MIXING TIMES AND APPLICATIONS OF FINITE MARKOV CHAINS

JHON LENIN ACUNA AREVALO

Abstract. This work explores the concept of mixing times for finite Markov chains and some specific examples. After introducing the theoretical framework, including definitions, key theorems, and proofs, we present concrete examples to illustrate the convergence to a stationary distribution. In particular, we analyze the mixing time of top-to-random shuffling and simple random walks on graphs, providing both theoretical derivations and computational simulations. The purpose of this study is to demonstrate how mixing time analysis offers a practical framework for understanding the speed of convergence in stochastic systems, bridging mathematical theory with computational experimentation.

Contents

1. Introduction

The study of Markov chains and their mixing behavior has its origins in the early 20th century, when Andrey Markov introduced stochastic processes to generalize independence in probability theory [3]. Since then, Markov chains have become a cornerstone of probability and combinatorics.

The concept of *mixing time*, which quantifies the speed at which a Markov chain approaches its stationary distribution, has attracted considerable attention because it bridges abstract mathematical theory with practical applications. Understanding how quickly a system "forgets" its initial state is not only of theoretical interest but also essential in areas such as randomized algorithms, Markov Chain Monte Carlo methods, and statistical mechanics.

Organization of the paper. In Section 2, we lay down the mathematical background, including measure-theoretic foundations, total variation distance, and the basic theory of Markov chains. Section 3 introduces the definition of mixing time and its fundamental properties. In Section 4, we discuss techniques for estimating mixing time, including strong stationary times and Monte Carlo methods. Section 5 presents a detailed analysis of the top-to-random shuffle, both from a theoretical and computational perspective.

2. Mathematical Background

2.1. **Measure-Theoretic Foundations.** To rigorously define probability distributions and concepts like total variation distance, we begin with the language of measure theory. More specifically, we begin with the so-called σ-algebra, which will give us the framework to define probability with all the well-known properties we are used to.

The definitions and results presented in this section follow the exposition of Lawler [1], with emphasis on the finite state space setting relevant to this work.

Definition 2.1 (σ-algebra)**.** Let Ω be a non-empty set. A collection $\mathcal{F} \subseteq 2^{\Omega}$, where 2^{Ω} is the collection of all subsets of Ω, is called a **σ-algebra** if it satisfies:

(1) $\Omega \in \mathcal{F}$,
(2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,
(3) $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

That is, $\mathcal{F}$ is closed under complementation and countable unions.

Definition 2.2 (Measurable Space)**.** A pair $(\Omega, \mathcal{F})$, where Ω is a set and $\mathcal{F}$ is a σ-algebra on Ω, is called a **measurable space**.

Definition 2.3 (Probability Measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a **probability measure** if:

(1) $\mathbb{P}(\Omega) = 1$,
(2) $\mathbb{P}$ is countably additive: for any countable collection $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ of pairwise disjoint sets,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Definition 2.4 (Probability Space)**.** A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where:

- Ω is the sample space (set of all outcomes),
- $\mathcal{F}$ is a σ-algebra of measurable events,
- $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.

In the setting of finite sample spaces, we can construct a foundational example of a probability space using the basic principles of discrete probability. Let us consider a finite set

$$\Omega = \{1, 2, \ldots, n\},$$

which we interpret as our sample space—the collection of all possible outcomes of a given random experiment. Associated with this sample space is the *power set* $\mathcal{F} = 2^{\Omega}$. This collection serves as a σ-algebra, meaning that it is closed under countable unions, intersections, and complements. Since Ω is finite, the distinction between countable and finite is trivial here, and the power set itself automatically satisfies the requirements of a σ-algebra.

To assign probabilities, we define a *probability mass function* (pmf) $p : \Omega \to [0, 1]$, which assigns a non-negative number to each outcome in Ω, such that the total probability over all outcomes sums to 1:

$$\sum_{i=1}^{n} p(i) = 1.$$

Using this function, we define the probability of any event $A \subseteq \Omega$ as:

$$\mathbb{P}(A) = \sum_{i \in A} p(i).$$

This definition satisfies the axioms of probability: non-negativity, normalization (i.e., $\mathbb{P}(\Omega) = 1$), and finite additivity (i.e., if A and B are disjoint events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$). Thus, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ forms a valid probability space.

Let us now illustrate this construction with some examples.

Example 1: A Fair Die. Consider the experiment of rolling a fair six-sided die. The sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

and since the die is fair, each outcome is equally likely:

$$p(i) = \frac{1}{6}, \quad \text{for } i = 1, 2, \ldots, 6.$$

If $A = \{2, 4, 6\}$ is the event that the die shows an even number, then

$$\mathbb{P}(A) = \mathbb{P}(\{2, 4, 6\}) = p(2) + p(4) + p(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Example 2: A Biased Coin. Suppose we have a biased coin for which the probability of heads is 0.7 and that of tails is 0.3. Then, the sample space is

$$\Omega = \{\mathrm{H}, \mathrm{T}\}, \quad p(\mathrm{H}) = 0.7, \quad p(\mathrm{T}) = 0.3.$$

If $A = \{\mathrm{H}\}$, then $\mathbb{P}(A) = p(\mathrm{H}) = 0.7$.

General Observations. In both of these examples, the discrete sample space allowed us to assign probabilities directly to individual outcomes and compute event probabilities via summation. In fact, in both finite and countably infinite sample spaces, it is common to use what is called the *discrete σ-algebra*. This is nothing more than the powerset of Ω, denoted 2^{Ω}, meaning that every subset of Ω is considered a measurable event. The name "discrete" reflects that we are distinguishing every single element of Ω.

In this setting, the probability measure $\mathbb{P}$ is completely characterized by its action on singleton sets:

$$\mathbb{P}(\{x\}) = p(x), \quad \text{for each } x \in \Omega,$$

and extended linearly to arbitrary subsets of Ω. In other words, once the probabilities of individual outcomes are specified, the probability of any event $A \subseteq \Omega$ is given by

$$\mathbb{P}(A) = \sum_{x \in A} p(x).$$

Example 3: Rolling Two Dice. Let us now consider a slightly more complex example. Suppose two fair six-sided dice are rolled. The sample space consists of ordered pairs:

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\},$$

so that $|\Omega| = 36$. Each outcome is equally likely with probability $p(i, j) = \frac{1}{36}$. Define the event

$$A = \{(i, j) \in \Omega : i + j = 7\},$$

i.e., the event that the sum of the two dice is 7. There are six such outcomes:

$$(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1).$$

Hence,

$$\mathbb{P}(A) = 6 \times \frac{1}{36} = \frac{1}{6}.$$

Such examples highlight the ease with which probability theory on finite sample spaces allows explicit computation. This construction of probability measures on finite spaces becomes the stepping stone towards the study of random variables, expectation, and convergence—core topics of both theoretical and applied probability.

2.2. **Total Variation Distance.** Given two probability distributions on the same sample space, how do we compare them? How close are they to each other? To measure this notion of closeness, we often rely on a metric known as the **total variation distance**. In the context of Markov chains, this metric plays a key role, as it provides the means to begin quantifying the number of steps required for the chain to converge to a certain limiting distribution—namely, the stationary distribution, which we will define soon.

Definition 2.5 (Total Variation Distance)**.** Let μ and ν be two probability measures on a common measurable space $(\Omega, \mathcal{F})$. The **total variation distance** between μ and ν is defined as:

$$\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

This definition captures the maximum discrepancy between the probabilities assigned by μ and ν to any measurable event $A \subseteq \Omega$.

Theorem 2.6 (Total Variation Distance for Finite or Countable Spaces [3])**.** *Suppose Ω is a finite or countable set. In this case, every probability measure on Ω is determined by a probability mass function (pmf). Then for two probability measures μ and ν with pmfs $\mu(x)$ and $\nu(x)$, we have*

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

Proof. Let $A^+ := \{x \in \Omega : \mu(x) \geq \nu(x)\}$ and $A^- := \Omega \setminus A^+$. Then

$$\sup_{A \subseteq \Omega} |\mu(A) - \nu(A)| = \mu(A^+) - \nu(A^+) = \sum_{x \in A^+} [\mu(x) - \nu(x)].$$

Note that for $x \in A^+$ the terms $\mu(x) - \nu(x)$ are nonnegative, while for $x \in A^-$ they are nonpositive. Since

$$\sum_{x \in \Omega} (\mu(x) - \nu(x)) = 0$$

because both distributions sum up to 1, it follows that the positive and negative parts of the sum must have the same absolute value:

$$\sum_{x \in A^+} [\mu(x) - \nu(x)] = - \sum_{x \in A^-} [\mu(x) - \nu(x)].$$

Therefore,

$$\sum_{x \in A^+} [\mu(x) - \nu(x)] = \tfrac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

□

Example 2.7. Let $\Omega = \{1, 2, 3\}$, and define two probability distributions

$$\mu = (0.5, 0.3, 0.2), \qquad \nu = \left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right).$$

Method 1.

$$\|\mu - \nu\|_{\mathrm{TV}} = \tfrac{1}{2}\Big(\big|0.5 - \tfrac{1}{3}\big| + \big|0.3 - \tfrac{1}{3}\big| + \big|0.2 - \tfrac{1}{3}\big|\Big).$$

Compute each term:

$$|0.5 - \tfrac{1}{3}| = \tfrac{1}{6}, \quad |0.3 - \tfrac{1}{3}| = \tfrac{1}{30}, \quad |0.2 - \tfrac{1}{3}| = \tfrac{2}{15}.$$

Adding:

$$\tfrac{1}{6} + \tfrac{1}{30} + \tfrac{2}{15} = \tfrac{10}{30} = \tfrac{1}{3}.$$

So

$$\|\mu - \nu\|_{\mathrm{TV}} = \tfrac{1}{2} \cdot \tfrac{1}{3} = \tfrac{1}{6} \approx 0.1667.$$

Method 2. Take

$$A^+ = \{x \in \Omega : \mu(x) \geq \nu(x)\} = \{1, 2\}.$$

Then

$$\mu(A^+) - \nu(A^+) = \big(0.5 + 0.3\big) - \left(\tfrac{1}{3} + \tfrac{1}{3}\right) = 0.8 - \tfrac{2}{3} = \tfrac{1}{6}.$$

No other subset yields a larger difference, so

$$\|\mu - \nu\|_{\mathrm{TV}} = \tfrac{1}{6} \approx 0.1667.$$

Both methods agree.

Remark 2.8. The total variation distance satisfies $0 \leq \|\mu - \nu\|_{\mathrm{TV}} \leq 1$, with:

- $\|\mu - \nu\|_{\mathrm{TV}} = 0 \iff \mu = \nu$,
- $\|\mu - \nu\|_{\mathrm{TV}} = 1 \iff \mu \perp \nu$.

The total variation distance can be interpreted as the *maximum probability of distinguishing* between two distributions in a single experiment.

2.3. **Markov Chains.** Markov chains constitute a fundamental and versatile class of stochastic processes used to model systems that evolve in a probabilistic manner over time. What distinguishes them is the *Markov property*: the future state depends only on the present state, not on the sequence of past states. Broadly speaking, a stochastic process is a collection of random variables indexed by time, representing the evolution of a system whose future behavior is subject to inherent randomness. In this section, we introduce the formal definition of discrete-time Markov chains, describe their transition dynamics, investigate their long-term behavior, and examine key structural properties that govern their evolution.

Definition 2.9 (Discrete-Time Markov Chain)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$, taking values in a countable state space $\mathcal{X} \subseteq \Omega$, is a **discrete-time Markov chain** if for all $t \in \mathbb{N}$ and all states $x_0, x_1, \ldots, x_t, x_{t+1} \in \mathcal{X}$, we have:

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t).$$

This is known as the **Markov property**. In simple words, it states that the value of the chain in the next step depends solely on the current state and not on the whole past. The way in which we achieved this is by using a probability matrix, aka **transition matrix**, in which each state has a given non-negative probability of going to all the others, including staying in the current one.

Definition 2.10 (Transition Matrix)**.** For a Markov chain $\{X_t\}$ on a finite or countable state space $\mathcal{X}$, the **transition matrix** $P = (P(x, y))_{x,y\in\mathcal{X}}$ is defined by:

$$P(x, y) := \mathbb{P}(X_{t+1} = y \mid X_t = x), \quad \text{for all } x, y \in \mathcal{X}.$$

Each row $P(x, \cdot)$ defines a probability distribution on $\mathcal{X}$, i.e.,

$$\sum_{y\in\mathcal{X}} P(x, y) = 1, \quad \forall x \in \mathcal{X}.$$

Definition 2.11 (Distribution at Time t)**.** Given an initial distribution μ_0 on $\mathcal{X}$, the distribution of the chain at time t, denoted μ_t, is given by:

$$\mu_t(y) = \mathbb{P}(X_t = y) = (\mu_0 P^t)(y) = \sum_{x\in\mathcal{X}} \mu_0(x) P^t(x, y),$$

where μ_0 is the initial distribution at time 0 and the matrix P^t represents the probability of transitioning from one state to another in t steps. What becomes important from now on is the idea of a **stationary distribution**, a probability distribution attained eventually by the chain which from then on remains the same no matter how many more steps we take (notice that each step corresponds to a right multiplication by the matrix P). As we shall see shortly, such a distribution always exists when the initial probability matrix P satisfies certain properties and enough time has elapsed. This concept is useful because it highlights the idea of **randomness**: there comes a point in time when the chain effectively "forgets" its initial state. In other words, regardless of the starting distribution μ_0, the chain converges to a well-defined distribution over the states that is as spread out as possible.

Definition 2.12 (Stationary Distribution)**.** A probability distribution π on $\mathcal{X}$ is called a **stationary distribution** for the Markov chain with transition matrix P

if

$$\pi = \pi P, \quad \text{that is,} \quad \pi(y) = \sum_{x \in \mathcal{X}} \pi(x) P(x, y), \quad \forall y \in \mathcal{X}.$$

In words, this means that if the chain starts out distributed according to π, then after one step (and in fact after any number of steps) it will still be distributed according to π. Once reached, the stationary distribution remains unchanged under the dynamics of the chain.

Definition 2.13 (Irreducibility)**.** A Markov chain with state space $\mathcal{X}$ and transition matrix P is said to be **irreducible** if for any $x, y \in \mathcal{X}$, there exists $t \in \mathbb{N}$ such that:

$$P^t(x, y) > 0.$$

That is, every state is reachable from every other state in a finite number of steps.

Definition 2.14 (Aperiodicity)**.** A state $x \in \mathcal{X}$ has **period** $d \in \mathbb{N}$ if

$$d = \gcd\{t \geq 1 : P^t(x, x) > 0\}.$$

In other words, the period measures the greatest common divisor of all possible return times to x. It captures the idea of how long it can take, in terms of step lengths, for the chain to return to the state once it has left. A chain is called **aperiodic** if every state has period 1, meaning that returns to each state can eventually occur at arbitrary times without being locked into a fixed cycle.

The stationary distribution of a Markov chain represents the "long-run" behavior of the system: the probabilities of finding the chain in each state after it has been running for a long time. For this concept to be meaningful, we must ensure that such a distribution is *unique* — otherwise, the long-term behavior would depend on which stationary distribution we happened to end up in, making predictions ambiguous. The natural question, then, is: under what conditions does a Markov chain have a single stationary distribution?

A particularly important class of chains, called *ergodic chains*, satisfies this property. We say that a Markov chain is *ergodic* if it is finite, irreducible and aperiodic. The theorem below formalizes the idea that these structural properties are exactly what is needed to guarantee the existence of a *unique* stationary distribution.

Theorem 2.15 (Convergence to Stationarity [3])**.** *Let $\{X_t\}$ be a Markov chain with finite state space $\mathcal{X}$ and transition matrix P. If the chain is **irreducible** and **aperiodic**, then there exists a unique stationary distribution π, and for any initial distribution μ_0,*

$$\lim_{t \to \infty} \|\mu_0 P^t - \pi\|_{\mathrm{TV}} = 0.$$

Proof. Let $\mathcal{X}$ be a finite state space and let P be the transition matrix of an irreducible Markov chain on $\mathcal{X}$. Since P is stochastic, the vector of all ones is a right eigenvector with eigenvalue 1, so 1 is an eigenvalue of P, and equivalently of $P^\top$ (the superscript refers to the transpose of the matrix). Because P is a nonnegative irreducible matrix, the Perron–Frobenius theorem guarantees that this eigenvalue has a strictly positive left eigenvector $v \gg 0$ satisfying $P^\top v = v$, and that the 1-eigenspace of $P^\top$ is one-dimensional. Normalizing v to sum to 1 produces a probability vector

$$\pi = \frac{v}{\sum_{x \in \mathcal{X}} v(x)},$$

which satisfies $\pi^\top P = \pi^\top$, so π is a stationary distribution. The one-dimensionality of the eigenspace implies that any other stationary distribution must be a scalar multiple of π, and the normalization condition $\sum_{x\in\mathcal{X}} \pi(x) = 1$ then forces equality. Thus π exists, is unique, and satisfies $\pi(x) > 0$ for all $x \in \mathcal{X}$ (see [3, Chs. 1–2] for further details).

Aperiodicity ensures that the chain is not trapped in a deterministic cycle. If the period of a state were $d > 1$, then returns to that state would only occur at times that are multiples of d, causing long-run behavior to oscillate between d distinct patterns. Aperiodicity rules out such rigid cycling, allowing the distribution of the chain to converge smoothly to the unique stationary distribution. In particular, for all $x, y \in \mathcal{X}$,

$$\lim_{t\to\infty} P^t(x, y) = \pi(y).$$

Hence, no matter the starting state, the probability of being in y after t steps converges to $\pi(y)$.

Finally, let μ_0 be any initial distribution on $\mathcal{X}$. The distribution of the chain at time t is

$$\mu_t(y) = \sum_{x\in\mathcal{X}} \mu_0(x)P^t(x, y).$$

Taking the limit and using $P^t(x, y) \to \pi(y)$ for each x, y gives

$$\lim_{t\to\infty} \mu_t(y) = \sum_{x\in\mathcal{X}} \mu_0(x)\,\pi(y) = \pi(y),$$

so $\mu_t \to \pi$ pointwise. Since $\mathcal{X}$ is finite, pointwise convergence of probability distributions implies convergence in total variation, and therefore

$$\lim_{t\to\infty} \|\mu_t - \pi\|_{\mathrm{TV}} = 0.$$

□

We have thus established that if a finite Markov chain is ergodic, then the transition probabilities converge to the stationary distribution π regardless of the initial state. This confirms both the existence and the uniqueness of π, and guarantees that in the long run the chain will behave according to this single distribution. With this foundational result, we can now explore a concrete example.

Example 1.16 (The Coupon Collector Problem). Consider a scenario where we repeatedly collect coupons, each labeled with a number from the set $\{1, 2, \ldots, n\}$. At each step, we choose one coupon uniformly at random and add it to our collection if it is not already present. The process continues until we have collected all n distinct coupons.

We can model this situation as a Markov chain whose state space is the set of all subsets of $\{1, 2, \ldots, n\}$:

$$\Omega = 2^{\{1,2,\ldots,n\}},$$

where each state represents the set of distinct coupons collected so far. The chain starts at the empty set $\varnothing$ (no coupons) and evolves by adding a randomly chosen coupon to the set. If the coupon is already in the set, the state remains unchanged for that step. Once the process reaches the full set $\{1, 2, \ldots, n\}$, it stays there forever; in this sense the chain has reached a terminal state.

The time it takes to collect all coupons is called the *cover time* (or *coupon collector time*), formally defined as

$$\tau := \min\{t \geq 0 : X_t = \{1, 2, \ldots, n\}\}.$$

This problem is of particular interest because it quantifies the number of steps required to "cover" the entire set of states in a specific sense, and its analysis connects naturally to the concepts of stationary distribution and convergence we have just established.

One of the key questions for the coupon collector problem is not just the expected time to collect all coupons, but also the probability that the process has not finished by a given time. The next lemma provides a *tail bound* for the cover time τ, showing that it is very unlikely for the process to take much longer than $n \log n$ steps.

Lemma 2.16 (Tail Bound for τ [2])**.** *For any $c > 0$,*

$$\mathbb{P}\big(\tau > n \log n + cn\big) \;\leq\; e^{-c}.$$

Proof. Let us first fix a particular coupon $i \in \{1, \ldots, n\}$. After $n \log n + cn$ steps, the probability that coupon i has not yet been collected is

$$\left(1 - \frac{1}{n}\right)^{n \log n + cn}.$$

Using the inequality $1 - x \leq e^{-x}$ with $x = \frac{1}{n}$, we have

$$\left(1 - \frac{1}{n}\right)^{n \log n + cn} \;\leq\; e^{-(n \log n + cn) \cdot \frac{1}{n}} = e^{-(\log n + c)} = \frac{e^{-c}}{n}.$$

Now apply the *union bound* over all n coupons: the probability that *some* coupon is still missing is at most

$$n \cdot \frac{e^{-c}}{n} = e^{-c}.$$

Since τ is defined as the first time when *no* coupon is missing, the probability that τ exceeds $n \log n + cn$ is bounded by e^{-c}, which is exactly the claim. □

This bound tells us that the distribution of τ is sharply concentrated around $n \log n$: adding only cn extra steps beyond $n \log n$ makes the probability of not having all coupons drop exponentially fast in c. This concentration property mirrors the rapid convergence to stationarity which we will see in other examples such as shuffling and random walks. This lemma will be used in the proof of Theorem 4.2 (top–to–random shuffle) to control the probability that the strong stationary time, which we will define soon, exceeds a given threshold.

3. Mixing Time

In the previous section, we established that an ergodic Markov chain converges to a unique stationary distribution. However, from both a theoretical and practical perspective, knowing *that* convergence happens is only part of the story. In real applications — whether simulating a physical system, designing a randomized algorithm, or sampling from a complicated probability distribution — we need to know *how long* it takes before the chain is "close enough" to stationarity for our purposes.

This is where the concept of *mixing time* becomes essential. The mixing time measures the number of steps required for the distribution of the chain to be within a chosen tolerance of the stationary distribution, no matter where the chain started.

Studying mixing times not only deepens our understanding of Markov chain dynamics, but also connects to a broad range of applications. In this section and the following one, we will introduce the formal definition of mixing time and discuss some properties.

The treatment of mixing time and its fundamental properties is based on the accounts given in Levin–Peres–Wilmer [3] and Sousi [2].

Definition 3.1 (Mixing Time)**.** Let $\{X_t\}_{t\in\mathbb{N}}$ be an ergodic Markov chain with transition matrix P on a finite state space $\mathcal{X}$, and let π be its unique stationary distribution. For $\varepsilon > 0$, the **mixing time** is defined as:

$$t_{\text{mix}}(\varepsilon) := \min\left\{ t \in \mathbb{N} : \max_{x\in\mathcal{X}} \left\| P^t(x,\cdot) - \pi \right\|_{\text{TV}} \leq \varepsilon \right\}.$$

That is, $t_{\text{mix}}(\varepsilon)$ is the smallest time t such that, starting from any initial state $x \in \mathcal{X}$, the distribution after t steps is within ε in total variation distance of the stationary distribution.

Lemma 3.2 (Subadditivity of Total Variation Distance [3])**.** *Let μ, ν, ρ be probability distributions on $\mathcal{X}$. Then:*

$$\|\mu - \rho\|_{\text{TV}} \leq \|\mu - \nu\|_{\text{TV}} + \|\nu - \rho\|_{\text{TV}}.$$

Proof. By the triangle inequality for real numbers:

$$|\mu(x) - \rho(x)| \leq |\mu(x) - \nu(x)| + |\nu(x) - \rho(x)|.$$

Summing over $x \in \mathcal{X}$ and multiplying by $\frac{1}{2}$, we obtain:

$$\|\mu-\rho\|_{\text{TV}} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x)-\rho(x)| \leq \frac{1}{2}\sum_{x\in\mathcal{X}} (|\mu(x)-\nu(x)| + |\nu(x)-\rho(x)|) = \|\mu-\nu\|_{\text{TV}} + \|\nu-\rho\|_{\text{TV}}.$$

□

Theorem 3.3 (Contraction of Total Variation Distance [2])**.** *Let P be the transition matrix of a Markov chain. Then for any two distributions μ, ν on $\mathcal{X}$, and for any $t \in \mathbb{N}$,*

$$\|\mu P^t - \nu P^t\|_{\text{TV}} \leq \|\mu - \nu\|_{\text{TV}}.$$

Proof. Let μ, ν be two distributions on $\mathcal{X}$. Then:

$$(\mu P^t)(y) = \sum_{x\in\mathcal{X}} \mu(x) P^t(x,y), \quad (\nu P^t)(y) = \sum_{x\in\mathcal{X}} \nu(x) P^t(x,y).$$

Taking the difference:

$$|(\mu P^t)(y) - (\nu P^t)(y)| = \left| \sum_{x\in\mathcal{X}} (\mu(x) - \nu(x)) P^t(x,y) \right| \leq \sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)| P^t(x,y).$$

Summing over y:

$$\sum_{y\in\mathcal{X}} |(\mu P^t)(y) - (\nu P^t)(y)| \leq \sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)| \sum_{y\in\mathcal{X}} P^t(x,y) = \sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)|.$$

Therefore,

$$\|\mu P^t - \nu P^t\|_{\text{TV}} \leq \|\mu - \nu\|_{\text{TV}}.$$

□

The contraction property tells us that the total variation distance between two distributions decreases (or at least does not increase) as the chain evolves. In other words, the chain *forgets* its initial state over time: two copies of the chain started from different initial distributions will get progressively closer together in distribution as we apply more steps of the transition matrix.

This observation is a cornerstone in the study of mixing times. It ensures that, although the exact distance to stationarity at time t can depend on the initial distribution, the *rate* at which convergence occurs is uniform: regardless of where the chain starts, the distance to stationarity decreases over time at the same asymptotic rate.

4. Estimation Techniques for Mixing Time

Since total variation distance is sometimes analytically intractable for large state spaces, we rely on probabilistic and computational techniques to estimate the mixing time. In this section, we present two important techniques: **strong stationary times**, and **Monte Carlo sampling**.
The material on this section draws on the discussions in Levin–Peres–Wilmer [3] and Berestycki [4].

Definition 4.1 (Stopping Time)**.** Let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov chain with state space $\mathcal{X}$. A random variable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ is called a **stopping time** with respect to the chain if for every $t \in \mathbb{N}$, the event $\{\tau = t\}$ depends only on the values $X_0, X_1, \ldots, X_t$.

That is, whether or not the process stops at time t is determined by observing the chain up to and including time t.

Definition 4.2 (Strong Stationary Time)**.** A stopping time τ for a Markov chain $\{X_t\}$ is called a **strong stationary time** if:

$$\mathbb{P}(X_\tau = y \mid \tau = t) = \pi(y), \quad \text{for all } y \in \mathcal{X},\ t \in \mathbb{N}.$$

That is, $X_\tau \sim \pi$ and is independent of the stopping time.

Theorem 4.3 (Total Variation Bound via Strong Stationary Time [4])**.** *If τ is a strong stationary time, then:*

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} \le \mathbb{P}_x(\tau > t).$$

Where $\mathbb{P}_x(\tau > t)$ denotes the probability that, when the chain starts from state x, the random time τ has not yet occurred by time t

Proof. Let $A \subseteq \mathcal{X}$ be any measurable set in the state space. Our goal is to bound the absolute difference between the probability that the Markov chain is in set A at time t, and the stationary probability of A.

We begin by decomposing the event $\{X_t \in A\}$ into two disjoint events depending on whether or not the strong stationary time exceeds t:

$$\mathbb{P}(X_t \in A) = \mathbb{P}(X_t \in A,\ \tau \le t) + \mathbb{P}(X_t \in A,\ \tau > t).$$

Now, observe the key idea behind a strong stationary time: it is a stopping time τ such that once the process is stopped at time τ, the state X_τ is exactly distributed according to the stationary distribution π and is independent of the

past. Moreover, the definition implies that after time τ, the process remains in stationarity. In particular, for all $s \leq t$,

$$\mathbb{P}(X_t \in A \mid \tau = s) = \mathbb{P}(X_s \in A \mid \tau = s) = \pi(A),$$

because $X_s = X_\tau \sim \pi$ and does not change in distribution afterward.

Thus, for the first term:

$$\mathbb{P}(X_t \in A,\, \tau \leq t) = \sum_{s=0}^{t} \mathbb{P}(\tau = s)\mathbb{P}(X_t \in A \mid \tau = s) = \sum_{s=0}^{t} \mathbb{P}(\tau = s)\pi(A) = \pi(A)\cdot\mathbb{P}(\tau \leq t).$$

Therefore,

$$\mathbb{P}(X_t \in A) = \pi(A) \cdot \mathbb{P}(\tau \leq t) + \mathbb{P}(X_t \in A,\, \tau > t).$$

Subtracting $\pi(A)$ from both sides, we find:

$$|\mathbb{P}(X_t \in A) - \pi(A)| = |\pi(A) \cdot \mathbb{P}(\tau \leq t) + \mathbb{P}(X_t \in A,\, \tau > t) - \pi(A)|\,.$$

This simplifies to:

$$|\mathbb{P}(X_t \in A) - \pi(A)| = |\mathbb{P}(X_t \in A,\, \tau > t) - \pi(A) \cdot \mathbb{P}(\tau > t)| \leq \mathbb{P}(\tau > t),$$

because the worst-case discrepancy cannot exceed the total probability mass of the event $\{\tau > t\}$.

Finally, taking the supremum over all measurable sets $A \subseteq \mathcal{X}$ gives the total variation bound:

$$\left\|P^t(x, \cdot) - \pi\right\|_{\mathrm{TV}} \leq \mathbb{P}_x(\tau > t).$$

□

Definition 4.4 (Empirical Monte Carlo Estimation)**.** Let $\{X_t^{(i)}\}_{i=1}^N$ be N independent runs of a Markov chain starting from state x. The **empirical distribution** $\widehat{\mu}_N^{(t)}$ at time t is defined by

$$\widehat{\mu}_N^{(t)}(A) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{X_t^{(i)} \in A\}}, \quad \text{for any } A \subseteq \mathcal{X}.$$

In the context of Markov chains, this construction is a way of *approximating the true distribution of the chain at time t*. Instead of calculating $\mu_t(A) = \mathbb{P}(X_t \in A)$ analytically, we simulate N independent copies of the chain, record which states they occupy at time t, and take the fraction that fall inside A. Thus $\widehat{\mu}_N^{(t)}$ is the simulated distribution of the chain at time t, and as N grows large it converges to the exact distribution μ_t.

This empirical distribution approximates $P^t(x, \cdot)$, and one may compute:

$$\left\|\widehat{\mu}_N^{(t)} - \pi\right\|_{\mathrm{TV}} \approx \frac{1}{2} \sum_{x \in \mathcal{X}_{\mathrm{obs}}} \left|\widehat{\mu}_N^{(t)}(x) - \pi(x)\right|,$$

where $\mathcal{X}_{\mathrm{obs}}$ is the set of observed states in the simulation.

This method will be our main computational tool for estimating mixing times in the examples that follow. By repeatedly simulating the Markov chain from a given initial state and recording the proportion of visits to each state at various times, we can construct empirical distributions and compare them to the stationary law using

total variation distance. This approach allows us to visualize the convergence process and to check how closely our theoretical bounds match what actually happens in practice.

In the remainder of this work, we will apply this method to concrete scenarios using Python implementations specifically designed for each case. For the interested reader, the complete source code used to generate our simulations and plots is available at:

https://github.com/JLenin312/Markov_Chains_Mixing_Times_Paper.git

This repository contains a notebook that can be adapted for further experimentation.

5. Top-to-Random Shuffling Model

In this section, we study a classical Markov chain on the symmetric group S_n, modeling the process of shuffling a deck of n cards. The chain is ergodic — finite, irreducible, and aperiodic — and thus converges to the uniform distribution on S_n. Our goal is to analyze its mixing times both theoretically and via simulation. The analysis follows the classical proofs as presented in Levin–Peres–Wilmer [3].

5.1. State Space and Stationary Distribution.

The state space of interest is the symmetric group:

$$S_n = \{\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\} \mid \sigma \text{ is a bijection}\}.$$

Each element $\sigma \in S_n$ corresponds to one possible ordering of the cards. The size of the state space is $|S_n| = n!$, which grows factorially in n.

The stationary distribution π for both models is the **uniform distribution** on S_n, i.e.,

$$\pi(\sigma) = \frac{1}{n!}, \quad \text{for all } \sigma \in S_n.$$

5.2. Top-to-Random Shuffle.

Definition 5.1 (Top-to-Random Shuffle)**.** At each step, the top card of the deck is removed and inserted uniformly at random into one of the n positions in the deck.

This process defines a Markov chain on S_n with transition matrix P. The chain is:

- **Irreducible:** From any permutation σ, we can reach any other $\tau \in S_n$ by successively moving cards from the top into the correct positions.
- **Aperiodic:** For any state $\sigma \in S_n$, there is a positive probability that the top card is reinserted into its original position, leaving the deck unchanged. Thus $P(\sigma, \sigma) > 0$, so the period of every state is 1, and the chain is aperiodic.

Theorem 5.2 (Mixing Time of Top-to-Random Shuffle [4])**.** *Let $\{X_t\}$ be the Markov chain on S_n induced by the top-to-random shuffle. Then for any $\varepsilon \in (0, 1)$,*

$$t_{\text{mix}}(\varepsilon) \leq n \log n + cn,$$

where $c = \log(1/\varepsilon)$.

Proof. We construct a **strong stationary time** τ as follows:

- Initially, all n cards are unmarked.

- Whenever a card is moved (i.e., selected from the top), if it is unmarked, mark it.
- Let τ be the first time at which all cards have been marked.

This process corresponds to the classical **coupon collector problem**, where each "coupon" (card) must be collected at least once. So by standard tail bounds for the coupon collector problem (Lemma 1.14):

$$\mathbb{P}(\tau > n \log n + cn) \leq e^{-c}.$$

Because τ is a strong stationary time, we can directly relate the tail probability of τ to the distance from stationarity. Recall Theorem 4.3:

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbb{P}_x(\tau > t),$$

which tells us that if the strong stationary time has almost certainly occurred by time t, then the chain is very close to its stationary distribution at that time.

In our case, we have already established that

$$\mathbb{P}(\tau > n \log n + cn) \leq e^{-c}.$$

This means that after $n \log n + cn$ steps, the probability that we have *not yet* reached stationarity is at most e^{-c}. In other words, the mixing time $t_{\mathrm{mix}}(\varepsilon)$ is the smallest time at which the chain is within ε of stationarity for **all** starting states. Setting $e^{-c} = \varepsilon$ (so $c = \log(1/\varepsilon)$) and by the tail bound:

$$t_{\mathrm{mix}}(\varepsilon) \leq n \log n + cn,$$

as claimed.

Thus, the combination of the strong stationary time construction and the coupon collector bound gives us a sharp, explicit upper bound on the mixing time. □

Remark 5.3. For $n = 52$ and $\varepsilon = \frac{1}{4}$, this yields $t_{\mathrm{mix}} \approx 278$, reflecting the logarithmic growth rate predicted by the bound $t_{\mathrm{mix}}(\varepsilon) \leq n \log n + cn$.

5.3. **Simulation Results for $n = 6$.** To complement this example, we report a Monte Carlo simulation for a deck of $n = 6$ cards (A simulation for $n = 52$ is computationally difficult since 52! is a very large number). Figure 1 shows the decay of $\|\widehat{\mu}_t - \pi\|_{\mathrm{TV}}$ as a function of the number of shuffles t (solid blue), together with the reference line $\varepsilon = \frac{1}{4}$ (red dashed). We see a short transient fluctuation during the first few shuffles, followed by a rapid drop; in this run, the curve falls below $1/4$ around $t \approx 9$ shuffles. This is consistent with the fact that for small n the empirical mixing time can be substantially better than the general upper bound

$$t_{\mathrm{mix}}(\varepsilon) \;\leq\; n \log n + n \log(1/\varepsilon).$$

For $n = 6$ and $\varepsilon = \frac{1}{4}$, the bound yields

$$n \log n + n \log(1/\varepsilon) \;=\; 6 \log 6 + 6 \log 4 \;\approx\; 19.07,$$

while the simulation indicates that $\|\widehat{\mu}_t - \pi\|_{\mathrm{TV}}$ typically drops below $1/4$ near $t \approx 9$.

This gap can be explained by the fact that the theoretical bound is designed to hold in a very general sense, namely for *all* values of n and for *all* possible starting states of the chain. Because of this generality, the bound cannot take advantage of the fact that for a fixed small n (say $n = 6$) and a specific starting state (sorted deck), convergence may occur much faster. The simulation, by contrast, reflects this faster convergence in the particular small state space we are testing.

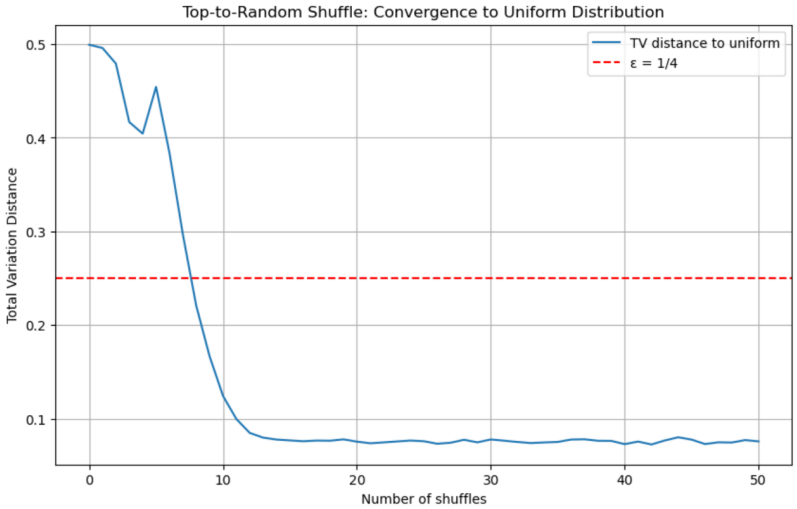

FIGURE 1. Top-to-random shuffle for $n = 6$: empirical total variation distance to the uniform distribution versus shuffle count t.

6. Simple Random Walk on the Complete Graph

In this section we analyze the simple random walk on the complete graph K_n and compute the total variation distance to stationarity exactly, which yields a closed-form expression for the mixing time. The analysis is inspired by the treatment in Berestycki [4].

6.1. Model and basic properties.

Let the state space be the vertex set $X = \{1, 2, \ldots, n\}$ of the complete graph K_n for $n \geq 3$. The simple random walk on K_n is the Markov chain with transition probabilities

$$P(x, y) = \begin{cases} \frac{1}{n-1}, & \text{if } y \neq x, \\ 0, & \text{if } y = x. \end{cases}$$

This finite Markov Chain satisfy our two important properties.

- *Irreducibility:* for any $x \neq y$ we have $P(x, y) = 1/(n-1) > 0$, hence all states communicate in one step.
- *Aperiodicity (when $n \geq 3$):* there are cycles of lengths 2 and 3 through each vertex, so the gcd of return times contains both 2 and 3 and is therefore 1.

Moreover, the uniform distribution $\pi(x) = 1/n$ for $x \in X$ is stationary, since for $x \neq y$,

$$\pi(x)P(x, y) = \frac{1}{n} \cdot \frac{1}{n-1} = \frac{1}{n} \cdot \frac{1}{n-1} = \pi(y)P(y, x),$$

and summing over x gives $\pi P = \pi$.

Theorem 6.1 ([3]). *Fix $n \geq 3$ and a starting state $x \in X$. Let $a_t := P^t(x, x)$ and, for any $y \neq x$, let $b_t := P^t(x, y)$. Then*

$$\left\| P^t(x, \cdot) - \pi \right\|_{\mathrm{TV}} = \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t, \qquad t = 0, 1, 2, \ldots$$

Consequently, for any $\varepsilon \in (0,1)$, the ε-mixing time is

$$t_{\text{mix}}(\varepsilon) = \min\Big\{t \in \mathbb{N} : \Big(1 - \frac{1}{n}\Big)\Big(\frac{1}{n-1}\Big)^t \leq \varepsilon\Big\} = \left\lceil \frac{\log\big(\frac{1-1/n}{\varepsilon}\big)}{\log(n-1)} \right\rceil.$$

Proof. Let X_t be a simple random walk on the complete graph K_n where, at each step, the chain jumps uniformly to a vertex *other than the current one*. The transition matrix P is given by:

$$P(x,y) = \begin{cases} 0 & \text{if } x = y, \\ \frac{1}{n-1} & \text{if } x \neq y. \end{cases}$$

This Markov chain is irreducible and aperiodic (for $n \geq 3$), and it has a unique stationary distribution π, where $\pi(y) = \frac{1}{n}$ for all $y \in \mathcal{X}$.

Let us fix an initial state x and define:

- $a_t := P^t(x,x)$ (the probability of being at the starting state at time t),
- $b_t := P^t(x,y)$ for $y \neq x$ (the probability of being in any other state, which are all equal by symmetry).

Note that the total probability at time t satisfies:

$$a_t + (n-1)b_t = 1.$$

Now, the total variation distance at time t is:

$$\big\|P^t(x,\cdot) - \pi\big\|_{\text{TV}} = \frac{1}{2}\sum_{y \in \mathcal{X}} \left|P^t(x,y) - \frac{1}{n}\right|.$$

Breaking this sum into the contribution from x and the other $n-1$ states:

$$= \frac{1}{2}\left(\left|a_t - \frac{1}{n}\right| + (n-1)\left|b_t - \frac{1}{n}\right|\right).$$

Using the fact that $a_t + (n-1)b_t = 1$, we can compute:

$$b_t = \frac{1 - a_t}{n-1}.$$

Thus:

$$\left|b_t - \frac{1}{n}\right| = \left|\frac{1-a_t}{n-1} - \frac{1}{n}\right|.$$

Let's now compute a_t directly. At $t = 0$, we are at x, so $a_0 = 1$. At $t = 1$, we have $a_1 = 0$ (no self-loops). But for $t \geq 2$, the recurrence relation can be derived as:

$$a_{t+1} = \sum_{y \neq x} \frac{1}{n-1} P^t(x,y) = \sum_{y \neq x} \frac{1}{n-1} b_t = \frac{(n-1)b_t}{n-1} = b_t.$$

Similarly, since $a_t + (n-1)b_t = 1$, we get:

$$b_t = \frac{1-a_t}{n-1}, \quad \Rightarrow \quad a_{t+1} = \frac{1-a_t}{n-1}.$$

This gives a linear recurrence:

$$a_{t+1} = \frac{1-a_t}{n-1}.$$

Solving this recurrence with $a_0 = 1$, we find:

$$a_t = \frac{1}{n} + \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t.$$

Then:

$$\left|a_t - \frac{1}{n}\right| = \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t.$$

Similarly,

$$\left|b_t - \frac{1}{n}\right| = \left|\frac{1-a_t}{n-1} - \frac{1}{n}\right| = \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t \cdot \frac{1}{n-1}.$$

So the total variation becomes:

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \frac{1}{2}\left[\left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t + (n-1)\cdot\left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t \cdot \frac{1}{n-1}\right]$$

$$= \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t.$$

This proves the main formula. To obtain the mixing time, we solve:

$$\left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t \le \varepsilon.$$

Taking logs:

$$t \ge \frac{\log\left(\frac{1-1/n}{\varepsilon}\right)}{\log(n-1)}.$$

So the minimal such integer is:

$$t_{\mathrm{mix}}(\varepsilon) = \left\lceil \frac{\log\left(\frac{1-1/n}{\varepsilon}\right)}{\log(n-1)} \right\rceil.$$

□

6.2. **Simulation Results on K_n for $n = 52$.** To complement, we carried out a Monte Carlo experiment for the simple random walk on the complete graph K_{52}, starting from a fixed vertex Figure 2.

By Theorem 5.1.,

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right)^t.$$

For $n = 52$ this becomes

$$\left\|P^t(x,\cdot) - \pi\right\|_{\mathrm{TV}} = \frac{51}{52}\cdot\frac{1}{51^t} = \begin{cases} \frac{51}{52} \approx 0.9808, & t = 0, \\ \frac{1}{52} \approx 0.01923, & t = 1, \\ \frac{1}{52\cdot 51} \approx 3.77 \times 10^{-4}, & t = 2, \\ \text{and decays by another factor } \frac{1}{51} \text{ each additional step.} \end{cases}$$

Thus, the exact TV distance plunges from ≈ 0.98 at $t = 0$ to ≈ 0.019 after a single step, and is already$< 10^{-3}$ by $t = 2$.

This example vividly contrasts two regimes: a large initial discrepancy at $t = 0$; and a near-instant approach to stationarity on K_{52}, where one step already leaves

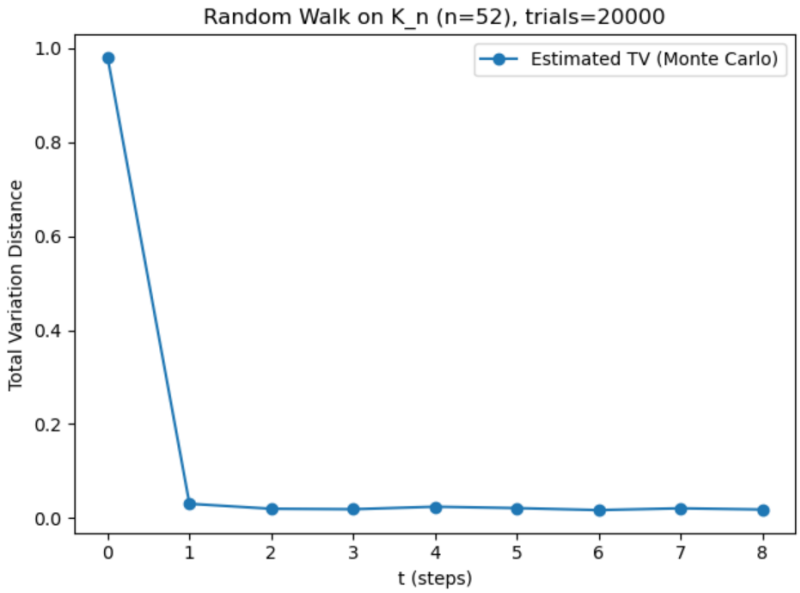

FIGURE 2. Random walk on K_n with $n = 52$: empirical total variation distance to the uniform distribution versus time t.

the walk within about 2% in TV of uniform, and two steps drive it to extreme precision.

ACKNOWLEDGMENTS

I would like to express my gratitude to my mentor, Brin Harper, for her invaluable guidance, encouragement, and insightful feedback throughout the course of this project. Her patience and support were fundamental in shaping both the direction and the clarity of my work. I am also grateful to Peter May and my classmates for making the REU 2025 an unforgettable experience filled with collaboration, learning, and inspiration.

7. BIBLIOGRAPHY

REFERENCES

[1] Gregory F. Lawler. *Introduction to Stochastic Processes*. Retrieved from: https://hamband.math.sharif.ir/wiki/_media///22635/14002/introduction_to_stochastic_processes_-_gregory_f._lawler.pdf

[2] Perla Sousi. *Mixing Times of Markov Chains*. Lecture notes, University of Cambridge. December 8, 2020. Retrieved from: https://www.statslab.cam.ac.uk/~ps422/mixing-notes.pdf

[3] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times* (2nd ed.). Retrieved from: https://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf

[4] Nathanaël Berestycki. *Mixing Times of Markov Chains: Techniques and Examples*. Lecture notes, University of Cambridge. November 22, 2016. Retrieved from: https://homepage.univie.ac.at/nathanael.berestycki/wp-content/uploads/2022/05/mixing3.pdf