):



Based on which groups of states can all reach each other, we can categorize the states into communication classes. Taking A as an example again, notice how even though we can directly reach states D and F, no other state returns back to A, so it is in a communication class with just itself. On the other hand for the states D, E, and F, we can make a complete cycle between these three states, meaning they all lead to one another and thus form a communication class. Similarly B and C form their own separate communication class. As a result, we can divide P into the following communication classes: $\{A\}$, $\{D, E, F\}$, and $\{B, C\}$.

Now that we have established communication classes, let us continue to a related topic: irreducible chains.

Definition 2.5. We call a Markov chain $\{X_n\}_{n\in\mathbb{N}}$ with state space S **irreducible** if:

$$\forall i, j \in S, \exists n \in \mathbb{N} : \quad p_{ij}^{(n)} > 0$$

Remark 2.6. Notice that by definition, a Markov chain is irreducible if and only if there is only one communication class in the entire state space.

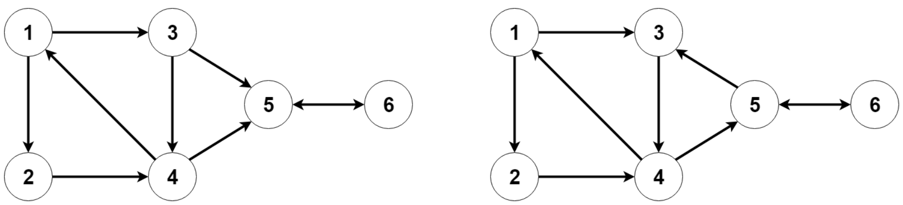


FIGURE 2. Left chain is not irreducible. Right chain is irreducible.

In Figure 2, the Markov chain on the left is not irreducible because starting from states 5 and 6, we cannot reach states 1-4. On the other hand, the Markov chain on the right is irreducible because from every state we can reach every other state.

Simply put, a Markov chain with state space S is irreducible if starting at any state in S, you can reach any other state in S in finite time.

2.2. **Recurrence, Transience, and Periodicity.**

Definition 2.7. We say that a state i of a Markov chain is **recurrent** if

$$p_{ii}^{(n)} = 1$$

for infinitely many n.

Definition 2.8. We say that a state i of a Markov chain is **transient** if

$$p_{ii}^{(n)} = 0$$

for infinitely many n.

In other words, a state of a Markov chain is recurrent if we eventually always come back to said state, and a state is transient if we eventually never come back to said state.

In relation to determining the recurrence and transience of the state, it makes sense to establish how long it takes for the Markov chain to return to the state, which we will call passage time.

Definition 2.9. We define the $(n+1)$-th **passage time** of state i recursively given the initial passage time:

$$T_i^{(0)} = 0 \quad \forall i$$

$$T_i^{(n+1)} = \min\{m \geq (T_i^{(n)} + 1) : X_m = i\} \quad \forall n \in \mathbb{N}$$

And we say $T^{(n)} = \infty$ if the chain never returns to i.

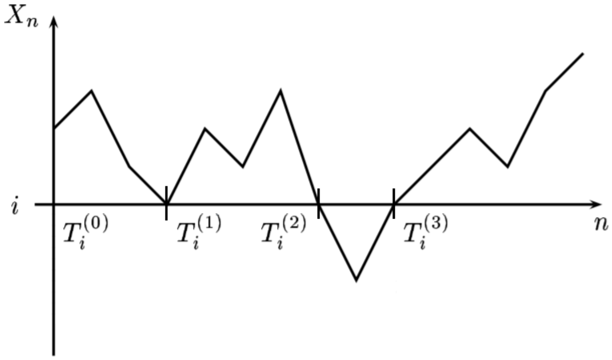


FIGURE 3. Graph showing passage time to state i. Image from [2].

We can then dichotomize recurrence and transience based on the probability of a finite passage time. But first, we must define a number of visits, V_i:

$$f_i^{(n)} = \begin{cases} 1 & X_n = i \\ 0 & X_n \neq i \end{cases} \qquad V_i = \sum_{n=0}^{\infty} f_i^{(n)}$$

and a return probability:

$$r_i = \mathbf{P}_i(T_i < \infty)$$

And now we must show a property of the return probability in terms of the number of visits.

Lemma 2.10. *Given any $n \in \mathbb{N}$, $\boldsymbol{P}_i(V_i > n) = r_i^n$.*

Proof. Suppose without loss of generality that $X_0 = i$, which we can do due to the Markov property. Then the following expressions are equivalent:

$$(V_i > n) = (T_i^{(n)} < \infty)$$

The rest of the proof proceeds by induction:
When $n = 0$, the claim holds by definition. Then suppose that the claim holds for n. Now we will show the claim holds for $n + 1$.

$$\begin{aligned}\mathbf{P}_i(V_i > n+1) &= \mathbf{P}_i(T_i^{(n+1)} < \infty) \\ &= \mathbf{P}_i((T^{(n)} < \infty) \cap ((T_i^{(n+1)} - T_i^{(n)}) < \infty)) \\ &= \mathbf{P}_i((T_i^{(n+1)} - T_i^{(n)}) < \infty | T^{(n)} < \infty)\mathbf{P}_i(T^{(n)} < \infty) \\ &= r_i r_i^n = r^{n+1}\end{aligned}$$

So by induction, the lemma holds. □

Now we can proceed with the theorem.

Theorem 2.11. *If $\boldsymbol{P}_i(T_i < \infty) < 1$, then i is transient. If $\boldsymbol{P}_i(T_i < \infty) = 1$, then i is recurrent.*

Proof. If $\mathbf{P}_i(T_i < \infty) < 1$, then by definition $r_i < 1$. Combining this with the previous lemma, we have

$$\sum_{n=0}^{\infty} \mathbf{P}_i(V_i > n) = \sum_{n=0}^{\infty} r_i^n = \frac{1}{1 - r_i} < \infty$$

It follows that $\mathbf{P}_i(V_i = \infty) = 0$, so i is transient.
If $\mathbf{P}_i(T_i < \infty) = 1$, then by the previous lemma

$$\mathbf{P}_i(V_i = \infty) = \lim_{n \to \infty} \mathbf{P}_i(V_i > n) = 1$$

So we have that i is recurrent. □

Related to the return time of a state, we now define the period of a state.

Definition 2.12. Letting c be any natural number, we say that a state i has a **period** $t > 1$ for $t \in \mathbb{N}$ if

$$\forall n : n \neq ct, \text{ for all } c \in \mathbb{N}, \quad \text{then } p_{ii}^{(n)} = 0$$

and t is the largest integer satisfying this property. If no such $t > 1$ satisfies this property, then we say that the state i is **aperiodic**.

Simply put, this means that given a period t and starting from a given state i, the Markov chain has a 0 probability of returning to i in n steps for any n that is not a multiple of t.

Corresponding to this definition, we say that a Markov chain $\{X_n\}_{n \in \mathbb{N}}$ is aperiodic if every state in $\{X_n\}$ is aperiodic.

Now combining this property with irreducibility, we will prove the existence of a multi-step transition matrix with all positive entries.

Lemma 2.13. *Suppose we have a Markov chain with transition matrix P and finite state space S. If P is both aperiodic and irreducible, then there exists some $N \in \mathbb{N}$ such that for all $n \geq N$:*

$$\forall i, j \in S, \quad p_{ij}^{(n)} > 0$$

Proof. Take any two states $i, j \in S$.
Since P is irreducible, there exists some $m_{ij} \in \mathbb{N}$ such that $p_{ij}^{(m_{ij})} > 0$.
Since P is also aperiodic, there exists some $K_i \in \mathbb{N}$ such that for all $k \geq K_i$, we have that $p_{ii}^{(k)} > 0$.
Then for all such k:

$$\begin{aligned} p_{ij}^{(k+m_{ij})} &= \sum_{l \in S} p_{il}^{(k)} p_{lj}^{(m_{ij})} \\ &\geq p_{ii}^{(k)} p_{ij}^{(m_{ij})} > 0 \end{aligned}$$

We can obtain the first inequality because all of the terms in the transition matrix are non-negative.
Thus we have that for each pair $i, j \in S$, there exists a $N_{ij} := m_{ij} + K_i$ such that for all $n \geq N_{ij}$, we have that $p_{ij}^{(n)} > 0$.
Let $N := \max_{i,j \in S} N_{ij}$, which we can do because S is finite. Then this property holds for all pairs i, j and we have that for all $n \geq N$, $p_{ij}^{(n)} > 0$ for all states i and j in S. □

2.3. **Stationary Distribution.** When studying the long-term behavior of a Markov chain, it is useful to understand the stationary distribution.

Definition 2.14. We say that a vector π with non-negative elements is a **stationary distribution** for a Markov chain with transition matrix P if:

$$\pi P = \pi$$

Once a row of a Markov chain reaches a stationary distribution, then for every future iteration of the process, the row will always maintain this same distribution.

Here is a basic example of how to calculate a stationary distribution of a Markov chain given the transition matrix.

Example 2.15. Suppose we have a Markov chain with two states and the following transition matrix, P:

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

Let us denote the stationary distribution π. In order for the stationary property to hold, we have the following equations:

$$\pi_1 = p\pi_1 + (1-q)\pi_2$$

$$\pi_2 = (1-p)\pi_1 + q\pi_2$$

And since π is a probability distribution, we have:

$$\pi_1 + \pi_2 = 1$$

Now solving this system of equations, we get that:

$$\pi_1 = \frac{q}{p+q} \text{ and } \pi_2 = \frac{p}{p+q}$$

We can check that this indeed is the stationary distribution through matrix multiplication:

$$\begin{aligned}\pi P &= \begin{bmatrix}\frac{q}{p+q} & \frac{p}{p+q}\end{bmatrix}\begin{bmatrix}1-p & p \\ q & 1-q\end{bmatrix} \\ &= \begin{bmatrix}\frac{(1-p)q}{p+q} + \frac{pq}{p+q} & \frac{pq}{p+q} + \frac{(1-q)p}{p+q}\end{bmatrix} \\ &= \begin{bmatrix}\frac{q}{p+q} & \frac{p}{p+q}\end{bmatrix} = \pi\end{aligned}$$

Now we will prove a theorem showing how the stationary distribution relates to the long-run behavior of a Markov chain.

Theorem 2.16. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a Markov chain with transition matrix P and finite state space S. Then suppose for some state $i \in S$:*

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j \quad \forall j \in S$$

Then π is a stationary distribution.

Proof. To show π is indeed a distribution:

$$\sum_{j\in S}\pi_j = \sum_{j\in S}\lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty}\sum_{j\in S} p_{ij}^{(n)} = 1$$

Since the elements in π sum to 1, we know it is a distribution. We can move the limit outside of the summation because S is finite.
Now to show π has the stationary property:

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} p_{ij}^{(n+1)} = \lim_{n\to\infty}\sum_{k\in S} p_{ik}^{(n)} p_{kj} = \sum_{k\in S}\lim_{n\to\infty} p_{ik}^{(n)} p_{kj} = \sum_{k\in S}\pi_k p_k j$$

The third equality comes from the definition of multiplying transition matrices, and we can once again move the limit with respect to the summation because S is finite. This equality implies

$$\pi = \pi P$$

from matrix multiplication. So π is indeed a stationary distribution, and we are done. □

So the limit of a row of a Markov chain as you allow the sequence to continue indefinitely is the stationary distribution. This leads to the question of whether the chain ever reaches this stationary distribution, and if it does, what circumstances are necessary for a Markov chain to converge.

3. Convergence and Mixing Time

Given a Markov chain, it is often useful for us to understand the long-term behavior of the chain. In particular, we want to show that Markov chains eventually converge to their stationary distributions and define how long this takes.

First, we should define a method to measure the distance between two probability distributions, called the total variation distance.

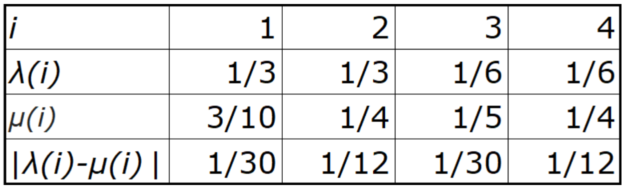
Definition 3.1. For two probability distributions, λ, μ, over state space S, the **total variation distance** is

$$||\lambda - \mu||_{TV} = \max_{I\subseteq S}|\lambda(I) - \mu(I)|$$

With this definition, the distance between two distributions is the largest difference in probability for a single event between both distributions. Let us demonstrate a very simple example of this distance between measures.

Example 3.2. Suppose we have two unfair, "four"-sided dice. The first die has $\frac{1}{3}$ probability to land on 1, $\frac{1}{3}$ probability to land on 2, $\frac{1}{6}$ probability to land on 3, and $\frac{1}{6}$ probability to land on 4. The second die has $\frac{3}{10}$ probability to land on 1, $\frac{1}{4}$ probability to land on 2, $\frac{1}{5}$ probability to land on 3, and $\frac{1}{4}$ probability to land on 4.

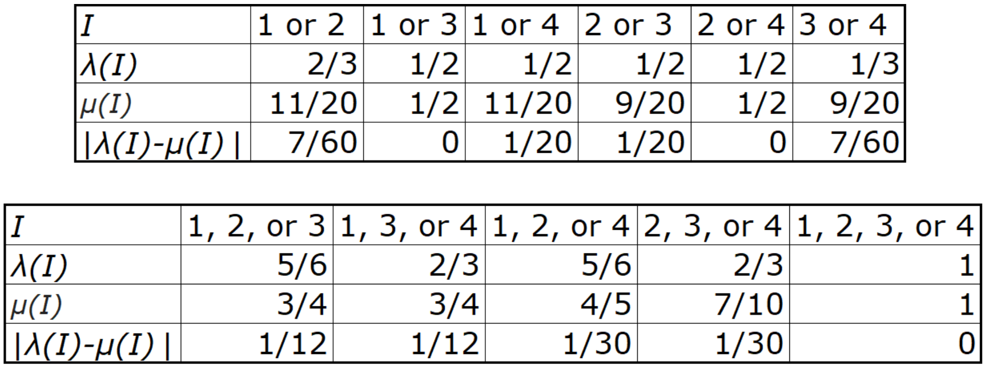
If we call the first die's probability distribution λ and the second die's probability distribution μ, we can generate the following table based on the absolute value of the difference of the two distribution for each state:



i	1	2	3	4
λ(i)	1/3	1/3	1/6	1/6
μ(i)	3/10	1/4	1/5	1/4
|*λ(i)-μ(i)* |	1/30	1/12	1/30	1/12

Figure 4. Differences between each state

But this only accounts for the difference between each state, whereas the total variation distance is the maximum in difference between each event. So we also have to account for combinations of states, such as rolling either a 1 or 2:



I	1 or 2	1 or 3	1 or 4	2 or 3	2 or 4	3 or 4
λ(I)	2/3	1/2	1/2	1/2	1/2	1/3
μ(I)	11/20	1/2	11/20	9/20	1/2	9/20
|*λ(I)-μ(I)* |	7/60	0	1/20	1/20	0	7/60

I	1, 2, or 3	1, 3, or 4	1, 2, or 4	2, 3, or 4	1, 2, 3, or 4
λ(I)	5/6	2/3	5/6	2/3	1
μ(I)	3/4	3/4	4/5	7/10	1
|*λ(I)-μ(I)* |	1/12	1/12	1/30	1/30	0

Figure 5. Differences between each event

Since total variation distance is the maximum difference between each event, in this example, if we let S be the state space, we have:

$$||\lambda - \mu||_{TV} = \max_{I \in S} |\lambda(I) - \mu(I)| = |\lambda(1 \text{ or } 2) - \mu(1 \text{ or } 2)| = \frac{7}{60}$$

Note that the events (1 or 2) and (3 or 4) both produce the maximum distance in the probability distributions and thus both result in the total variation distance.

Before we get into convergence and mixing, let us first establish some basic properties of the total variation distance.

Lemma 3.3. *For two probability distributions λ and μ on state space S, we have:*

$$||\lambda - \mu||_{TV} = \frac{1}{2}\sum_{i\in S}|\lambda(i) - \mu(i)|$$

Proof. Choose any event $I \subseteq S$ and let $J = \{i : \lambda(i) \geq \mu(i)\}$. Then we have that:

$$\lambda(I) - \mu(I) \leq \lambda(I \cap J) - \mu(I \cap J) \leq \lambda(J) - \mu(J)$$

The first inequality is true because we are removing all $i \in (I \cap J^c)$ from the left hand side, and so every i we are removing satisfies $\lambda(i) < \mu(i)$, so the difference can only increase. Similarly for the second inequality, we are only adding $i \in J$, thus regardless of what states we are now including, the difference again can only increase.

Likewise, we have:

$$\mu(I) - \lambda(I) \leq \mu(I \cap J^c) - \lambda(I \cap J^c) \leq \mu(J^c) - \lambda(J^c)$$

Notice that if we take the difference between the right hand sides of both of the inequalities, we get:

$$\begin{aligned}\lambda(J) - \mu(J) - [\mu(J^c) - \lambda(J^c)] &= [\lambda(J) + \lambda(J^c)] - [\mu(J) + \mu(J^c)]\\ &= 1 - 1 = 0\end{aligned}$$

We get the third equality because λ and μ are probability distributions. This shows that $\lambda(J) - \mu(J) = \mu(J^c) - \lambda(J^c)$. And since these both act as upper bounds for the difference between any given state of λ and μ, we can take $I = J$ to get:

$$\lambda(J) - \mu(J) = \mu(J^c) - \lambda(J^c) = \max_{I\subseteq S}|\lambda(I) - \mu(I)| = ||\lambda - \mu||_{TV}$$

Finally, notice that given our definition of J, we have:

$$[\lambda(J) - \mu(J)] + [\mu(J^c) - \lambda(J^c)] = \sum_{i\in S}|\lambda(i) - \mu(i)|$$

This is true because the first summand on the left hand side gives the sum of all $(\lambda(i)-\mu(i))$ for $\lambda(i) \geq \mu(i)$ and the second summand gives the sum of all $(\mu(i)-\lambda(i))$ for $\mu(i) > \lambda(i)$ and thus we get the entire sum across S. Thus

$$||\lambda - \mu||_{TV} = \frac{1}{2}[\lambda(J) - \mu(J) + \mu(J^c) - \lambda(J^c)] = \frac{1}{2}\sum_{i\in S}|\lambda(i) - \mu(i)|$$

□

This gives us a method of defining the total variation distance through summing distances across the whole state space rather than on just a single state.

From the example above, we can use this method to obtain the total variation distance:

$$\begin{aligned}||\lambda - \mu||_{TV} &= \frac{1}{2}\sum_{i\in S}|\lambda(i) - \mu(i)|\\ &= \frac{1}{2}(\frac{1}{30} + \frac{1}{12} + \frac{1}{30} + \frac{1}{12}) = \frac{7}{60}\end{aligned}$$

and we get the same result.

Furthermore, this lemma gives us an important property:

Remark 3.4. From Lemma 3.2, we can see that total variation distance follows the triangle inequality. That is, for probability distributions λ, μ, ν:

$$||\lambda - \mu||_{TV} \leq ||\lambda - \nu||_{TV} + ||\nu - \mu||_{TV}$$

Now that we have described a method to measure distance, we can prove convergence.

3.1. **Convergence Theorem.**

Theorem 3.5. *Suppose $\{X_n\}_{n \in \mathbb{N}}$ is a Markov chain with state space S, transition matrix P, and stationary distribution π. If P is aperiodic and irreducible, then there exists $C > 0, \alpha \in (0, 1)$ such that:*

$$\max_{i \in S} ||P^t_{i,\cdot} - \pi||_{TV} \leq C\alpha^t$$

Proof. Because P is both aperiodic and irreducible, we know by Lemma 2.13 that there exists some n such that P^n has all positive elements. Since P^n is strictly positive, there exists $\delta > 0$ small enough such that:

$$p^{(n)}_{ij} \geq \delta\pi_j \quad \forall i, j \in S$$

Now let us define a matrix Π of the same size as P with all rows as the stationary distribution:

$$\Pi_{i,\cdot} = \pi \quad \forall i \in S$$

If we let $\gamma = 1 - \delta$, then we can define a new stochastic matrix Q with:

$$P^n = (1 - \gamma)\Pi + \gamma Q$$

Notice that if we have a stochastic matrix R satisfying $\pi R = \pi$, it follows that $R\Pi = \Pi$ and $\Pi R = \Pi$.

To continue the proof, we need to show the following claim:

Lemma 3.6. $P^{nm} = (1 - \gamma^m)\Pi + \gamma^m Q^m$

Proof. Continue by induction on m. If $m = 1$ then the claim holds by definition. Now, assuming the claim is true for m, we want to prove the claim for $m + 1$:

$$\begin{aligned}
P^{n(m+1)} &= P^{nm}P^n \\
&= ((1 - \gamma^m)\Pi + \gamma^m Q^m)P^n \\
&= (1 - \gamma^m)\Pi P^n + \gamma^m Q^m P^n \\
&= (1 - \gamma^m)\Pi P^n + \gamma^m Q^m((1 - \gamma)\Pi + \gamma Q) \\
&= (1 - \gamma^m)\Pi P^n + (1 - \gamma)\gamma^m Q^m\Pi + \gamma^{m+1}Q^{m+1} \\
&= (1 - \gamma^{m+1})\Pi + \gamma^{m+1}Q^{m+1}
\end{aligned}$$

Where the second equality holds by the inductive hypothesis, the third equality holds by distributing through by P^n, the fourth equality holds by expanding P^n, and the final equality holds by the property that $\Pi P^n = \Pi$ and $Q^n\Pi = \Pi$

So by induction, the claim holds. □

Now for some k we can multiply both sides by P^k to get:

$$P^{nm+k} = (1 - \gamma^m)\Pi P^k + \gamma^m Q^m P^k$$

Since $\Pi P^k = \Pi$, we can distribute and rearrange the terms to get:

$$P^{nm+k} - \Pi = \gamma^m(Q^m P^k - \Pi)$$

Then for some $i \in S$, we can sum on the ith row on both sides and divide by 2 to get:

$$\frac{1}{2}\sum_{j\in S}(P_{ij}^{nm+k} - \pi_j) = ||P_{i,\cdot}^{nm+k} - \pi||_{TV} = \frac{1}{2}\gamma^m \sum_{j\in S}(Q_{ij}^m P_{ij}^k - \pi_j)$$

Notice that since the summation on the right is a total variation distance between two distributions, the sum is at most 1. Therefore, we get the inequality:

$$||P_{i,\cdot}^{nm+k} - \pi||_{TV} \leq \gamma^m$$

If we take $\alpha = \gamma^{\frac{1}{n}}$ and $C = \frac{1}{\gamma}$, then the original statement holds. □

With this theorem, it holds that Markov chains with these properties will eventually converge. While we now know the convergence of these chains, it also helps to define a method of measuring how long it takes for these chains to converge to their stationary distributions.

3.2. **Mixing Time.**

Definition 3.7. The **mixing time** is a function that measures how much time, t, it takes for a Markov chain with transition matrix P and state space S to reach a total variation distance of within a parameter ϵ of its stationary distribution π.

$$t_{mix}(\epsilon) := \min\{t : \max_{x\in S} ||P_{x,\cdot}^t - \pi||_{TV} < \epsilon\}$$

This general equation gives a measure of how long it takes for all rows of the multi-step transition matrix of a Markov chain to reach within an epsilon neighborhood of the stationary distribution.

Note that it is often common to define the standard mixing time as:

$$t_{mix} := t_{mix}(1/4)$$

Now let us define a method to estimate mixing time.

3.3. **Coupling.**

Definition 3.8. Let μ and ν be probability distributions. A **coupling** of μ and ν is a pair of random variables (X, Y) on the same state space such that:

$$\mathbf{P}(X = x) = \mu(x) \quad \text{and} \quad \mathbf{P}(Y = y) = \nu(y)$$

For mixing times, coupling is a tool that allows us to bound, and thus come up with an estimate for, the mixing time of a Markov chain.

But before we can get to that, let us prove a theorem relating coupling to total variation distance.

Theorem 3.9. *Let a pair of random variables (X, Y) be a coupling of two distributions μ and ν. Then we have:*

$$||\mu - \nu||_{TV} \leq \boldsymbol{P}(X \neq Y)$$

Proof. Let S be the state space of both μ and ν. Take any event $I \subseteq S$. Since (X, Y) is a coupling of μ and ν, we have:

$$\begin{aligned}\mu(I) - \nu(I) &= \mathbf{P}(X \in I) - \mathbf{P}(Y \in I)\\ &\leq \mathbf{P}(X \in I) - [\mathbf{P}(Y \in I) - \mathbf{P}[(X \notin I) \cap (Y \in I)]]\\ &= \mathbf{P}[(X \in I) \cap (Y \notin I)]\\ &\leq \mathbf{P}(X \neq Y)\end{aligned}$$

Since this inequality holds for any event $I \subseteq S$, the result also holds for the total variation and we are done. $\square$

With this theorem we have that the total variation distance is related to coupling in the sense that the probability that the random variables are different is always greater than or equal to the total variation distance. This can give us upper bounds on total variation distance.

We will now explore more on how coupling can estimate mixing times, but first, it is convenient to define two functions:

$$d(t) := \max_{i \in S} ||P^t_{i,\cdot} - \pi||_{TV}$$
$$\rho(t) := \max_{i,j \in S} ||P^t_{i,\cdot} - P^t_{j,\cdot}||_{TV}$$

where P is the transition matrix of a Markov chain and i, j are states in the state space S.

Notice that $d(t)$ is the expression we are trying to get within an epsilon neighborhood for mixing time, and $\rho(t)$ can be thought of as the maximum total variation distance between two separate Markov sequences at time t that have the same transition matrix P but not necessarily the same initial distribution.

Theorem 3.10. *Using $d(t)$ and $\rho(t)$ as defined above:*

$$d(t) \leq \rho(t) \leq 2d(t)$$

Proof. For the second inequality, since Remark 3.3 establishes the triangle inequality for total variation distance, we have:

$$\begin{aligned} \rho(t) &= \max_{i,j \in S} ||P^t_{i,\cdot} - P^t_{j,\cdot}||_{TV} \\ &= \max_{i,j \in S} ||P^t_{i,\cdot} - \pi + \pi - P^t_{j,\cdot}||_{TV} \\ &\leq \max_{i,j \in S} (||P^t_{i,\cdot} - \pi||_{TV} + ||P^t_{j,\cdot} - \pi||_{TV}) \\ &= 2d(t) \end{aligned}$$

Now to show the first inequality, notice that if we have state space S, for any set $A \subseteq S$:

$$\pi_A = \sum_{j \in S} \pi_j P^t_{j,A}$$

We get this from the stationary property of the distribution along with matrix multiplication.

Since this holds for arbitrary A, we have:

$$\begin{aligned} d(t) &= \max_{i \in S, A \subseteq S} |P^t_{i,A} - \pi_A| \\ &= \max_{i \in S, A \subseteq S} |\sum_{j \in S} \pi_j [P^t_{i,A} - P^t_{j,A}]| \\ &\leq \max_{i \in S, A \subseteq S} \sum_{j \in S} \pi_j |P^t_{i,A} - P^t_{j,A}| \\ &\leq \max_{i \in S} \sum_{j \in S} \pi_j ||P^t_{i,\cdot} - P^t_{j,\cdot}||_{TV} \\ &\leq \rho(t) \sum_{j \in S} \pi_j = \rho(t) \end{aligned}$$

The first inequality comes form the triangle inequality, and the second inequality comes from the definition of total variation distance being the maximum of that expression. The final line comes from the fact that since π is a probability distribution, all of its elements are less than or equal to 1, and all sum up to 1.
With this inequality established, we are done. □

So now we have established another upper bound for the mixing time, and now we'll show how coupling ties into this.

To proceed, we define a **Markovian coupling**, where we couple two Markovian chains $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$, such that X_n and Y_n have the same state space and transition matrix, though they do not necessarily have the same initial distribution.

For this paper, we will say that once the two Markovian chains of a coupling meet at the same state at the same time, that they will run congruously from there on out:

$$\mathbf{P}(X_t = Y_t | X_s = Y_s) = 1 \quad \forall t \geq s \tag{3.11}$$

Let us proceed by defining a method to measure the time it takes for a Markovian coupling to reach this state:

Definition 3.12. For a Markovian coupling $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$, let the **coalescence time** τ_{couple} be:

$$\tau_{couple} := \min\{t : X_n = Y_n, \quad \forall n \geq t\}$$

Example 3.13. Let us demonstrate an example of the coalescence time of a Markovian coupling using the simple random walk example from above.

Let $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$, be a Markovian coupling such that both X_n and Y_n have the coin-flip distribution stated above, where for both sequences there is a $\frac{1}{2}$ probability the next step will be an increment of 1 and a $\frac{1}{2}$ probability the next step will be a decrement of 1. This time, however, let $X_0 = 2$ and $Y_0 = -2$ so that the sequences do not start together. Also let our simulation satisfy equation (3.11) from above.

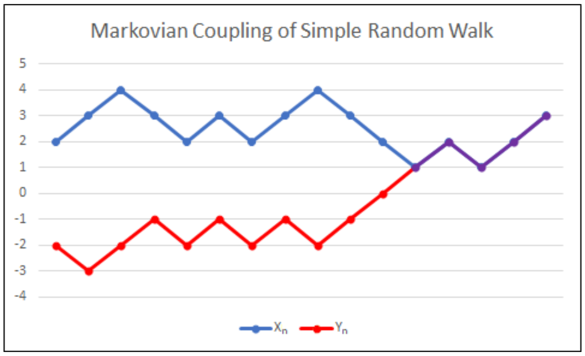


Figure 6. A simulated Markovian coupling of 15 steps

As you can see from the simulation in Figure 6, the two simple random walks meet at $n = 11$, and from our stipulation in equation (3.11), once the chains meet, they continue to run congruously. This also means that in this example, our Markovian coupling has a coalescence time of $\tau_{couple} = 11$.

Now we can provide another upper bound for mixing time through Markovian coupling.

Theorem 3.14. *Suppose we have a Markovian coupling* $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ *such that* $X_0 = x$ *and* $Y_0 = y$*. We have:*

$$||P^t_{x,\cdot} - P^t_{y,\cdot}||_{TV} \leq \boldsymbol{P}(\tau_{couple} > t)$$

Proof. Notice that since $X_0 = x$ and $Y_0 - y$, (X_n, Y_n) is a coupling of $(P^t_{x,\cdot}, P^t_{y,\cdot})$. Also note that because of our previous assumption from equation (3.11), we have that $\mathbf{P}(\tau_{couple} > t) = \mathbf{P}(X_t \neq Y_t)$. Together, we get:

$$||P^t_{x,\cdot} - P^t_{y,\cdot}||_{TV} \leq \mathbf{P}(X_t \neq Y_t) = \mathbf{P}(\tau_{couple} > t)$$

as desired. □

Corollary 3.15. *Suppose we have a Markovian coupling* $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ *over state space* S *such that* $X_0 = x$ *and* $Y_0 = y$*. Then we have:*

$$d(t) \leq \max_{x,y\in S} \boldsymbol{P}(\tau_{couple} > t)$$

Proof. Take the maximum of both sides in Theorem 3.14 and directly apply to the left inequality of Theorem 3.10. By combining these two inequalities, we obtain the desired result. □

As a result of this corollary, we get:

$$t_{mix} \leq 4 \max_{x,y\in S} \mathbf{E}(\tau_{couple})$$

As shown by these theorems, coupling provides us different methods of finding upper bounds for mixing times and thus can give quantitative estimates.

Acknowledgments. I would like to thank my mentor, Kevin Casto, for guiding me through this entire paper. From choosing the topic to helping me understand more about Markov chains, Kevin was always available to give me guidance as to what direction I should take this paper, and he was supportive along the entire way. I would also like to thank Professor Peter May for organizing this REU program and giving me the opportunity to participate.

References

[1] David A. Levin, Yuval Peres, Elizabeth L. Wilmer. *Markov Chains and Mixing Times, second edition.* http://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf

[2] Gordan Žitković. Introduction to Stochastic Processes.
http://www.ma.utexas.edu/users/gordanz/lecture_notes_page.html

[3] Gregory F. Lawler. *Introduction to Stochastic Processes.* Chapman & Hall. 1995.

[4] James Norris. *Markov Chains.* Cambridge University Press. 1998.
http://www.statslab.cam.ac.uk/ james/Markov/

[5] Nathanaël Berestycki. *Mixing Times of Markov Chains: Techniques and Examples: A Crossroad between Probability, Analysis and Geometry.*
http://www.statslab.cam.ac.uk/ beresty/teach/Mixing/mixing3.pdf

[6] William Feller. *An Introduction to Probability Theory and Its Applications.* Volume I. Third Edition. John Wiley & Sons, Inc. 1968.