Mixing Times of Markov Chains: An Examination of Techniques and Applications

Author: Aditya Vaze

Advisor: Dr. Christopher Hoffman

May 30, 2011
1 Motivation

A Markov chain is a set of random variables, \( \{X_t\} \) with \( t = 0, 1, \ldots \) with the property

\[
P(X_t = j | X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}) = P(X_t = j | X_{t-1} = i_{t-1})
\]

This implies that the future is conditionally independent of the past. This property is referred to as the \textit{Markov Property}.

Markov chains are critical in analyzing trends, patterns and predicting events. They serve as mathematical models of random processes where given the initial parameters, predictions can be made about future states. Applications include the natural and social sciences, queuing theory, economics, finance, and computer science, among others. Here, as an introduction, I present a practical application of Markov chains in relation to random sampling.

Suppose we are interested in the average income of a voter in the US. Let \( \mu : S \to \mathbb{R} \), be a probability distribution (\( \sum_{s \in S} \mu(s) = 1 \)) where \( S \) represents a finite set (the population of eligible voters in the US). Here, \( \mu \) describes the probability of selecting a given voter from the set \( S \). Define a function \( f : S \to \mathbb{R} \), where \( f \) represents a voter’s income. We are interested in calculating

\[
E(f) = \sum_{s \in S} f(s) \mu(s).
\]

The Monte Carlo method is as follows. Suppose \( \{X_i\}_{i=1}^n \) are independent identically distributed random variables. \( P(X_i = s) = \mu(s), \forall s \in S \). Then, \( E(f) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n f(X_i) \) by the Law of Large Numbers. If we sample long enough, i.e. as \( n \) approaches the population of the US, we approximate the expectation well. However, this method is infeasible because of the difficulties associated with independent sampling.

The problem of independent sampling is compounded in cases where there is no central register of the eligible population for a given study. For example, we may be interested in finding the average income of jazz musicians. Often, once a particular individual is found, he is asked for references to others potentially eligible participants for the study. The problem with this methodology is that a musician may give references to other musicians in his income bracket. For example, a highly successful musician may give references to other highly successful musicians. This creates serious issues
in independent sampling that are difficult to overcome.

Instead, we can utilize the Markov Chain Monte Carlo method. We set the stationary distribution of the chain as the distribution from which we want to sample. Let \( \{Y_i\}_{i=1}^{\infty} \) be a Markov chain on \( S \) with stationary distribution \( \pi \), where \( \pi = \mu \), and mixing time \( N \). (We can construct a Markov chain with stationary distribution \( \pi \) using the Metropolis-Hastings Algorithm.) Pick any \( s_0 \in S \) and let \( Y_1 = s_0 \). Set \( X_i' = Y_{i,N} \). Then \( X_i' \) is nearly independent and identically distributed, though not exactly because of the difficulty in perfectly sampling from \( \mu \). Then, \( E(f) \approx \frac{1}{N} \sum_{i=1}^{N} f(X_i') \). Once the chain has been run to its mixing time, regardless of the initial conditions, the distribution we are sampling from is independent and very close to \( \mu \).

The main issue with this method lies in discovering how long the Markov chain should be run. We will show that any finite irreducible, aperiodic chain has a unique stationary distribution \( \pi \) and that as the chain is run \( (t \to \infty) \), it will reach \( \pi \). We are interested in calculating the mixing time, or the time \( t \), after which the Markov chain will approximate its stationary distribution. This thesis will present techniques used in computing mixing times, the rationale behind these techniques, along with examples to illustrate real-life applications.

2 Definitions

Here, I will clarify some of the definitions that will be used throughout the thesis.

1) \( \Omega \) = the state space of a Markov Chain, \( \pi \) = stationary distribution of a Markov chain

2) \( P_{ij} = P(X_{i+1} = j|X_i = i) \), where \( P \) represents the transition matrix of the Markov chain. \( P_{ij} \) can also be written as \( P(i,j) \).

3) The period of the chain is \( \gcd(T(x)) \), where \( T(x) = \min(t \geq 1 : P^t(x,x) > 0) \). A chain is called aperiodic if all states have period 1.

4) A chain is irreducible if \( \forall x,y \in \Omega, \exists t \geq 0 \), such that \( P^t(x,y) > 0 \).

5) An aperiodic, irreducible Markov Chain is called ergodic.

5) \( d(t) = \max_{x \in \Omega} ||P^t(x,\cdot) - \pi||_{TV} \)

6) The mixing time is defined as: \( t_{mix}(\varepsilon) = \min\{t : d(t) \leq \varepsilon \} \)
3 Total Variation

The total variation technique provides a good metric for measuring the distance between distributions. We will eventually use it to find the distance between a Markov chain and the chain's stationary distribution.

The total variation between two probability distributions $\mu$ and $\nu$ on $\Omega$ is defined as

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

**Theorem:** $\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$, where $\mu$ and $\nu$ are two probability distributions defined on $\Omega$. [1]

**Proof:** Let $B = \{x : \mu(x) \geq \nu(x)\}$ and let $A$ be any event of $\Omega(A \subseteq \Omega)$.

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

The first inequality holds because $A \cap B$ only contains $x$ such that $\mu(x) \geq \nu(x)$. In other words, the intersection eliminates events $x \in A$ such that $\mu(x) - \nu(x) \leq 0$. The second inequality holds because adding more elements of $B$ cannot decrease the probability because of the definition of the set.

Similarly, we have $\mu(A) - \nu(A) \geq \mu(B^c) - \nu(B^c)$ because the right hand side will always be negative. This implies $\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c)$.

We can partition $\Omega$ into 2 areas: $B$ and $B^c$ with the above definitions. Define Region I as the area between $\mu(x)$ and $\nu(x)$ for $x \in B$, and define Region II as the area between $\mu(x)$ and $\nu(x)$ for $x \in B^c$.

$\mu(B) + \mu(B^c) = 1$ and $\nu(B) + \nu(B^c) = 1$ because they are probability distributions.

So, $\mu(B) + \mu(B^c) - \nu(B) - \nu(B^c) = 0 \Rightarrow \nu(B^c) - \mu(B^c) = \mu(B) - \nu(B)$.

Taking $A = B$ (or $B^c$), we have $|\mu(A) - \nu(A)| = \mu(B) - \nu(B) = \nu(B^c) - \mu(B^c)$.

So, $\|\mu - \nu\|_{TV} = \frac{1}{2} (\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|$

Q.E.D.
4 Coupling

Another technique that we will examine is coupling. Coupling reduces the comparison of two probability distributions to a comparison of two random variables.

A coupling of two probability distributions $\mu$ and $\nu$ is a pair of random variables $(X, Y)$ defined on one probability space such that $P(X = x) = \mu(x)$ and $P(Y = y) = \nu(y)$ for all $x$ and $y$.

A coupling of Markov Chains with transition matrix $P$ is defined as a process $(X_t, Y_t)_{t=0}^\infty$ with the property that $(X_t)$ and $(Y_t)$ are Markov Chains with transition matrix $P$. We will later use coupling to prove that Markov chains converge to their stationary distribution.

**Theorem:** Let $\mu$ and $\nu$ be two probability distributions on $\Omega$.

1. For any coupling of $\omega$ of $(\mu, \nu)$, if $(X, Y)$ is a random variable distributed according to $\omega$, then:

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

**Proof:** Define $a \wedge b = \min\{a, b\}$. For any $z \in \Omega$,

$$\mu(z) = P(X = z) = P[(X = z) \wedge (Y = X)] + P[(X = z) \wedge (Y \neq X)]$$

$$\leq P(Y = z) + P[(X = z) \wedge (Y \neq X)] = \nu(z) + P[(X = z) \wedge (Y \neq X)]$$

So, $\mu(z) - \nu(z) \leq P[(X = z) \wedge (Y \neq X)]$

An exactly symmetric argument gives us, $\nu(z) - \mu(z) \leq P[(Y = z) \wedge (X \neq Y)]$ for any $z \in \Omega$. 
Thus,

\[2\|\mu - \nu\|_{TV} = \sum_{z \in \Omega} |\mu(z) - \nu(z)| = \sum_{z \in \Omega, \mu(z) \geq \nu(z)} |\mu(z) - \nu(z)| + \sum_{z \in \Omega, \mu(z) < \nu(z)} |\mu(z) - \nu(z)|\]

\[\leq \sum_{z \in \Omega, \mu(z) \geq \nu(z)} P[(X = z) \wedge (Y \neq X)] + \sum_{z \in \Omega, \mu(z) < \nu(z)} P[(Y = z) \wedge (X \neq Y)]\]

\[\leq \sum_{z \in \Omega, \mu(z) \geq \nu(z)} P[(X = z) \wedge (Y \neq X)] + \sum_{z \in \Omega, \mu(z) < \nu(z)} P[(Y = z) \wedge (X \neq Y)]\]

\[\leq P(X \neq Y) + P(X \neq Y)\]

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

Q.E.D

**Corollary:** There exists an optimal coupling \(\omega^*\) of \((\mu, \nu)\) such that

\[P(X \neq Y) = \|\mu - \nu\|_{TV}\]

**Theorem:** Let \((X_t, Y_t)\) be a coupling with starting states \(X_0 = x\) and \(Y_0 = y\). Let

\[\tau_{\text{couple}} = \min\{t \geq 1 : X_t = Y_t\}\]

This is called the coupling time. Then,

\[\|P^t(x, \cdot) - P^t(y, \cdot)\|_{TV} \leq P(\tau_{\text{couple}} > t)\]

**Proof:** We know that \(P^t(x, z) = P(X_t = z)\) and \(P^t(y, z) = P(Y_t = z)\). So \((X_t, Y_t)\) is a coupling of \(P^t(x, \cdot)\) and \(P^t(y, \cdot)\).

We use the fact that for two probability distributions \(\mu\) and \(\nu\) on \(\Omega\), we have

\(\|\mu - \nu\|_{TV} \leq P(X_t \neq Y_t)\) where \((X_t, Y_t)\) is a coupling of \(\mu\) and \(\nu\).

Then, \(\|P^t(x, \cdot) - P^t(y, \cdot)\|_{TV} \leq P(X_t \neq Y_t)\).

The chains are coupled at \(\tau_{\text{couple}}\) so \(P(X_t \neq Y_t) = P(\tau_{\text{couple}} > t)\). The result follows.

Q.E.D
5 Convergence Theorem

Theorem: Let $P$, a transition matrix, be irreducible and aperiodic and have stationary distribution \( \pi \). Let \( \{X_n\} \) be a Markov Chain with transition matrix \( P \) and stationary distribution \( \lambda \).

Then \( P(X_n = j) \to \pi_j \) as \( n \to \infty \) for \( \forall j \). In particular, \( P^n_{ij} \to \pi_j \) as \( n \to \infty \) for all \( i,j \). \([1,3]\)

Proof: Let \( \{Y_n\} \) be a Markov Chain with transition matrix \( P \) and stationary distribution \( \pi \). Fix a point \( b \) and let \( T = \inf\{n \geq 1 : X_n = Y_n = b\} \). The first step is to show that \( P(T < \infty) = 1 \). We can use a coupling for this. Set \( W_n = (X_n, Y_n) \) on \( \Omega \times \Omega \) where the transition probabilities are as follows: 

\[
\tilde{p}_{i,j}(i,l) = p_{ij}p_{lk}.
\]

Because \( P \) is aperiodic, \( \forall i,j,k,l \) we have \( \tilde{p}^n_{i,j}(i,l) = p^n_{ij}p^n_{lk} > 0 \) for large enough \( n \). \( \tilde{P} \), the transition matrix for \( W \), is irreducible and has a stationary distribution \( \tilde{\pi}_{(i,k)} = \pi_i \pi_k \).

This means that \( \tilde{P} \) is recurrent, or that there is a non-zero probability that the chain will return to a state \( i \). Since \( T \) is the first time that the Chain \( W_n \) reaches state \( (b,b) \), we observe that \( P(T < \infty) = 1 \).

Next we can construct a process \( \{Z_n\} \) such that \( Z_n = \begin{cases} X_n & n < T \\ Y_n & n \geq T \end{cases} \).

\( \{Z_n\} \) has transition matrix \( P \) because each update of the chain is independent of the previous updates by the Markov Property. Furthermore, it does not matter if \( \{Z_n\} \) is updating according to \( \{Y_n\} \) because \( \{Y_n\} \) also has transition matrix \( P \). Since \( \{Z_n\} \) has the same initial distribution as \( \{X_n\} \), and we now know it has a transition matrix \( P \), it will have \( \lambda \) as its stationary distribution. So, \( \{Z_n\} \) is a Markov Chain with stationary distribution \( \lambda \) and transition matrix \( P \).

So we have, \( P(Z_n = j) = P(X_n = j \text{ and } n < T) + P(Y_n = j \text{ and } n \geq T) \)

\[
\|P(X_n = j) - \pi_j\|_{TV} = |P(Z_n = j) - P(Y_n = j)|
\]

\[
= |P(X_n = j \text{ and } n < T) - P(Y_n = j \text{ and } n < T)| \leq P(n < T) \to 0 \text{ as } n \to \infty.
\]

Q.E.D
6 Strong Stationary Times

This section will explore strong stationary times, which provide a useful approach to bound the rate of convergence to stationarity for Markov Chains. We will begin by providing three useful definitions.

Let $\{X_t\}_{t=0}^{\infty}$ be $\Omega$-valued random variable. Then a stopping time is a random variable $T \in \mathbb{N}$ such that the event $\{T = t\}$ depends only on $X_0, X_1, ..., X_t$.

A strong stationary time $\tau$, for an irreducible Markov Chain $\{X_t\}$, with stationary distribution $\pi$, is a stopping time, dependent on the initial state $x$, such that the distribution of $X_\tau$ is $\pi$.

Equivalently,

$$P_x(X_\tau = y) = \pi(y)$$

(Note that the subscript $x$ below the probability represents the initial state of the chain.)

A stopping time $\tau$ for $(X_n : n \geq 0)$ is a strong stationary time if for all $x$

$$P(X_t = x|\tau = t) = \pi(x)$$

This is the same as saying $X_\tau$ has distribution $\pi$ and is independent of $\tau$. Next, we will examine how to utilize strong stationary times to bound convergence time.

**Theorem:** If $T$ is a strong stationary time for $\{X_n\}$ with stationary distribution $\pi$, then for any initial state $x$, \[1,14\]

$$\|P_x^t - \pi\|_{TV} \leq P(T > t | X_0 = x)$$

**Proof:** We know that

$$\|P_x^t - \pi\|_{TV} = \max_{A \subset \Omega} |P_x^t(A) - \pi(A)|$$

For simplicity, let $T_x$ be the strong stationary time for the Markov Chain starting at $x$.

$$P_x^t(A) = P(X_t \in A) = P(X_t \in A, T_x > t) + \sum_{t' \leq t} P(X_{t'} \in A, T_x = t')$$

We are partitioning the probabilities here based on $t$. 

7
\[ P(X_t \in A_t | T_x > t)P(T_x > t) + \pi(A) \sum_{t' \leq t} P(T = t') \]

This follows from the definition of strong stationary time.

\[ = P(X_t \in A_t | T_x > t)P(T_x > t) + \pi(A)[1 - P(T_x > t)] \]

This is because \( P(T_x > t) + \sum_{t' \leq t} P(T = t') = 1 \). Rearrangement yields,

\[ \pi(A) + P(T_x > t)[P(X_t \in A_t | T_x > t) - \pi(A)] \]

Notice that the part in the parentheses is a difference in probabilities and so is bounded in absolute value by 1. So,

\[ |P_x^t(A) - \pi(A)| = |P(T_x > t)[P(X_t \in A_t | T_x > t) - \pi(A)]| \leq P(T_x > t) = P(T > t | X_0 = x) \]

Q.E.D

7 Gambler's Ruin

This section introduces some of the basic concepts of Markov Chains and presents the classic Gambler's Ruin problem.

Consider a gambler who starts with \$i, 0 \leq i \leq N. For each successive gamble, he either wins \$1 or loses \$1, independent of past performance, with probability \( p = 1 - p \) respectively. Let \( R_n \) be his fortune after the \( n \)th gamble, \( n \geq 0 \).

\( R_n \) is a random walk where \( R_n = X_1 + ... + X_n \). \{X_n\} forms a sequence of independent, identically distributed random variables with \( P(X = 1) = p \) and \( P(X = -1) = q = 1 - p \). The chain is absorbed at either \( R_n = 0 \) or \( R_n = N \).

Alternatively we can define \{\( R_n \)\} as a Markov Chain on state space \( \Omega = \{0, ..., N\} \). Here,

\( P(R_{n+1} = i+1 | R_n = i) = p_{i,i+1} = p \) and \( P(R_{n+1} = i-1 | R_n = i) = p_{i,i-1} = q (0 < i < N) \).
Since 0 and N are absorbing states, we define \( p_{00} = p_{NN} = 1. \)

Let \( \tau_i = \{ n \geq 0 : R_n \in \{0, N\} \mid R_0 = i \} \)

Define \( P_i = P(R_{\tau_i} = N) \). \( P_i \) represents the probability of the gambler reaching \( N \) when starting at \( R_0 = i \). \( P_0 = 0 \) because this represents the probability of the gambler reaching \( N \) when he starts at \( S_0 \). \( P_N = 1 \) because this represents the probability of the gambler reaching \( N \) when he starts at \( S_N \).

We can calculate \( P_i \) by conditioning it on the first outcome of the gamble. For example, if \( X_1 = 1 \) (winning the first round), the gambler’s fortune increases to \( R_1 = i + 1 \). We can employ the Markov property and say that now the gambler will reach \( N \) with probability \( P_{i+1} \). A similar situation holds if \( X_1 = -1 \) (losing the first round). The gambler’s fortune decreases to \( R_1 = i - 1 \) and he will reach \( N \) with probability \( P_{i-1} \). We can set this scenario as follows: \( P_i = pP_{i+1} + qP_{i-1} \)

Since \( p + q = 1 \), \( pP_1 + qP_1 = pP_{i+1} + qP_{i-1} \).

This implies \( P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1}) \).

We see that \( P_2 - P_1 = \frac{q}{p}(P_1 - P_0) = \frac{q}{p}P_1 \) since \( P_0 = 0 \).

Similarly, \( P_3 - P_2 = \frac{q}{p}(P_2 - P_1) = \frac{q}{p}^2P_1 \)

This pattern holds for \( 0 < i < N \), where \( P_{i+1} - P_i = \frac{q}{p}^iP_i \).

Now we want to find \( P_i - P_1 \):

\[
P_i - P_1 = \sum_{k=1}^{i} (P_{k+1} - P_k) = \sum_{k=1}^{i} \left( \frac{q}{p} \right)^k P_1.
\]

This implies \( P_{i+1} = P_1 \sum_{k=0}^{i} \left( \frac{q}{p} \right)^k \). The equation can be expanded by observing that this is a geometric sum.

So, \( P_{i+1} = P_1 \frac{1 - \left( \frac{q}{p} \right)^{i+1}}{1 - \frac{q}{p}} \), \( p \neq q \)

We can use the fact that \( P_N = 1 = P_1 \frac{1 - \left( \frac{q}{p} \right)^N}{1 - \frac{q}{p}} \) to calculate that \( P_1 = \frac{1 - \frac{q}{p}}{1 - \left( \frac{q}{p} \right)^N} \)

Thus, \( P_i = \frac{1 - \left( \frac{q}{p} \right)^i}{1 - \left( \frac{q}{p} \right)^N}, p \neq q \)

A nearly identical derivation holds for \( p = q = 0.5 \), a game where the probability of winning and losing are equal (flipping a coin, for example) and yields, \( P_i = \frac{i}{N} \).
Thus,

\[ P_i = \begin{cases} 
\frac{1 - (\frac{3}{2})^i}{1 - \left(\frac{3}{2}\right)^N} & \text{if } p \neq q \\
\frac{1}{N} & \text{if } p = q = 0.5
\end{cases} \]

8 Top to Random Shuffle

This section will demonstrate some of techniques discussed above by examining the Top to Random Shuffle. The Top to Random Shuffle is a slow method of mixing cards. Consider a deck of \( n \) cards. Take the top card and randomly insert it somewhere in the deck.

Consider a sequence of card \( \{a_n\} \). Then for all \( i = 1, \ldots, (n - 1) \),

\[ P(a_1, \ldots, a_n \rightarrow a_2, \ldots, a_i, a_1, a_{i+1}) = \frac{1}{n} \]

**Theorem:** The mixing time of the top to random shuffle of an \( n \) card deck is \( n \log n \). \([9,11]\)

**Proof:** The first step is to show that after a certain time the deck is indeed mixed. Let \( T_1 \) be the first time a card is inserted under the bottom card \( B \). Let \( T_{n-1} \) be the first time \( n - 1 \) cards are under \( B \). This necessarily implies that \( B \) is now on top.. Let \( T = T_{n-1} + 1 \) be the time when card \( B \) is randomly inserted somewhere in the deck. We can check through induction that at this point there are \( n! \) permutations for the deck.

Label the \( i^{th} \) card to be put below \( B \) as \( C_i \). We claim that all \( j! \) arrangements of \( C_1, \ldots, C_j \) are equally likely. The claim is trivial if \( j = 1 \). There are \( j + 1 \) positions under \( B \) that \( C_{j+1} \) could be placed. All are equally likely and independent of the arrangement of cards \( C_1, \ldots, C_j \) by the Markov property. Thus, the \( (j + 1) \) cards have \( (j + 1)! \) equally likely arrangements.

This means that deck is distributed uniformly with each card having probability \( \frac{1}{n} \) of being at a given location. This distribution is the stationary distribution of the chain. We see that \( T \) is **strong stationary time**. Thus, at this time, the Markov Chain has reached its stationary distribution. The deck is mixed.
The next step is to compute how many shuffles it requires for this mixing to occur. We can compute 
\( E(T_{n-1}) \). Let \( T_0 = 0 \) and we observe

\[
T_{n-1} = T_1 + (T_2 - T_1) + (T_3 - T_2) + ... + (T_{n-1} - T_{n-2})
\]

Set \( X_i = T_i - T_{i-1} \) for \( i = 1, ..., n - 1 \). Then

\[
E(T_{n-1}) = E(X_1 + X_2 + ... + X_{n-1}) = E(X_1) + E(X_2) + ... + E(X_{n-1})
\]

by linearity of expectation. Thus,

\[
E(T) = E(T_{n-1} + 1) = E(1) + E(T_{n-1}) = 1 + \sum_{k=1}^{n-1} E(X_k)
\]

At this point it becomes necessary to compute the waiting times. We want to calculate \( E(X_k) \). We know that \( P(X_0 = 0) = 0, P(X_k = 1) = \frac{k}{n}, P(X_k = 2) = (1 - \frac{k}{n}) \frac{k}{n} \). We quickly see that \( X_k \) is a geometric random variable with parameter \( \frac{k}{n} \). So \( P(X_k = j) = (1 - \frac{k}{n})^{j-1} (\frac{k}{n}) \). Then

\[
E(X_k) = \sum_{j=1}^{\infty} (1 - \frac{k}{n})^{j-1} (\frac{k}{n}) = \frac{k}{n}
\]

Thus,

\[
E(T) = 1 + \sum_{k=1}^{n-1} \frac{k}{n} = \sum_{k=1}^{n} \frac{k}{n} \approx n \log n
\]

Next, we can use the coupling method. This time we will look at Random-to-Top Shuffling.

Let \( \{X_t\} \) denote a Markov Chain on the symmetric group \( S_n \). We will be using this group to look at the permutations of an \( n \)-card deck. Denote the initial state \( X_0 \) as the order of the card \( \{1, 2, ..., n\} \). Let \( \{Y_t\} \) be another Markov Chain starting from stationarity. This means the initial state \( Y_0 \) will be uniformly distributed on the set of all \( N! \) permutations. The coupling \( (X_t, Y_t) \) progresses as follows. At each time \( t \) select a random integer \( A_t \) between 1 and \( n \), independent of all previous choices. Then, for each \( t \), select the \( A_t^{th} \) card from both the decks and move it to the the top. Let \( \tau = \min\{t \geq 1 : X_t = Y_t\} \). This is the coupling time or the time when the two Markov chains coincide. At this stage the decks are matched. For each successive \( t \), the decks will coincide i.e. they
are coupled. Using the second proof from the coupling section, we see that total variation between the
distribution of \( X_t \) and its stationary distribution, the distribution of \( Y_t \), is bounded above by \( P(\tau > t) \).
At this point, we will not go further and use the fact that if \( t = (1 + \epsilon)E(\tau) \), then \( P(\tau > t) \to 0 \). This
part of the proof does not utilize Markov Chain techniques but instead uses Chebyshev’s Inequality.
We have shown that at \( \tau \), the total variation goes to 0 and that the decks are mixed.

9 Random Transpositions

This section will provide an example of where the method of strong stationary times gives a tighter
bound than coupling.

A random transposition is a simple card shuffling technique. At a time \( t \), a person chooses a card
\( X_t \in \{1, 2, \ldots, n\} \) (from a deck of \( n \) cards) and a position \( Y_t \in \{1, 2, \ldots, n\} \), independently of \( X_t \). The
person then transposes card \( X_t \) with the card in position \( Y_t \). If \( X_t = Y_t \) then no exchange is made.

First let us examine coupling.

We can couple two decks by using the same choices \( X_t \) and \( Y_t \) to shuffle both decks. Let \( \{\sigma_t\} \) and
\( \{\sigma_t^*\} \) represent the chains for each deck. We can examine coupling by counting how many cards are
in the same position in both decks. Let \( a_t \) be the number of cards in the same positions in \( \{\sigma_t\} \) and
\( \{\sigma_t^*\} \). There are three cases:

1. If card \( X_t \) is in the same position in both decks and this card also occupies the same position \( Y_t \) in
both decks, then \( a_{t+1} = a_t \). This corresponds to no switch has been made.
2. If card \( X_t \) is in different positions in both the decks but the card at position \( Y_t \) is the same, then
\( a_{t+1} = a_t \). This is because the transposition breaks one match but immediately creates another one,
namely the one at position \( Y_t \).
3. If card \( X_t \) is in different positions in both the decks and the cards at position \( Y_t \) are different, then
at least one new alignment is made.
**Theorem:** Let $\tau$ be the time required for the decks to couple. Then, regardless of their initial configurations, $^{[1,5]}$

$$E(\tau) < \frac{n^2}{6}n^2$$

**Proof:** We can decompose $\tau$ into the sum of the number of transpositions depending on the value of $a_t$.

$$\tau = \tau_1 + \tau_2 + \ldots + \tau_n$$

where $\tau_i$ is the number of transpositions from when $a_t$ is above $i - 1$ and when $a_t$ is above $i$. For a $t$ such that $a_t = i$, there are $n - i$ unaligned cards. The probability of increasing the number of matches is the product of selecting an unmatched card and an unmatched position. Thus, the probability of increasing the number of matches is $\frac{(n-i)^2}{n^2}$.

We observe that $\tau_{i+1}$ is a geometric variable with probability of success $\frac{(n-i)^2}{n^2}$. So,

$$E(\tau_{i+1}|a_t = i) = \frac{1}{\frac{(n-i)^2}{n^2}} = \frac{n^2}{(n-i)^2}$$

When there does not exist a $t$ such that $a_t = i$ then $\tau_{i+1}$ is clearly 0. Thus,

$$E(\tau) < n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2} < n^2 \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}n^2$$

Q.E.D

Now we will examine the use of strong stationary times in random transpositions.

**Theorem:** In a random transposition shuffle of $n$ cards, let $R_t$ and $L_t$ be cards chosen by the right and left hands at time $t$. At $t = 0$, no cards are marked. Then at time $t$, mark card $R_t$ if both the following hold: $^{[1,5,8]}$

1. $R_t$ is unmarked
2. Either $L_t$ is unmarked or $L_t = R_t$.

Let $\tau$ be the first time when all $n$ cards are marked. Then, $\tau$ is a strong stationary time for the chain.
Proof: Let $V_t \subseteq \{1, \ldots, n\}$ be the set of cards marked at (or before) time $t$. Let $U_t \subseteq \{1, \ldots, n\}$ be the set of positions occupied by $V_t$ after the $t^{th}$ transposition. By induction, we will show that given $t$, $V_t$ and $U_t$, all permutations of $V_t$ on $U_t$ are equally likely. This is equivalent to saying the deck of cards has a uniform distribution, which is the stationary distribution for a mixed deck.

At $t = 1$, given a card $V_1$, there are $n$ equally likely positions to place it. The probability of getting a particular $U_t$ is $\frac{1}{n}$. The claim is thus true for $t = 1$.

Now assume the claim is true for $t$ and the shuffler chooses cards $R_{t+1}$ and $L_{t+1}$. Now we have two cases:

1. If no new card is marked, then $V_{t+1} = V_t$. This could happen if cards $L_{t+1}$ and $R_{t+1}$ are different and unmarked, in which case $V_{t+1} = V_t$ and $U_{t+1} = U_t$. Another possibility is if $L_{t+1}$ and $R_{t+1}$ have already been marked. Then, $U_{t+1} = U_t$ and we observe that by definition, these cards are in $V_t$. We transpose these two cards randomly and see that all permutations of $V_t$ are equally likely. Lastly, $R_{t+1}$ is already marked and $L_{t+1}$ is unmarked. Here, we need to update the set $U_{t+1}$. We delete the integer (from the set) corresponding to the position of $R_{t+1}$ and add the integer corresponding to the position of $L_t$. If $U_t$ is fixed then all $R_{t+1} \in U_t$ are equally likely, along with the permutations of $V_t$ on $U_t$. Thus, once the specified integer is deleted from the set, and the specific integer is added to the set (corresponding to the transposition), all permutations of $V_t$ on $U_{t+1}$ are equally likely.

2. If $R_t$ does get marked, then $L_{t+1}$ is equally likely to be any element of $V_{t+1} = V_t \cup \{R_{t+1}\}$. Notice that this corresponds to adding the element $R_{t+1}$ to the set of marked cards, giving the set of marked cards at time $t + 1$. Similarly, $U_{t+1}$ is $U_t$ with the position of $R_{t+1}$ at time $t$. Fixing $L_{t+1}$ and specifying $V_t$ on $U_t$ uniquely determines the permutation of $V_{t+1}$ on $U_{t+1}$. Thus, once again, all permutations are equally likely.

By induction on $t$, we see that $t = \tau$ is a strong stationary time.

Q.E.D
**Theorem:** For the above stopping time $\tau$, $E(\tau) = 2n(\log n + O(1))$ \footnote{1}

**Proof:** As in the coupling method, we can write $\tau$ as

$$\tau = \tau_0 + \ldots + \tau_{n-1}$$

where $\tau_k$ is the number of transpositions after the $k^{th}$ card is marked to reach the point where $(k+1)^{th}$ card is marked. $\tau_k$ is a geometric variable with probability of success $\frac{(k+1)(n-k)}{n^2}$. (To see why this is true, observe that at $\tau_0$, the probability is $\frac{1}{n}$. The only way to mark the first card is if $L_1 = R_1$. At $\tau_1$, the probability is $\frac{2(n-1)}{n^2}$. You have 1 card marked, and a probability $\frac{n-1}{n}$ of choosing an unmarked card. The 2 correspond to either fixing the right hand card and choosing an unmarked card with your left hand, or vice versa.) Since the $\tau_i$ are clearly independent,

$$E(\tau) = \sum_{k=0}^{n-1} \frac{n^2}{(k+1)(n-k)} = \sum_{k=0}^{n-1} \frac{n^2}{n+1} \left( \frac{1}{k+1} + \frac{1}{n-k} \right)$$

We can use the fact that

$$\sum_{j=1}^{n} \frac{1}{j} = \log n + O(1)$$

Thus, after some algebra, $E(\tau) = 2n(\log n + O(1))$.

Q.E.D

These examples are instances where the strong stationary time technique gives a better upper bound than coupling. However, they use a naïve approach to coupling, and by no means does coupling always give a worse bound. Optimal couplings can be found that give better bound than strong stationary times. However, finding such couplings is not always feasible and if they are found, they may be difficult to analyze.

### 10 Riffle Shuffle

We will now examine the Riffle Shuffle, the method most often used to shuffle standard 52-card decks. The deck is cut into 2 piles and the bottom card from each pile is dropped to reconstitute the deck. Here are three methods to model the riffle shuffle of an $n$ card deck.
1. Let $M$ be binomial random variable with parameters $(n, \frac{1}{2})$. Split the deck into two piles, the first having the top $M$ cards and the second having the remaining $n - M$ cards. We can select $M$ positions in an $n$ card deck, and assign $n - M$ cards to the remaining positions. This method will preserve the relative order of the cards. There are $\binom{n}{M}$ methods of riffling the cards.

2. Let $M$ be binomial random variable with parameters $(n, \frac{1}{2})$. Split the deck into two piles, the first having the top $M$ cards and the second having the remaining $n - M$ cards. Drop a card from a pile in relation to the number of cards in the other pile. At a given moment, if the left pile contains $a$ cards and the right pile contains $b$ cards then drop a card from the left pile with probability $\frac{a}{a+b}$ and a card from the right pile with probability $\frac{b}{a+b}$.

3. Select randomly and independently an element from the set $\{0, 1\}$, $n$ times. Assign a value of 0 or 1 to each card accordingly. Place all cards with label 0 at the top of the deck, keeping the relative order intact.

We can consider a riffle shuffle as a permutation $\sigma$ on the symmetric group $S_n$.

Methods 1 and 2 generate the following distribution $Q$.

$$Q(\sigma) = \begin{cases} \frac{n+1}{2^n} & \sigma = \text{id} \\ \frac{1}{2^n} & \text{if } \sigma \text{ has 2 rising sequences} \\ 0 & \text{else} \end{cases}$$

A rising sequence of a permutation $\sigma$ is number of sequences in $\sigma$ that are in the correct relative order.

Method 3 generates the inverse distribution $\hat{Q}$ where $\hat{Q}(\rho) = Q(\rho^{-1})$. We will be using Method 3 for the riffle shuffle analysis. First, we need to cite an important theorem.
**Theorem:** Let $P$ be a transition matrix of a random walk on group $G$, with increment distribution $\mu$ and let $\hat{P}$ be that of the walk on $G$ with increment distribution $\hat{\mu}$, the inverse distribution of $\mu$. Let $\pi$ be the stationary distribution on $G$. Then for any $t \geq 0$: \[ ||P^t(id, \cdot) - \pi||_{TV} = ||\hat{P}^t(id, \cdot) - \pi||_{TV} \]

This theorem says that if a random walk and its inverse start on the identity, they have the same distance from uniformity after $t$ steps. This justifies using Method 3 for the analysis.

We will now define the *inverse shuffle*. For each shuffle, we label the cards 0 or 1 independently with probability $\frac{1}{2}$. We then pull the cards labeled 0 and put them at the top of the deck, preserving their relative order. (Notice that a given permutation, regardless of whether it was derived from a standard shuffle or an inverse shuffle, has probability $\frac{1}{2^n}$.) For the subsequent shuffle, this process is repeated and the *bit*, $\{0, 1\}$, for that shuffle is placed on the left of the bit for the previous shuffle. For example, a card with label 100 after three shuffles, was assigned 0 on the first shuffle, 0 on the second shuffle and 1 on the last shuffle. After $t$-shuffles, each card will have $t$-bit label that is not necessarily unique. Cards with the same label will be in the same relative order.

**Theorem:** Let $\tau$ be the number of inverse shuffles required for all cards to have different labels. Then $\tau$ is a strong stationary time. \[1\]

**Proof:** At $\tau$, each card has label that differs by at least 1 bit and the left-most bit, the one corresponding to the most current shuffle, determines the relative order of the cards. Since the bits are generated independently, each label can be assigned to a card with equal probability. Equivalently, the final order of the cards depends on the label, not on the original identity permutation. So, since the labels are chosen randomly, the deck, conditioned on $\tau$, is in a random permutation. The deck at $\tau$ has a uniform distribution.

Q.E.D
Next we will examine the mixing time for the riffle shuffle.

**Theorem:** For the riffle shuffle of a deck, the mixing time $t_{mix} \leq 2 \log_2 \left( \frac{4n}{3} \right)$ for sufficiently large $n$. [1,9]

**Proof:** We can examine an $n$ card deck and observe that $\tau$ is strong stationary time, by the previous theorem. Thus, $\tau$ is a stopping time. For, $\tau \leq t$, different labels have been assigned to the $n$ cards after $t$ shuffles.

We have:

$$P(\tau \leq t) = \prod_{k=0}^{n-1} \left( 1 - \frac{k}{2^t} \right)$$

This is referring to the probability of having unique labels by time $t$. Observe that this is similar to the classic birthday problem. (What is the probability of 2 people in a room of $n$ people sharing a birthday?) Here, the “people” are cards and there can be $2^t$ labels, which represent possible birthdays. We are trying to find the probability that all “birthdays” are distinct.

Let $t = 2 \log_2 \left( \frac{n}{e} \right)$. Thus, $2^t = \frac{n^2}{2^t}$. Thus, with a little algebra, we have

$$\log \prod_{k=0}^{n-1} \left( 1 - \frac{k}{2^t} \right) = - \sum_{k=0}^{n-1} \left( \frac{c^2 k}{n^2} + O \left( \frac{k}{n^2} \right)^2 \right) = - \frac{c^2 n(n-1)}{2n^2} + O \left( \frac{n^3}{n^4} \right) = - \frac{c^2}{2} + O \left( \frac{1}{n} \right)$$

So,

$$P(\tau \leq t) = \prod_{k=0}^{n-1} \left( 1 - \frac{k}{2^t} \right) = e^{-\frac{c^2}{2} + O \left( \frac{1}{n} \right)}$$

Thus,

$$\|Q^t - U\|_{TV} = \|\hat{Q}^t - U\|_{TV} \leq P(\tau > t) = 1 - P(\tau \leq t) = 1 - e^{-\frac{c^2}{2} + O \left( \frac{1}{n} \right)}$$

where $U$ is the stationary distribution (in this case, the uniform distribution). The first equality holds because a distribution and its inverse have the same distance from stationarity after $t$ steps. The first inequality was proven in Section 6. As $n \to \infty$,

$$\|Q^t - U\|_{TV} \leq 1 - e^{-\frac{c^2}{2}}$$

18
By the definition of mixing, the right hand side must be less than or equal to 0.25. Taking values of $c$ such that $c < \sqrt{2 \log \frac{4}{3}} \approx 0.7585$ approximately yields $t_{mix} = t_{mix}(\frac{1}{c})$. (A standard value for $c$ is 0.75). Substituting this value in $t = 2 \log_2 \left( \frac{n}{c} \right)$ gives the required result.

Q.E.D

Thus, for a standard $n = 52$ deck, using $c = 0.75$, we get $t = 12.23$. So, after $t = 13$ riffle shuffles gives an upper bound for deck mixing. The next theorem will give a lower bound for the mixing time.

**Theorem:** Fix $\epsilon > 0$ and $\delta > 0$. Consider a riffle shuffle on an $n$ card deck. For sufficiently large $n$, $|\Omega|$,

$$t_{mix}(\epsilon) \geq (1 - \delta) \log_2 n$$

**Proof:** First, we will need a general result about lower bounds on mixing times. Let $d_{out}(x) = |\{y : P(x, y) > 0\}|$. This is the number of states accessible from $x$ after 1 step. Let $\Delta = \max_{x \in \Omega} d_{out}(x)$.

Let $\Omega^t_x$ be the states accessible from $x$ after $t$ steps. Thus, $|\Omega^t_x| \leq \Delta^t$.

If $\Delta^t < (1 - \epsilon)|\Omega|$, then

$$\epsilon < 1 - \frac{\Delta^t}{|\Omega|} \leq P_t(x, \Omega^t_x) - \pi(\Omega^t_x) \leq \|P^t(x, \cdot) - \pi\|_{TV}$$

Thus,

$$t_{mix}(\epsilon) \geq \frac{\log(|\Omega|(1 - \epsilon))}{\log \Delta}$$

There are a maximum of $2^n$ states accessible after 1 riffle shuffle. Thus, $\log_2 \Delta \leq n$. We know that $|\Omega| = n!$

Now we can use Stirling's formula to write $\log_2 n!$ as $\log_2 n! = [1 + o(1)]n \log_2 n$.

We can now substitute these results in the above formula. This yields the theorem, for all $\delta > 0$, given that $n$ is sufficiently large.
Q.E.D

11 Spectral Gaps

This section will briefly introduce the relation between the spectral gap of a Markov Chain and its mixing time. This will be the last technique that we cover.

Let \( P \) be a transition matrix for an ergodic Markov Chain. Let \( \{\lambda_i\} \) be its eigenvalues such that \(|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|\). Since the chain is ergodic, we know that it has a stationary distribution \( \pi \). Furthermore, we know that it has an eigenvalue of 1, corresponding to the eigenvector \( \pi \). By the Perron-Frobenius Theorem, we conclude that for a finite, irreducible, aperiodic Markov Chain, \( \lambda_1 = 1 \) and that \( |\lambda_i| < 1 \) for \( 2 \leq i \leq n \). Now, we give the definition of a reversible transition matrix \( P \). A transition matrix \( P \) is reversible with respective to its stationary distribution \( \pi \) if \( \pi(x)P(x,y) = \pi(y)P(y,x) \) for all \( x, y \in \Omega \).

Define \( \lambda_* = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \lambda \neq 1\} \). Then, the absolute spectral gap is defined as \( \gamma_* = 1 - \lambda_* \).

The spectral gap of a reversible chain is defined as \( \gamma = 1 - \lambda_2 \).

The relaxation time \( t_{rel} = \frac{1}{\gamma} \).

An important theorem that relates the above terminology to Markov Chain mixing is given below:

**Theorem:** Let \( P \) be the transition matrix for a reversible, aperiodic and irreducible Markov Chain defined on \( \Omega \). Let \( \pi_{\text{min}} = \max_{x \in \Omega} \pi(x) \). Then, \( \frac{1}{2\epsilon} \leq t_{\text{mix}}(\epsilon) \leq \log(\frac{1}{\epsilon \pi_{\text{min}}})t_{\text{rel}} \).

This theorem provides the upper and lower bounds on mixing time based upon the relaxation time and the stationary distribution. If we can calculate, or at bound, the size of the spectral gap of the Markov Chain, then we can determine the mixing time. We can also change the value of \( \pi_{\text{min}} \) thereby
changing spectral gap. This will help us to determine the effects of altering the state space of the chain.

12 Conclusion

We have seen ways to bound the mixing times of Markov chains through four techniques. Initially, we discussed the method of total variation and how it bounds the distance between two probability distributions. This section hinted at the possibility of bounding a distribution with a Markov Chain’s stationary distribution.

Next, we moved on to coupling. This probabilistic technique allows us to reduce the comparison of two probability distributions to that of two random variables. In this section, we proved two theorems that would be referenced throughout the paper. We followed these two techniques with one of the most important theorems regarding Markov Chains: the Convergence Theorem. This theorem proved that convergence does occur for aperiodic and irreducible chains. At this point, we became more interested in estimating the rate at which convergence to the stationary distribution occurs. We then proceeded to explore strong stationary times. Here, we saw a way to analyze chains in the context of their stationary distributions.

The next few sections of the paper were devoted to exploring the applications of Markov Chains to card shuffling games. The classic Gambler’s Ruin example helped to see how the basic properties of Markov chains can be implemented in real-life problems. Subsequently, we examined the top-to-random and random transposition methods of card shuffling and gave results about the number of shuffling required to achieve mixing. Here, we also saw how different techniques can give better bounds for the same problem. Finally, we decided to give a more practical analysis of shuffling by analyzing the riffle shuffle. In the last section, we briefly explored another technique known as the spectral gap.

As we have seen, Markov Chains are indeed an exciting topic in mathematics. This paper only gives an introduction to this area and there is much left to uncover. We can go on to study topics such as hitting and cover times, discuss more spectral techniques by applying concepts of conductance and contractions. We can also study chains with a not necessarily finite state space or even continuous time chains. Markov chains are a lively area of probability theory and play a central role in applied fields.
They are powerful tools in understanding and interpreting stochastic processes. Their importance and the range of their applications will only increase in the future.

References


