FINITE MARKOV CHAINS AND THE TOP-TO-RANDOM SHUFFLE

PHILIP LIANG

ABSTRACT. In this paper, I present an introduction to Markov chains, basic tools to analyze them, and an example, the top-to-random shuffle. I cover the existence and uniqueness of stationary distributions, the Convergence Theorem, total variation distance, mixing time, and strong stationary times. Using these tools, I show that the top-to-random shuffle on a deck of $n$ cards mixes the deck in approximately $n \log n$ shuffles.

CONTENTS

1. Introduction 1
2. Stationary Distributions of Markov Chains 2
3. The Convergence Theorem 5
4. Approximately when does the top-to-random shuffle fully randomize a deck? 8
5. So when are we sure that we’ve shuffled the deck enough? 13
6. Are we sure it takes so long? 16
Acknowledgments 18
References 18

1. INTRODUCTION

Markov chains are memoryless stochastic processes that can be used to model processes such as card shuffling. There are several natural questions to ask in this context. We would want to know if the shuffling process brings the deck to a uniform equilibrium distribution. Furthermore, the number of shuffles required to make the deck approximately random is important. In this paper, I answer these two questions about the top-to-random shuffle.

One key result that I prove is that irreducibility (all states can eventually reach each other) and aperiodicity are sufficient conditions for the existence of an equilibrium distribution. Furthermore, I show that it is possible to bound from above the mixing time (the fixed number of transitions needed to make a distribution approximately stationary) with a bound on the upper tail of the strong stationary time (the random number of steps required for a chain to reach its stationary distribution). By characterizing a strong stationary time for the top-to-random shuffle, I place an upper bound on its mixing time. Moreover, I place a matching lower bound on the mixing time that shows the upper bound is off only by a constant for

Date: DEADLINES: Draft AUGUST 19 and Final version AUGUST 30, 2013.
sufficiently large decks. Along the way, I also characterize the coupon collector’s
time and the Pólya’s Urn model. My exposition is largely based off of that of Levin,
Peres, and Wilmer in *Markov Chains and Mixing Times*.

2. Stationary Distributions of Markov Chains

Note on notation: \( P_x \{ \text{Event} \} \) denotes the probability that \{ Event \} occurs
given that the process started at state \( x \).

Roughly speaking, a finite Markov chain is a stochastic process on a finite state
space \( \Omega \) that has no memory. Thus, its movement is determined only by its current
state. More precisely:

**Definition 2.1.** A sequence of random variables is a Markov chain with state
space \( \Omega \) if for all \( x,y \in \Omega \), \( t \geq 1 \), and events \( H_{t-1} = \bigcap_{s=0}^{t-1}\{ X_s = x_s \} \) satisfying
\( P\{ H_{t-1} \cup \{ X_t = x \} \} > 0 \), we have
\[
(2.2) \quad P\{ X_{t+1} = y \mid H_{t-1} \cap \{ X_t = x \} \} = P\{ X_{t+1} = y \mid X_t = x \}.
\]

In this paper, we will assume that the state space \( \Omega \) of our Markov chains is
finite. The probability distribution of a random variable can be described by a row
vector where the \( i \)th entry corresponds to the probability of being in the
\( i \)th state.

We can define the transition matrix \( P \) of a Markov chain as an \(|\Omega| \times |\Omega|\) matrix
\( P(x,y) = P\{ X_{t+1} = y \mid X_t = x \} \). Since Markov chains have no memory, the
distribution at time \( t + 1 \) is a function only of the distribution at time \( t \), and thus
a matrix is sufficient to describe the evolution of the Markov chain.

**Definition 2.3.** A stationary distribution for a Markov chain \( P \) is a probability
distribution \( \pi \) such that \( \pi = \pi P \).

We wish to investigate the stationary distributions of Markov chains. Now, we
will describe conditions that ensure that there exists a unique stationary dis-
tribution that is the long-term limiting distribution of the chain. Firstly, any two
states must be able to reach each other. This property is called **irreducibility**.

**Definition 2.4.** A chain \( P \) is called irreducible if for any two states \( x,y \in \Omega \),
there exists a non-negative integer \( t \) such that \( P^t(x,y) > 0 \).

Secondly, we will need our chain to be aperiodic.

**Definition 2.5.** Let \( T(x) = \{ t \geq 1 \mid P^t(x,x) > 0 \} \). Define the period of state \( x \)
as the greatest common divisor of \( T(x) \).

The following proposition allows us to define the period of an irreducible chain:

**Proposition 2.6.** If \( P \) is an irreducible Markov chain, then \( \gcd(T(x)) = \gcd(T(y)) \)
for any \( x,y \in \Omega \).

**Proof.** Let \( x,y \) be any states in \( \Omega \). Since \( P \) is irreducible, there exist non-negative
integers \( r \) and \( l \) such that \( P^r(x,y) > 0 \) and \( P^l(y,x) > 0 \). Define \( m = r + l \) and
observe that \( m \in T(x) \cap T(y) \). Then, for all \( a \in T(x) \), \( a + m \in T(y) \). Since \( \gcd \)
\( T(y) \) divides any element of \( T(y) \), \( \gcd(T(y)) \) divides \( a \). Therefore, \( \gcd(T(x)) \leq \gcd(T(y)) \).
Similarly, \( \gcd(T(y)) \leq \gcd(T(x)) \). Therefore, for any two states \( x \) and \( y \) of an
irreducible chain, \( \gcd(T(x)) = \gcd(T(y)) \).

Therefore, we can define the period of a irreducible chain in the following
manner:
Definition 2.7. For an irreducible Markov chain $P$ on $\Omega$, pick an arbitrary state $x \in \Omega$. Then, the period of $P$ is the period of $x$.

Definition 2.8. A Markov chain is aperiodic if it has period 1.

Now, we will prove that these conditions guarantee the existence of a unique stationary distribution. To do so, we will need some more definitions.

Definition 2.9. A first return time $\tau^+_x$ for the Markov chain $(X_0, X_1, \ldots)$ is defined by

$$\tau^+_x = \min\{t \geq 1 : X_t = x\}.$$  

Now, we will show that the expectation of the first return time for irreducible chains is finite.

Lemma 2.10. For any states $x$ and $y$ of an irreducible chain, $E_x(\tau^+_y) < \infty$.

Proof. Since $P$ is irreducible, there exists an integer $r > 0$ and a real number $\epsilon > 0$ such that for all states $z, w \in \Omega$, there exists a $j \leq r$ with $P^j(s, w) > \epsilon$. Hence, for any value of $X_t$, the probability of hitting state $Y$ between times $t$ and $t + r$ is at least $\epsilon$. Hence, for $k > 0$, we have

$$P_x\{\tau^+_y > kr\} \leq (1 - \epsilon)P_x\{\tau^+_y > (k - 1)r\}.$$  

Repeatedly applying the previous equation yields

$$P_x\{\tau^+_y > kr\} \leq (1 - \epsilon)^k.$$  

When $Y$ is a non-negative integer-valued random variable, we have

$$E(Y) = \sum_{t \geq 0} P\{Y > t\}.$$  

Also, since $P_x\{\tau^+_y > t\}$ is a decreasing function of $t$, (2.11) suffices to bound all terms of the corresponding expression for $E_x(\tau^+_y)$:

$$E_x(\tau^+_y) = \sum_{t \geq 0} P_x\{\tau^+_y > t\} \leq \sum_{k \geq 0} rP_x\{\tau^+_y > kr\} \leq r \sum_{k \geq 0} (1 - \epsilon)^k < \infty.$$  

□

Now, we are ready to state a sufficient condition for the existence of a stationary distribution.

Proposition 2.12. Let $P$ be the transition matrix of an irreducible Markov chain. Then, there exists a stationary distribution $\pi$ such that $\pi(x) > 0$ for all $x \in \Omega$ and $\pi(x) = \frac{1}{E_x(\tau^+_x)}$.

Proof. Let $z \in \Omega$ be an arbitrary state of the Markov chain. We will examine the average time the chain spends in each state between visits to $z$. We define

$$\tilde{\pi}(y) = E_x(\text{number of visits to } y \text{ before returning to } z)$$

$$= \sum_{t=0}^{\infty} P_x\{X_t = y, \tau^+_z > t\}.$$
For any state \( y \), \( \tilde{\pi}(y) \leq \mathbf{E}_z(\tau^+_z) \). Hence, by the previous lemma, for all \( y \in \Omega \), \( \tilde{\pi}(y) < \infty \). Now, we check that \( \tilde{\pi} \) is stationary:

\[
(2.13) \quad \sum_{x \in \Omega} \tilde{\pi}(x)P(x, y) = \sum_{x \in \Omega} \sum_{t=0}^{\infty} \mathbf{P}_z\{X_t = x, \tau^+_x > t\}P(x, y).
\]

Because the event \( \{\tau^+_x \geq t+1\} = \{\tau^+_x > t\} \) is determined by \( X_0, \ldots, X_t \),
\[
\mathbf{P}_z\{X_1 = x, X_{t+1} = y, \tau^+_z \geq t+1\} = \mathbf{P}_z\{X_{t+1} = x, \tau^+_z \geq t + 1\}
= \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau^+_z \geq t\}
= \tilde{\pi}(y) - \mathbf{P}_z\{X_0 = y, \tau^+_z > 0\} + \sum_{t=1}^{\infty} \mathbf{P}_z\{X_t = y, \tau^+_z = t\}.
\]

We now consider two cases.

**Case 1:** \( y = z \): Then, \( X_0 = z \) and \( X_{\tau^+_z} = z \), so the last two terms of are both 1 and cancel each other out.

**Case 2:** \( y \neq z \): Then, the last two terms of are both 0.

Therefore, \( \tilde{\pi} = \tilde{\pi}P \).

We define \( \pi(x) = \frac{\tilde{\pi}(x)}{\mathbf{E}_x(\tau^+_x)} \). Since \( \sum_x \tilde{\pi}(x) = \mathbf{E}_x(\tau^+_x) \), \( \pi \) is a probability measure. Furthermore, \( \pi \) satisfies \( \pi = \pi P \).

In particular, for any \( x \in \Omega \), \( \pi(x) = \frac{1}{\mathbf{E}_x(\tau^+_x)} \).

Now, we want to find sufficient conditions that guarantee that we have a unique stationary distribution. We define harmonic functions as ones that are invariant under left multiplication by \( P \).

**Definition 2.14.** A function \( h : \Omega \to \mathbb{R} \) is harmonic at \( x \) if

\[
h(x) = \sum_{y \in \Omega} P(x, y)h(y).
\]

**Lemma 2.15.** If \( P \) is irreducible, a function \( h \) which is harmonic at every point in \( \Omega \) is constant.

**Proof.** Since \( \Omega \) is finite, there exists a state \( x_0 \) such that \( h(x_0) \) is maximal. Define \( M = h(x_0) \). Suppose for contradiction that \( z \) is a state such that \( P(x_0, z) > 0 \) and \( h(z) < M \). Then,

\[
h(x_0) = P(x_0, z)h(z) + \sum_{y \neq z} P(x_0, y)h(y) < M,
\]

which is a contradiction. Hence, for all states \( z \) such that \( P(x_0, z) > 0 \), \( h(z) = M \).

Fix any \( y \in \Omega \). Since \( P \) is irreducible, there exists a sequence \( x_0, x_1, \ldots, x_n = y \) such that \( P(x_i, x_{i+1}) > 0 \). Repeating the above argument proves \( h(y) = h(x_{n-1}) = \cdots = h(x_0) = M \). Hence, \( h \) is constant.

A direct consequence of this lemma is that irreducible chains have a unique stationary distribution.
Proposition 2.16. If $P$ is the transition matrix of an irreducible Markov chain, then there exists a unique probability distribution $\pi$ such that $\pi = \pi P$.

Proof. We have already shown the existence of such a distribution. By the previous lemma, the kernel of $P - I$ has dimension 1, so the column rank of $P - I$ is $|\Omega| - 1$. Since the row and column ranks of any square matrix are equal, the row vector equation $v = vP$ has a one-dimensional space of solutions. This space contains only one vector whose entries sum to 1. \qed

3. The Convergence Theorem

To proceed, we will need to show that all irreducible, aperiodic Markov chains converge to their stationary distributions. This result is known as the Convergence Theorem. First, we will quantify the distance between two distributions. We define our metric, the total variation distance as follows:

Definition 3.1. The total variation distance between two probability distributions $\mu$ and $\nu$ is the maximum difference between the probabilities that the distributions assign to a single event:

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

Now, we derive an alternate definition of total variation distance.

Proposition 3.2. If $\mu$ and $\nu$ are probability distributions on $\Omega$, then

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

Proof. Define $B = \{x : \mu(x) \geq \nu(x)\}$ and let $A \subseteq \Omega$ be any event. Then, $\mu(A) - \nu(A) \leq \mu(A \cap B)$ because for all $x \in A \cap B^c$, $\mu(x) - \nu(x) < 0$, so the difference in probability cannot decrease when such elements are eliminated. Furthermore,

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

because including more elements of $B$ cannot decrease the difference in probability. Similarly,

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

By subtracting the right-hand sides of (3.3) and (3.4), we can see that they are the same. Furthermore, when we take $A = B$, then $|\mu(A) - \nu(A)|$ is equal to the upper bound. Thus,

$$||\mu - \nu||_{TV} = \frac{1}{2} [\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)]$$

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

\qed

We will need a couple more results to prove the Convergence Theorem. Now, we will prove a useful lemma.

Lemma 3.5. If $S \subseteq \mathbb{N}$ has $\gcd(S) = g_s$, then there exist an integer $m_s$ such that for all $m \geq m_s$, the product $mg_s$ can be written as a linear combination of elements of $S$ with non-negative integer coefficients.
Proof. If $S \subset \mathbb{N}$ is empty, the proof is trivial. If $S \subset \mathbb{N}$ is nonempty, we can define $g^*_S$ as the smallest natural number which is an integer combination of the elements of $S$. Then, $g^*_S$ divides every element of $S$ (if it doesn’t, we get a contradiction when we consider the remainder). Furthermore, since a common divisor of elements divides any integer combination of them, $g_S$ divides $g^*_S$. Hence, $g_S = g^*_S$.

For all subsets $S \subset \mathbb{N}$, there exists a finite subset $F \subset S$ such that $\gcd(S) = \gcd(F)$. Therefore, it suffices to prove the result for finite subsets $F \subset \mathbb{N}$. We proceed by induction on the size of $F$.

**Base Case**: Sets of size 1 trivially satisfy the claim.

We now check sets of size 2. Suppose $F = \{a, b\} \subset \mathbb{N}$ and define $g = \gcd(F)$. Given $m > 0$, there exist integers $c^*$ and $d^*$ such that $mg = c^*a + d^*b$. Since for all $k$, $mg = (c^* + kb)a + (d^* - ka)b$, there exist $c, b$ such that $0 \leq c < b$ and $mg = ca + db$. Furthermore, if $mg > (b-1)a - b$, then $d \geq 0$. Therefore, we can take $m = \frac{ab - a - b}{g} + 1$.

**Inductive hypothesis**: Suppose that for all sets $F$ of size $k$ there exists $m_F$ such that for all $m \geq m_F$, $mg$ can be written as a linear combination of elements of $F$ with non-negative integer coefficients.

**Inductive step**: Let $F'$ be a finite subset of $\mathbb{N}$ with size $k + 1$ and $\gcd(F) = g_{F'}$. Then, $F' = F \cup \{a\}$ for some $a \in \mathbb{N}$ and some subset $F \subset \mathbb{N}$ of size $k$.

Define $g = \gcd(\{a\} \cup F)$. Then, it follows from the definition of gcd that $g = \gcd(a, g_F)$. Suppose that $n$ satisfies $ng \geq m_{\{a,g_F\}}g + m_Fg_F$. By the base case, there exist integers $c, d \geq 0$ such that $ng = m_{\{a,g_F\}}g + m_Fg_F$. Hence, $ng = ca + (d + m_F)g_F$. By the definition of $m_F$, $ca + (d + m_F)g_F = ca + \sum_{f \in F} c_f f$ for some integers $c_f \geq 0$. Therefore, we can define $m_{\{a,g_F\}} = m_{\{a,g_F\}} + \frac{m_Fg_F}{g}$.

\[ \square \]

**Proposition 3.6**: If $P$ is aperiodic and irreducible, then there exists an integer $r$ such that $P^r(x, y) > 0$ for all $x, y \in \Omega$.

**Proof.** For $x \in \Omega$, $T(x) = \{ t \geq 1 : P^t(x, x) > 0 \}$. Since $P$ is aperiodic, gcd $T(x) = 1$. Furthermore, if $s, t \in T(x)$, $P^{s+t}(x) \geq P^{s-r}(x, x)P^r(x, y) > 0$ and hence $s + t \in T(x)$. Therefore, $T(x)$ is closed under addition. By the previous lemma, any set of non-negative integers that is closed under addition and has gcd 1 contains all but finitely many non-negative integers. Thus, there exists a $t(x)$ such that $t \geq t(x)$ implies $t \in T(x)$. Since $P$ is irreducible, we know that for all $y \in \Omega$, there exists $r = r(x, y)$ such that $P^r(x, y) > 0$. Thus, for $t \geq t(x) + r$,

\[ P^t(x, y) \geq P^{t-r}(x, x)P^r(x, y) > 0. \]

Define $t'(x) = t(x) + \max_{y \in \Omega} r(x, y)$. Then, for all $t \geq t'(x)$ and for all $y \in \Omega$, $P^t(x, y) > 0$. Finally, if $t \geq \max_{x \in \Omega} t'(x)$, then $P^t(x, y) > 0$ for all $x, y \in \Omega$. \[ \square \]

A **stochastic matrix** is a matrix that could be the transition matrix of a Markov chain.
Definition 3.7. A matrix $M$ is stochastic all its entries are non-negative and for any $x \in \Omega$, $\sum_{y \in \Omega} P(x, y) = 1$.

Theorem 3.8. (Convergence Theorem) Suppose that $P$ is an irreducible and aperiodic chain with stationary distribution $\pi$. Then, there exist constants $\alpha \in (0, 1)$ and $C > 0$ such that

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

Proof. Since $P$ is irreducible and aperiodic, by 3.6 there exists an $r$ such that $P^r$ has strictly positive entries. Define $P$ as the matrix with $|\Omega|$ rows, each of which is the row vector $\pi$ and $\Pi = \Pi$ (3.9)

It is straightforward to check that $M\Pi = \Pi$ for all stochastic matrices $M$ and $\Pi M = \Pi$ for any matrix $M$ such that $\pi M = \pi$. Next, we use induction to demonstrate that for all $k \geq 1$,

$$P^r k = (1 - \theta^k)\Pi + \theta^k Q^k. \quad (3.10)$$

We have already shown the base case. Suppose (3.10) holds for $k = n$. Then,

$$P^r(n+1) = P^r n P^r = [(1 - \theta^n)\Pi + \theta^n Q^n]P^r. \quad (3.11)$$

Therefore,

$$[(1 - \theta^n)\Pi + \theta^n Q^n]P^r = (1 - \theta^n)\Pi P^r + \theta^n Q^n P^r$$

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

$$= [1 - \theta^n]\Pi + (1 - \theta)\theta^n \Pi + \theta^n Q^n + 1$$

$$= \Pi + \theta \Pi - \theta^{n+1} \Pi + \theta^{n+1} Q^{n+1}$$

$$= (1 - \theta^{n+1})\Pi + \theta^{n+1} Q^{n+1}.$$

By mathematical induction, $P^r k = (1 - \theta^k)\Pi + \theta^k Q^k$ is true for all $k \in \mathbb{N}$. Therefore,

$$P^j P^r k = P^j[(1 - \theta^k)\Pi + \theta^k Q^k]$$

$$\Rightarrow P^j P^r k = (1 - \theta^k)P^j \Pi + \theta^k P^j Q^k$$

$$\Rightarrow P^{j + rk} = \Pi - \theta^k \Pi + \theta^k P^j Q^k$$

$$\Rightarrow P^{j + rk} - \Pi = \theta^k (Q^k P^j - \Pi).$$

Let $x_0$ be an arbitrary state of the chain. Then,

$$P^{j + kr}(x_0) - \pi = \theta^k (Q^k P^j - \Pi)$$

$$\Rightarrow |P^{j + kr}(x_0) - \pi| = \theta^k |Q^k P^j - \Pi|$$

$$\Rightarrow \frac{1}{2}|P^{j + kr}(x_0) - \pi| = \theta^k \frac{1}{2}|Q^k P^j x_0 - \pi|$$

$$\Rightarrow ||P^{k+j}(x_0, \cdot) - \pi||_{TV} = \theta^k ||Q^k P^j(x_0, \cdot) - \pi||_{TV}$$

$$\Rightarrow ||P^{k+j}(x_0, \cdot) - \pi||_{TV} \leq \theta^k$$

(The largest possible total variation distance is 1).
4. APPROXIMATELY WHEN DOES THE TOP-TO-RANDOM SHUFFLE FULLY RANDOMIZE A DECK?

We now investigate the top-to-random method of shuffling a deck of \( n \) cards. The top-to-random shuffle takes the top card of a deck and places it uniformly and randomly in the deck. First, we will check that the top-to-random shuffle will eventually mix up the deck. To do so we will derive and use some more general results.

**Definition 4.1.** Given a group \( G \) and a probability distribution \( \mu \), we define the random walk on \( G \) with increment distribution \( \mu \) as follows: The random walk is a Markov chain with transition matrix \( P \) with entries \( P(g, hg) = \mu(h) \) for all \( g, h \in G \).

We now show that the uniform distribution is stationary for all random walks on finite groups.

**Proposition 4.2.** Let \( P \) be the transition matrix of a random walk on a finite group \( G \) and let \( U \) be the uniform probability distribution on \( G \). Then \( U \) is a stationary distribution for \( P \).

**Proof.** Let \( \mu \) be the increment distribution of the random walk. For any \( g \in G \),

\[
\sum_{h \in G} U(h)P(g, h) = \frac{1}{|G|} \sum_{k \in G} P(k^{-1}g, g) \quad \text{(re-index by setting } k = gh^{-1})
\]

\[= \frac{1}{|G|} \sum_{k \in G} \mu(k) \]

\[= \frac{1}{|G|} \]

\[= U(g).\]

Hence, the uniform distribution is stationary. \( \Box \)

Next, we will show that the top-to-random shuffle is a random walk on the group \( S_n \). Here, the distribution \( \mu(h) \) is defined by \( \mu(h) = \frac{1}{n} \) if \( h \) is a permutation that preserves the relative order of the bottom \( n-1 \) cards and 0 otherwise. Thus, the top-to-random shuffle is a random walk on a group. Therefore, it has uniform stationary distribution. Since the shuffle is irreducible (from any ordering of the deck, any other ordering can eventually be reached) and aperiodic (since if the top card is sent to the top card, the deck’s state does not change), by the Convergence Theorem, the shuffle converges to its uniform stationary distribution. Next, we investigate how long the deck takes to reach the stationary distribution. First, we prove the following initial result.

**Proposition 4.3.** Let \( (X_t) \) be the random walk on \( S_n \) corresponding to the top-to-random shuffle on \( n \) cards. Given at time \( t \) that there are \( k \) cards under the original bottom card, each of the \( k! \) possible orderings of these cards is equally likely.

**Proof.** We proceed by induction on \( t \) starting from \( t = 0 \). Let \( P(t) \): if at time \( t \) there are \( n \) cards under the original bottom card, each of the \( n! \) possible orderings of those cards is equally likely.
Base case: \( t = 0 \): The claim is trivially valid.

Inductive hypothesis: Assume that at time \( t \) there are \( n \) cards under the original bottom card and each of the \( n! \) possible orderings is equally likely.

Inductive step:
Case 1: A card is placed under the original bottom card: Then, since the card is placed uniformly at random under the original bottom card and all time \( t \) orderings of the original bottom cards were equally likely, all \((n+1)!\) orderings of the \( n+1 \) bottom cards are now equally likely.

Case 2: The card is placed on top of the original bottom card: Then, the distribution of orderings of cards under the original bottom card is unchanged.

Now, we are able to characterize the amount of time needed for the top-to-random shuffle to reach the uniform distribution. It will be convenient to name the time where a Markov chain reaches its stationary distribution.

Definition 4.4. Given a sequence \( (X_t)_{t=0}^{\infty} \) of \( \Omega \)-valued random variables, a stopping time \( \tau \) is a \( \{0, 1, 2, \ldots, \infty\} \)-valued random variable such that for each \( t \in \{0, 1, 2, \ldots\} \), there exists a set \( B_t \subset \Omega^{t+1} \) such that \( \{\tau = t\} = \{(X_0, X_1, \ldots, X_t) \in B_t\} \).

Now, we can define the strong stationary time as a time where a chain reaches its stationary distribution.

Definition 4.5. A strong stationary time for a Markov chain \( (X_t) \) with stationary distribution \( \pi \) is a randomized stopping time \( \tau \), possibly depending on the starting position \( x \), such that \( \Pr_x(\tau = t, X_\tau = y) = \Pr_x(\tau = t) \pi(y) \).

We will need one more lemma to find our desired stopping time for the top-to-random shuffle.

Lemma 4.6. (Pólya’s Urn model): Once \( k \) of the top \( n-2 \) cards have placed under the second-to-bottom card via the top-to-random shuffle, the number of cards between the bottom and second-to-bottom card is uniformly distributed on \( \{0, \ldots, k\} \).

Proof. We first analyze the placement of the \( n-2 \) top cards relative to the bottom card. When one of these cards reaches the top, it can either be placed above the second-to-bottom card, between the second-to-bottom and bottom cards, or under the bottom card. If it is placed above the second-to-bottom card, then it must reach the top before the second-to-bottom card reaches the top. Therefore, by the time the second-to-bottom card reaches the top, all the \( n-2 \) top cards have been placed under it. Observe that there is initially only one spot under the bottom card and one spot between the second-to-bottom and bottom cards. Furthermore, given that a card is shuffled under the second-to-bottom card, \( j \) cards are between the second-to-bottom and bottom cards, and \( k-j \) cards are below the bottom card, the probability that the card is placed between the second-to-bottom and bottom cards is \( \frac{1}{k+2} \).
When \( k \) cards are under the second-to-bottom card, define \( B_k \) as the number of cards between the second-to-bottom and bottom cards plus one. Then, \( B_1 \) is 1 and
\[
P\{B_{k+1} = j + 1 \mid B_k = j\} = \frac{j}{k + 2}
\]
and
\[
P\{B_{k+1} = j \mid B_k = j\} = \frac{k + 2 - j}{k + 2}.
\]
We now find an alternate characterization of \( B_k \). Let \( U_0, U_1, \ldots, U_k \) be independent and identically distributed random variables that are each uniformly distributed on the interval \([0, 1]\). Define
\[
L_k = |\{ j \in \{0, 1 \ldots, k\} \mid U_k \leq U_0\}|
\]
as the number of \( U_i \) less than or equal to \( U_0 \). Then, the event \( \{L_k = j, L_{k+1} = j + 1\} \) occurs exactly when \( U_0 \) is the \((j + 1)\)th smallest and \( U_{k+1} \) is one of the \( j + 1 \) smallest among \( \{U_0, U_1, \ldots, U_k\} \). Notice that there are \( j(k!) \) orderings of \( \{U_0, U_1, \ldots, U_{k+1}\} \) making up this event. Since all \((k + 2)!\) orderings are equally likely,
\[
P_x\{L_k = j, L_{k+1} = j + 1\} = \frac{j(k!)}{(k + 2)!} = \frac{j}{(k + 2)(k + 1)}.
\]
Since each relative ordering of \( U_0, \ldots, U_k \) is equally likely, \( P\{L_k = j\} = \frac{1}{k + 1} \).
Together with the previous equation, this implies that
\[
P\{L_{k+1} = j + 1 \mid L_k = j\} = \frac{j}{k + 2}.
\]
Since \( L_{k+1} \in \{j, j + 1\} \), if \( L_k = j \), then
\[
P\{L_{k+1} = j \mid L_k = j\} = \frac{k + 2 - j}{k + 2}.
\]
Notice that \( L_1 \) and \( B_1 \) have the same distribution. By the previous two equations, the sequences \( L_k \) and \( B_k \) have the same transition probabilities. Hence, the sequences \( L_k \) and \( B_k \) have the same distribution. Therefore, \( L_k \) and \( B_k \) have the same distribution. Since the position of \( U_0 \) is uniform among the \( k + 1 \) possible positions, \( L_k \) is uniform on \( \{1, \ldots, k + 1\} \). Hence, \( B_k \) is uniform on \( \{1, \ldots, k + 1\} \). □

**Theorem 4.7.** The time \( \tau \) at which the card initially one card from the bottom rises to the top, plus one more shuffle, is a strong stationary time.

**Proof.** By the previous proposition, when the card initially one card from the bottom rises to the top, the relative order of the cards that started more than one card from the bottom is uniform. Furthermore, shuffling once preserves the relative uniform ordering of these cards.

By the previous lemma, when the original second-to-bottom card reaches the top, the number of these cards above the bottom card is uniformly distributed on \( \{0, \ldots, n - 2\} \). Therefore, at this time, the position of the original bottom card (with 1 denoting the top position and \( n \) denoting the bottom position) is uniform on \( \{2, \ldots, n\} \). There are two ways the original bottom card can reach position \( j \) from the top in one shuffle: either it starts in position \( j \) and the original second-to-bottom card is shuffled on top of it or it starts in position \( j + 1 \) and the original second-to-bottom card is shuffled under it.
Case 1: \( j = 1 \): In this case, the original bottom card must have been in position 2 when the original second-to-bottom card reaches the top. Then, there is a \( \frac{n-1}{n} \) probability that the final shuffle places the original second-to-bottom card in the second position or lower. Let \( b \) denote the position of the original bottom card. Then,
\[
P\{b = j\} = \frac{1}{n-1} \cdot \frac{n-1}{n} = \frac{1}{n}.
\]

Case 2: \( 2 \leq j \leq n-1 \): In this case, the original bottom card either started in position \( j \) and the final shuffle placed a card on top of it, or it started in position \( j + 1 \) and the final shuffle placed a card under it. There is a \( \frac{j}{n} \) chance of a shuffle placing a card on top of a card in position \( j \) and a \( \frac{n-(j+1)}{n} \) chance of a shuffle placing a card below one in position \( j + 1 \).
Hence,
\[
P\{b = j\} = \frac{1}{n-1} \cdot \frac{j}{n} + \frac{1}{n-1} \cdot \frac{n-j-1}{n} = \frac{j}{(n-1)n} + \frac{n-j-1}{(n-1)n} = \frac{n-1}{(n-1)n} = \frac{1}{n}.
\]

Case 3: \( j = n \): In this case, the original bottom card must have started in position \( n \) and the final shuffle must have placed a card on top of it. Then,
\[
P\{b = j\} = \frac{1}{n-1} \cdot \frac{n-1}{n} = \frac{1}{n}.
\]

Therefore, one shuffle after the original second-to-bottom card reaches the top, the distribution of the original bottom card is uniform. Since the other cards have a uniformly distributed relative order, it follows that the distribution of deck permutations is uniform. Hence, the time at which the card initially one card from the bottom rises to the top, plus one more shuffle, is a strong stationary time. \( \square \)

**Theorem 4.8.** Let \( \tau \) be the strong stationary time in the last theorem. Then,
\[
E(\tau) = n \sum_{k=2}^{n-1} \frac{1}{k}.
\]

**Proof.** Note: Credits to Proposition 2.3 of [1] for the idea behind this proof.

Case 1: \( n = 2 \): The second-to-bottom card starts on top, so \( \tau = 1 = 2 \sum_{k=2}^{2} \frac{1}{k} \).

Case 2: \( n \geq 3 \): We can compute the \( E(\tau) \) by writing \( \tau \) as the sum of geometric random variables. Define \( \tau_k \) to be the time the second-to-bottom card reaches the \( k^{th} \) position from the top. Since \( \tau = \tau_1 + 1 \) and the second-to-bottom card starts in position \( n-1 \),
\[
\tau = \tau_1 + 1 = 1 + (\tau_1 - \tau_2) + (\tau_2 - \tau_3) + \cdots + (\tau_{n-2} - \tau_{n-1}).
\]
Note that each shuffle can place the card in \( n \) possible positions and the number of ways to shuffle a card under one in position \( k \) is \( n-k+1 \). Therefore, when the second-to-bottom card is in the \( k^{th} \) position for \( k \geq 2 \), the next shuffle has a \( \frac{n-k+1}{n} \) probability of shuffling under the original second-to-bottom and hence moving it up to the \( k-1^{th} \) position. Each
card has the same \(\frac{n-k+1}{n}\) probability of being shuffled under the second-to-bottom card until the second-to-bottom card is pushed up one position. Therefore, \(\tau_{k-1} - \tau_k\) is a geometric random variable with success probability \(\frac{n-k+1}{n}\) and expected value \(\frac{n}{n-k+1}\). Hence,

\[
E(\tau) = 1 + E(\tau_1 - \tau_2) + E(\tau_2 - \tau_3) + \cdots + E(\tau_{n-2} - \tau_{n-1})
\]

\[
= 1 + \sum_{k=2}^{n-1} E(\tau_{k-1} - \tau_k)
\]

\[
= 1 + \sum_{k=2}^{n-1} \frac{n}{n-k+1}
\]

\[
= n \cdot \frac{1}{n} + n \sum_{k=2}^{n-1} \frac{1}{k}
\]

\[
= n \sum_{k=2}^{n} \frac{1}{k}.
\]

\[\square\]

We now introduce the coupon collector distribution. Suppose there are \(n\) types of coupons and a collector is trying to collect them all. Suppose that each coupon the collector finds is independent of all other coupons he finds and the probability that a coupon is any given type is \(\frac{1}{n}\). Define the coupon collector random variable \(c\) as the number of coupons the collector has when he first reaches a full collection of \(n\) coupons. By comparison with the coupon collector distribution, we can obtain an upper bound on \(\tau\).

**Lemma 4.9.** For all \(t\), \(P(c > t) \geq P(\tau > t)\).

**Proof.** We can also write \(c\) as a sum of geometric random variables. Let \(c_k\) denote the number of coupons that the collector has when he first has \(k\) distinct types of coupons. Then,

\[
c = c_n
\]

\[
= (c_n - c_{n-1}) + (c_{n-1} - c_{n-2}) + \cdots + (c_2 - c_1) + c_1
\]

\[
= (c_n - c_{n-1}) + (c_{n-1} - c_{n-2}) + \cdots + (c_2 - c_1) + 1.
\]

If the collector has \(m\) coupon types, the probability of reaching \(m+1\) coupon types with the next coupon is \(\frac{n-m}{n}\). Therefore, \(c_{m+1} - c_m\) is a geometric random random variable with expected value \(\frac{n}{n-m}\).

Define \(g(p)\) as the geometric random variable with success probability \(p\). Then,

\[
c = g\left(\frac{1}{n}\right) + \cdots + g\left(\frac{n}{n}\right)
\]

\[
= g\left(\frac{1}{n}\right) + \sum_{k=2}^{n} g\left(\frac{k}{n}\right).
\]
Furthermore,
\[ \tau = 1 + (\tau_1 - \tau_2) + (\tau_2 - \tau_3) + \cdots + (\tau_{n-2} - \tau_{n-1}) \]
\[ = g\left(\frac{n}{n}\right) + g\left(\frac{n-1}{n}\right) + g\left(\frac{n-2}{n}\right) + \cdots + g\left(\frac{2}{n}\right) \]
\[ = \sum_{k=2}^{n} g\left(\frac{k}{n}\right). \]
Hence, for all \( t \), \( P(c > t) \geq P(\tau > t) \). \( \square \)

Now, we bound the upper tail of \( c \).

**Proposition 4.10.** Let \( c \) be a coupon collector random variable. For any \( \alpha > 0 \),
\[ P\{\tau > [n \log n + \alpha n]\} \leq e^{-\alpha}. \]

**Proof.** Define \( A_i \) to be the event that the \( i \)th coupon type does not appear among the first \( [n \log n + cn] \) coupons drawn. Then,
\[ P\{c > [n \log n + \alpha n]\} = P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i). \]
Since each trial has probability \( 1 - \frac{1}{n} \) of not drawing coupon \( i \) and trials are independent, the right-hand side of the above expression is bounded above by
\[ \sum_{i=1}^{n} \left(1 - \frac{1}{n}\right)^{\lfloor n \log n + \alpha n\rfloor} \leq n \exp\left(-\frac{n \log n + \alpha n}{n}\right) = e^{-\alpha}, \]
and hence \( P\{c > [n \log n + \alpha n]\} \leq e^{-\alpha} \). \( \square \)

**Theorem 4.11.** For all \( \alpha > 0 \),
\[ (4.12) \quad P\{\tau > [n \log n + \alpha n]\} \leq e^{-\alpha}. \]

**Proof.** This is immediate from the preceding two results. \( \square \)

**Remark** Since \( E(\tau) \) is approximately \( n \log n \), we have shown that \( \tau \) is unlikely to be much larger than its expected value.

5. **So when are we sure that we’ve shuffled the deck enough?**

We have now characterized the random time it takes for the deck to be perfectly randomized with the top-to-random shuffle. If we wanted to shuffle a deck, we would want to know the fixed amount of shuffles it would take to make the deck close to random. Fortunately, we can use our knowledge of the shuffle’s strong stationary time to answer this question by bounding the mixing time from above. First, we will need to define and use a convenient alternate method of describing a chain.

**Definition 5.1.** A random mapping representation of a transition matrix \( P \) on state space \( \Omega \) is a function \( f : \Omega \times \Lambda \to \Omega \), along with a \( \Lambda \)-valued random variable \( Z \), such that
\[ P\{f(x, Z) = y\} = P(x, y). \]

**Proposition 5.2.** Every transition matrix on a finite state space has a random mapping representation.
Proof. Let \( P \) be the transition matrix of a Markov chain with state space \( \Omega = \{x_1, \ldots, x_n\} \). Take \( \Lambda = [0, 1] \) and choose auxiliary random variables \( Z, Z_1, Z_2, \ldots \) uniformly from this interval. Define \( F_{j,k} = \sum_{i=1}^{k} P(x_j, x_i) \) and define
\[
f(x_j, z) = x_k \quad \text{when} \quad F_{j,k-1} < z \leq F_{j,k}.
\]
Then,
\[
\mathbb{P}\{f(x_j, Z) = x_k\} = \mathbb{P}\{F_{j,k-1} < Z \leq F_{j,k}\} = P(x_j, x_i).
\]

Lemma 5.3. Let \((X_t)\) be an irreducible Markov chain with stationary distribution \( \pi \). If \( \tau \) is a strong stationary time for \((X_t)\), then for all \( t \geq 0 \),
\[
\mathbb{P}_x\{\tau \leq t, X_t = y\} = \mathbb{P}\{\tau \leq t\} \pi(y).
\]

Proof. Let \( Z_1, Z_2, \ldots \) be the i.i.d sequence used in the random mapping representation of \((X_t)\). For any \( s \leq t \),
\[
\mathbb{P}_x\{\tau = s, X_t = y\} = \sum_{z \in \Omega} \mathbb{P}_x\{X_t = y \mid \tau = s, X_s = z\} \mathbb{P}_x\{\tau = s, X_s = z\}.
\]
Since \( \tau \) is a stopping time for \((Z_t)\), there exists a set \( B \in \Omega^t \) such that such that the event \( \{\tau = s\} \) equals \( \{(Z_1, \ldots, Z_s) \in B\} \). Furthermore, for nonnegative integers \( r \) and \( s \), there exists a function \( f_r : \Omega^{r+1} \to \Omega \) such that
\[
X_{s+r} = f_r(X_s, Z_{s+1}, \ldots, Z_{s+r}).
\]
Since the random vectors \((Z_1, \ldots, Z_s)\) and \((Z_{s+1}, \ldots, Z_t)\) are independent,
\[
\mathbb{P}_x\{X_t = y \mid \tau = s, X_s = z\} = \mathbb{P}_x\{f_{t-s}(z, Z_{s+1}, \ldots, Z_t) = y \mid (X_1, \ldots, X_s) \in B, X_s = z\} = P^{t-s}(z, y).
\]
By the definition of strong stationary time and the above equality, (5.5) can be rewritten as
\[
\mathbb{P}_x\{\tau = s, X_t = y\} = \sum_{z \in \Omega} P^{t-s} \pi(z) \mathbb{P}_x\{\tau = s\}.
\]
Since \( \pi = \pi P^{t-s} \),
\[
\sum_{z \in \Omega} P^{t-s} \pi(z) \mathbb{P}_x\{\tau = s\} = \pi(y) \mathbb{P}_x\{\tau = s\}.
\]
Summing over all \( s \leq t \) establishes \( \mathbb{P}_x\{\tau \leq t, X_t = y\} = \mathbb{P}\{\tau \leq t\} \pi(y) \), our desired result. \( \square \)

We now introduce a metric of how isolated a point is in a Markov chain.

Definition 5.6. The separation distance \( s_x(t) \) is defined by
\[
s_x(t) = \max_{y \in \Omega} \mathbb{P}_x\left\{1 - \frac{P^t(x, y)}{\pi(y)}\right\}.
\]
Next, we quantify how disconnected the states of a given Markov chain are.

Definition 5.7. We define \( s(t) = \max_{x \in \Omega} s_x(t) \).

To establish the next proposition, we will need two more lemmas.

Lemma 5.8. If \( \tau \) is a strong stationary time, then \( s_x(t) \leq \mathbb{P}_x\{\tau > t\} \).
Proof. Fix any $x \in \Omega$. For any $y \in \Omega$,
\[
1 - \frac{P^t(x, y)}{\pi(y)} = 1 - \frac{P_x\{X_t = y\}}{\pi(y)} \\
\leq 1 - \frac{P_x\{X_t = y, \tau \leq t\}}{\pi(y)} \\
= 1 - \frac{\pi(y)P_x\{\tau \leq t\}}{\pi(y)} = 1 - \pi(y)P_x\{\tau \leq t\}.
\]
Hence, $s_x(t) \leq P_x\{\tau > t\}$. □

We now define precise notion of what “close to stationary” means and the amount of time it takes a Markov chain to approach its equilibrium distribution. The distance from stationary quantifies how close to stationary a Markov chain must be after it has been mixing for time $t$.

**Definition 5.9.** The distance from stationary $d(t)$ of a time $t$ is defined by
\[
d(t) = \max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV}.
\]

**Definition 5.10.** The mixing time $t_{mix}(\epsilon)$ is defined by
\[
t_{mix}(\epsilon) = \min\{t : d(t) \leq \epsilon\}.
\]

**Lemma 5.11.** The separation distance $s_x(t)$ satisfies $\|P^t(x, \cdot) - \pi\|_{TV} \leq s_x(t)$ and hence $d(t) \leq s(t)$.

Proof. Observe that
\[
\|P^t(x, \cdot)\|_{TV} = \sum_{y \in \Omega, P^t(x, y) < \pi(y)} \left[\pi(y) - P^t(x, y)\right] \\
= \sum_{y \in \Omega, P^t(x, y) < \pi(y)} \pi(y) \left[1 - \frac{P^t(x, y)}{\pi(y)}\right] \\
\leq \max_y \left[1 - \frac{P^t(x, y)}{\pi(y)}\right] = s_x(t).
\]
□

The required proposition follows from these lemmas.

**Proposition 5.12.** If $\tau$ is a strong stationary time, then
\[
d(t) = \max_{x \in \Omega} \|P^t(x, \cdot)\|_{TV} \leq \max_{x \in \Omega} P_x\{\tau > t\}.
\]

Proof. By the definition of $d(t)$, $d(t) = \max_{x \in \Omega} \|P^t(x, \cdot)\|_{TV}$. By Lemma 6.13, $d(t) \leq s(t) = \max_{x \in \Omega} s_x(t)$. By Lemma 6.11, $s_x(t) \leq P_x\{\tau > t\}$. Hence, $\max_{x \in \Omega} \|P^t(x, \cdot)\|_{TV} = d(t) \leq \max_{x \in \Omega} P_x\{\tau > t\}$. □

Now, we can identify a sufficient number of shuffles for a desired degree of randomness.
Theorem 5.14. For the top-to-random shuffle, for all \( n \), \( t_{\text{mix}}(\epsilon) \leq n \log n + n \log \epsilon^{-1} \).

Proof. Notice that
\[
d([n \log n + \alpha n]) \leq \max_{x \in \Omega} P_x \{ \tau > [n \log n + \alpha n] \} \tag{5.12}
\]
Therefore,
\[
d([n \log n + n \log \epsilon^{-1}]) \leq e^{-\alpha} \tag{4.12}.
\]

Thus, \( t_{\text{mix}}(\epsilon) \leq n \log n + n \log \epsilon^{-1} \).

It is common to use \( t_{\text{mix}}(\frac{1}{4}) \) as a notion of how long it takes to reach an approximately random distribution. For a 52-card deck and the top-to-random shuffle, the bound we just derived yields \( t_{\text{mix}}(\frac{1}{4}) \leq 277 \). Since that is a large number of shuffles, we will check whether this shuffling process indeed takes this long.

6. Are we sure it takes long?

So far, we have shown that it takes at most approximately \( n \log n \) shuffles to mix up the deck. In this section, we show that the time required to randomize the deck, up to a constant, is at least \( n \log n \) for large decks. The following is a useful inequality that we will need for the next theorem.

Lemma 6.1. (Chebyshev’s Inequality) For any real number \( k > 0 \) and random variable \( X \) with finite expectation \( \mu \) and finite variance \( \sigma^2 \),
\[
P\{|X - \mu| \geq k\sigma\} \leq \frac{1}{k^2}.
\]

Proof. Let \( A \) be any event. Define the indicator variable \( I_A \) as \( I_A = 1 \) if \( A \) occurs and \( I_A = 0 \) otherwise. Then,
\[
P\{|X - \mu| \geq k\sigma\} = E(I_{|X - \mu| \geq k\sigma}) = E(I_{|X - \mu|/k\sigma \geq 1}) \leq E\left(\frac{(X - \mu)^2}{k\sigma}\right) = \frac{1}{k^2} E\left(\frac{(X - \mu)^2}{\sigma^2}\right) = \frac{1}{k^2}.
\]

\( \square \)

Theorem 6.2. Let \((X_t)\) be the top-to-random shuffle on \( n \) cards. For any \( \epsilon > 0 \), there exists a constant \( \alpha_0 \) such that for all \( \alpha > \alpha_0 \) and sufficiently large \( n \),
\[
d(n \log n - \alpha n) \geq 1 - \epsilon.
\]
Furthermore, there exists a constant \( \alpha_1 \) such that for all sufficiently large \( n \),
\[
t_{\text{mix}} \geq n \log n - \alpha_1 n.
\]
Proof. Define the event $A_j$ as

$$A_j = \{\text{The original bottom } j \text{ cards are in their original relative order}\}.$$ 

Let $\text{id}$ denote the identity permutation. We will bound $||P^t(\text{id}, \cdot) - \pi||_{\text{TV}}$ from below. Let $\tau_j$ be the time required for the card initially $j$th from the bottom to reach the top and let $\tau_{j,i}$ be the time it takes for the card initially $j$th from the bottom to ascend from position $i$ from the bottom to position $i+1$. Then,

$$\tau_j = \sum_{i=j}^{n-1} \tau_{j,i}.$$ 

The variables $\{\tau_{j,i}\}_{i=j}^{n-1}$ are independent and $\tau_{j,i}$ has a geometric distribution with parameter $p = \frac{i}{n}$. Hence, $E(\tau_{j,i}) = \frac{n}{i}$ and $\text{Var}(\tau_{j,i}) < \frac{n^2}{i^2}$. Thus,

$$E(\tau_j) = \sum_{i=j}^{n-1} \frac{n}{i} \geq n \left(\log n - \log j - 1\right)$$

and

$$\text{Var}(\tau_j) \leq n^2 \sum_{i=j}^{\infty} \frac{1}{i(i-1)} \leq \frac{n^2}{j-1}.$$ 

Suppose $\alpha \geq \log j + 2$. Then,

$$P\{\tau_j < n \log n - \alpha n\} \leq P\{\tau_j - E(\tau_j) < n \log n - \alpha n - n(\log n - \log j - 1)\}$$

(6.3)

$$= P\{\tau_j - E(\tau_j) < n \log n - \alpha n - n \log n + n \log j + n\}$$

$$= P\{\tau_j - E(\tau_j) < \alpha n + n \log j + n\}$$

$$= P\{\tau_j - E(\tau_j) < -n(\alpha - \log j - 1)\}$$

$$\leq P\{|\tau_j - E(\tau_j)| < n(\alpha - \log j - 1)\}$$

$$\leq \frac{\text{Var } \tau_j}{(n(\alpha - \log j - 1))^2}$$

(6.4)

$$\leq \frac{n^2}{j-1}$$

$$= \frac{1}{(j-1)(\alpha - \log j - 1)^2}$$

$$\leq \frac{1}{j-1}.$$ 

Define $t_n(\alpha) = n \log n - \alpha n$. If $\tau_j \geq t_n(\alpha)$, then the original $j$ cards remain in their original relative order at time $t_n(\alpha)$, so for $\alpha > \log j + 2$,

$$P^{t_n(\alpha)}(\text{id}, A_j) \geq P\{\tau_j \geq t_n(\alpha)\}$$

$$\geq 1 - \frac{1}{j-1}.$$ 

On the other hand, for the uniform stationary distribution,

$$\pi(A_j) = \frac{1}{j!} \leq \frac{1}{j-1}.$$
Hence, for $\alpha \geq \log j + 2$,
\[
d(t_n(\alpha)) \geq \|P^{t_n(\alpha)}(\text{id}, \cdot) - \pi\|_{TV}
\geq P^{t_n(\alpha)}(\text{id}, A_j) - \pi(A_j)
\geq 1 - \frac{2}{j - 1}.
\]

Suppose $n \geq e^{\alpha - 2}$. Then, define $j = e^{\alpha - 2}$. Define
\[
g(\alpha) = 1 - \frac{1}{e^{\alpha - 2} - 1}.
\]

Since $d(t_n(\alpha)) > g(\alpha)$,
\[
\liminf_{n \to \infty} d(t_n(\alpha)) \geq g(\alpha),
\]
and as $\alpha \to \infty$, $g(\alpha) \to 1$. 

**Acknowledgments.** It is a pleasure to thank my mentor, Bowei Zheng, for answering my questions about Markov chains and helping me with LaTeX. Credits to [1] for much of the material in this paper. Credits to [2] for the proof of Chebyshev’s inequality. Theorem 4.7, Theorem 4.8, Lemma 4.9, and Theorem 4.11 are original mathematics.

**References**
