1 Introduction

Markov Chain Monte Carlo (MCMC) methods address the problem of sampling from a given distribution by first constructing a Markov chain whose stationary distribution is the given distribution, and then sampling from this Markov chain. Since there are broad classes of Markov chains for which the distribution over states converges to the stationary distribution, MCMC methods such as the Metropolis-Hastings algorithm eventually produce samples from the desired distribution (or some distribution “close” to it). A natural question that arises here is:

*How long do we have to run a Markov chain before its distribution over states is “close” to the stationary distribution?*

In the Markov chain literature, the length of time a Markov chain must run until its distribution is close to the stationary distribution is known as its **mixing time**. As mentioned in the lecture notes, mixing times determine the **burn-in period** of MCMC algorithms (i.e. they determine how many samples must be discarded before useful samples are produced). In the sequel, we will try to answer this question.

1.1 Basics of Markov Chains

We begin by recalling some basic definitions of Markov chains.

**Definition 1** (Markov Chain). A Markov chain is a discrete-time stochastic process \( \{x_n : n \geq 0\} \) with each random variable taking values in a countable state space \( \mathcal{X} \), that satisfies the Markov property:

\[
P (x_n = x_n|x_{n-1} = x_{n-1}, \ldots , x_0 = x_0) = P (x_n = x_n|x_{n-1} = x_{n-1})
\]

for every \( n \geq 1 \) and \( x_0, \ldots , x_n \in \mathcal{X} \) such that \( P (x_0 = x_0, \ldots , x_{n-1} = x_{n-1}) > 0 \). We say \( \{x_n : n \geq 0\} \) is a finite state Markov chain if \( \mathcal{X} \) is a finite set, and we say it is time-homogeneous if for every \( n \geq 1 \) and every \( x, y \in \mathcal{X} \):

\[
P (x_n = y|x_{n-1} = x) = P (x_1 = y|x_0 = x).
\]
We will only consider time-homogeneous finite state Markov chains in our discussion. So, we will refer to a time-homogeneous finite state Markov chain as an “MC” from hereon. Without loss of generality, let $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ with $|\mathcal{X}| \geq 2$, and let $\mathcal{P}$ denote the simplex of all probability distributions on $\mathcal{X}$. We will assume that all distributions in $\mathcal{P}$ are row vectors, i.e. each $\mu \in \mathcal{P}$ can be represented as:

$$\mu = [\mu(1) \ \mu(2) \ \cdots \ \mu(|\mathcal{X}|)].$$

Observe that given the initial state, the distribution of an MC can be succinctly described by its one-step transition probabilities:

$$\forall x, y \in \mathcal{X}, \ W(x, y) \triangleq \mathbb{P}(x_1 = y | x_0 = x)$$

which we usually stack into an $|\mathcal{X}| \times |\mathcal{X}|$ stochastic matrix $W$ whose $(x, y)$th element is $W(x, y)$ for all $x, y \in \mathcal{X}$ (as shown in the lecture notes). $W$ is an entry-wise non-negative matrix whose rows sum to 1. In particular, we will denote the $x$th row of $W$ as $W(x, \cdot) \in \mathcal{P}$ (which is the conditional distribution of the next state given the current state is $x \in \mathcal{X}$).\(^1\)

It is straightforward to verify that for every $n \geq 1$ and every $x, y \in \mathcal{X}$:

$$W^n(x, y) = \mathbb{P}(x_n = y | x_0 = x)$$

which shows that $W^n(x, \cdot) \in \mathcal{P}$ is the conditional distribution of the $n$th state given the initial state is $x \in \mathcal{X}$ (Chapman-Kolmogorov equation). Moreover, if the initial distribution of the MC is $p_{x_0} \in \mathcal{P}$, then the distribution of $x_n$ for every $n \geq 1$ can be obtained by:

$$p_{x_n} = p_{x_0} W^n.$$

Typically, we study properties of MCs that only depend on the transition probabilities. As a result, we usually do not specify an initial distribution, and represent an MC with its stochastic transition probability matrix $W$.

We next present some more definitions relevant to our discussion.

**Definition 2** (Irreducibility and Aperiodicity). An MC with stochastic transition probability matrix $W$ is called irreducible if for every pair of states $x, y \in \mathcal{X}$, there exists some $n \geq 0$ such that $W^n(x, y) > 0$. If $W$ is irreducible, then we say it is aperiodic if every state $x \in \mathcal{X}$ has period $d_x \triangleq \gcd\{n \geq 1 : W^n(x, x) > 0\} = 1$.

These definitions turn out to be the precise conditions under which we observe the behavior of MCs converging to their stationary distributions over time. Intuitively, an MC is irreducible if it is possible to get to any state from any other state after a finite sequence of transitions. This condition allows probabilities to “flow” to different states even if all the probability is initially concentrated at a particular state. However,

\(^1\)Note that we use slightly different notation from the lecture notes here for ease of exposition. For example, we do not use boldface for matrices and vectors.
irreducibility does not preclude the MC with $\mathcal{X} = \{0, \ldots, |\mathcal{X}| - 1\}$ where transitions happen from state $x \in \mathcal{X}$ to state $x + 1 \pmod{|\mathcal{X}|}$ with probability 1. This MC is irreducible, and each of its states has period $|\mathcal{X}|$. If we start this chain at state 0 (i.e. $\mathbb{P}(x_0 = 0) = 1$), we cannot hope for this probability to “diffuse” to all states over time, because the probability mass will periodically cycle over all the states. The condition of aperiodicity is needed to preclude such chains. As we will see, irreducibility and aperiodicity together, allow us to prove convergence to stationary distributions over time. We remark that although Definition 2 requires us to check that $d_x = 1$ for every $x \in \mathcal{X}$ to deduce aperiodicity, it suffices to only check this for one state. Indeed, it is a simple exercise to prove that $d_x = d_y$ for any two states $x, y \in \mathcal{X}$ of an irreducible MC (try it!).

We now state some well-known results about MCs. The proofs are omitted since many readers are probably familiar with these results.

**Theorem 1** (Properties of Markov Chains). Let $W$ be the stochastic transition probability matrix of an MC. Then, the following are true:

1. There exists a stationary distribution $\pi \in \mathcal{P}$ such that $\pi W = \pi$.
2. If $W$ is irreducible, then the stationary distribution $\pi$ is unique and entry-wise strictly positive.
3. If $W$ is irreducible and aperiodic, then there exists some $n \geq 0$ such that $W^n(x, y) > 0$ for all $x, y \in \mathcal{X}$ (i.e. $W^n$ is entry-wise strictly positive).

We remark that the first result is actually an immediate consequence of Brouwer’s fixed-point theorem. It also admits a short proof using linear programming duality, but a probabilistic proof requires some work. For those familiar with matrix theory, we also remark that the third result simply says that irreducible and aperiodic stochastic matrices are primitive matrices.

## 2 Total Variation Distance

In order to prove convergence to stationary distributions, we require a notion of distance between distributions. The classical choice for this is the so called total variation distance (which you were introduced to in the problem sets).

**Definition 3** (Total Variation Distance). Given two distributions $\mu, \nu \in \mathcal{P}$, we define the total variation distance between them as:

$$\|\mu - \nu\|_{TV} \triangleq \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|$$

where $\mu(A) = \sum_{x \in A} \mu(x)$ for any event $A \subseteq \mathcal{X}$. 
This definition perceives distributions (or probability measures) as maps from the set of all events to \([0, 1]\), and measures the maximum deviation between the two distributions over all events. We will prove a few more equivalent characterizations of total variation distance. To this end, we first introduce the notion of couplings.

### 2.1 Coupling

Coupling is a powerful proof technique in probability theory, and we will see some uses of it in the ensuing sections. For now, we formally define it below.

**Definition 4 (Coupling).** Given two probability distributions \(\mu, \nu \in \mathcal{P}\), a coupling between them corresponds to a pair of random variables \((x, y)\) (defined on the same probability space) with joint distribution \(p_{x,y}\) on \(X \times X\) whose marginal distributions satisfy \(p_x = \mu\) and \(p_y = \nu\).

There are several possible couplings between any two \(\mu, \nu \in \mathcal{P}\). For example, we can always define the independent coupling where \(x\) and \(y\) are independent random variables with \(p_{x,y}(x, y) = \mu(x)\nu(y)\) for every \(x, y \in X\). This is typically not a very useful coupling. If \(\mu = \nu\), then we can also define the “identical” coupling with \(x = y\) and \(p_x = \mu\). As we will see next, a particular mixture of these two couplings is closely related to the total variation distance.

### 2.2 Equivalent Characterizations of Total Variation Distance

The next result presents some equivalent characterizations of total variation distance.

**Theorem 2 (Characterizations of Total Variation Distance).** For any two probability distributions \(\mu, \nu \in \mathcal{P}\), we have:

\[
\|\mu - \nu\|_{TV} = \sum_{x \in X : \mu(x) \geq \nu(x)} \mu(x) - \nu(x)
\]

\[
= \frac{1}{2} \|\mu - \nu\|_1
\]

\[
= \min \{ \mathbb{P}(x \neq y) : (x, y) \text{ coupling of } \mu \text{ and } \nu \}
\]

where the first equality illustrates that the event \(S \triangleq \{x \in X : \mu(x) \geq \nu(x)\}\) achieves the maximum in the definition of total variation distance, the second equality is the \(\ell^1\)-norm characterization (recall that \(\ell^1\)-norm is defined as \(\|x\|_1 \triangleq \sum_{i=1}^n |x_i|\) for any \(x \in \mathbb{R}^n\)), and the final equality is the optimal coupling representation.

**Proof.** To prove the first characterization, let \(S \triangleq \{x \in X : \mu(x) \geq \nu(x)\}\). Then, for any event \(A \subseteq X\), we have:

\[
\mu(A) - \nu(A) \leq \mu(S) - \nu(S)
\]
because \( x \in A \setminus S \Rightarrow \mu(x) - \nu(x) < 0 \) and \( x \in S \setminus A \Rightarrow \mu(x) - \nu(x) \geq 0 \). Likewise, we also get \( \nu(A) - \mu(A) \leq \nu(S^c) - \mu(S^c) = \mu(S) - \nu(S) \), which implies that:

\[
|\mu(A) - \nu(A)| \leq \mu(S) - \nu(S).
\]

We can maximize over all \( A \subseteq \mathcal{X} \) on the left hand side and obtain \( \|\mu - \nu\|_{TV} = \mu(S) - \nu(S) \) (where equality is achieved by \( A = S \)). Since we have:

\[
\mu(S) - \nu(S) = \sum_{x \in \mathcal{X} : \mu(x) \geq \nu(x)} \mu(x) - \nu(x)
\]

this proves the first characterization.

To prove the \( \ell^1 \)-norm characterization, notice that \( \|\mu - \nu\|_{TV} = \nu(S^c) - \mu(S^c) = \mu(S) - \nu(S) \) also gives us:

\[
\|\mu - \nu\|_{TV} = \frac{1}{2} (\mu(S) - \nu(S) + \nu(S^c) - \mu(S^c)) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)| = \frac{1}{2} \|\mu - \nu\|_1.
\]

To prove the optimal coupling representation, observe that for any coupling \((x, y)\) of \( \mu \) and \( \nu \) where \( \mu = p_x \) and \( \nu = p_y \), and any event \( A \subseteq \mathcal{X} \), we have:

\[
\mu(A) - \nu(A) = \mathbb{P}(x \in A) - \mathbb{P}(y \in A) = \mathbb{P}(x, y \in A) + \mathbb{P}(x \in A, y \notin A) - \mathbb{P}(y \in A) \leq \mathbb{P}(x \in A, y \in A) + \mathbb{P}(x \in A, y \notin A) - \mathbb{P}(x \in A, y \in A) = \mathbb{P}(x \in A, y \notin A) \leq \mathbb{P}(x \neq y).
\]

Likewise, \( \nu(A) - \mu(A) \leq \mathbb{P}(x \neq y) \), and hence, \( |\mu(A) - \nu(A)| \leq \mathbb{P}(x \neq y) \), which implies that:

\[
\|\mu - \nu\|_{TV} \leq \min \{\mathbb{P}(x \neq y) : (x, y) \text{ coupling of } \mu \text{ and } \nu\}.
\]

So, it suffices to construct a particular joint distribution \( p_{x,y} \) that achieves equality here by making \( x \) equal to \( y \) as much as possible. Let \( a \land b \triangleq \min\{a, b\} \) for any \( a, b \in \mathbb{R} \). Notice that:

\[
\|\mu - \nu\|_{TV} + \sum_{x \in \mathcal{X}} \mu(x) \land \nu(x) = \sum_{x \in S} \mu(x) - \nu(x) + \sum_{x \in \mathcal{X} \setminus S} \mu(x) \land \nu(x) = \sum_{x \in \mathcal{X}} \mu(x) = 1
\]

which means we can define \( \delta \triangleq 1 - \|\mu - \nu\|_{TV} = \sum_{x \in \mathcal{X}} \mu(x) \land \nu(x) \). Furthermore, define the probability distributions (check this!):

\[
\forall x \in \mathcal{X}, \quad p_1(x) = \frac{\mu(x) \land \nu(x)}{\delta},
\]

\[
\forall x \in \mathcal{X}, \quad p_2(x) = \frac{\mu(x) - \nu(x)}{1 - \delta} \mathbb{1}_S(x),
\]

\[
\forall x \in \mathcal{X}, \quad p_3(x) = \frac{\nu(x) - \mu(x)}{1 - \delta} \mathbb{1}_{S^c}(x).
\]
Let \( Z \) be a Bernoulli random variable with \( \mathbb{P}(Z = 1) = 1 - \mathbb{P}(Z = 0) = \delta \). Given \( Z = 1 \), let \((x, y)\) be coupled identically with conditional distribution:

\[
\forall x, y \in \mathcal{X}, \quad p_{x,y|z}(x, y|1) = p_1(x)\mathbb{1}_{x=y}
\]

so that \( x = y \). Given \( Z = 0 \), let \((x, y)\) be coupled independently with conditional distribution:

\[
\forall x, y \in \mathcal{X}, \quad p_{x,y|z}(x, y|0) = p_2(x)p_3(y)
\]

where \( x \neq y \) as \( p_2 \) and \( p_3 \) have disjoint supports. It is straightforward to verify that \( p_x = \delta p_1 + (1 - \delta)p_2 = \mu \) and \( p_y = \delta p_1 + (1 - \delta)p_3 = \nu \), which means \( p_{x,y} \) is a valid coupling with \( \mathbb{P}(x \neq y) = 1 - \delta = \|\mu - \nu\|_{TV} \). This completes the proof.

In the first characterization, the event \( S \) has a useful interpretation in terms of binary hypothesis testing as explored in the problem sets. Indeed, if \( \mu \) and \( \nu \) are likelihoods corresponding to two equiprobable hypotheses, then \( S \) is precisely the decision region where the maximum likelihood (ML) decision rule chooses \( \mu \), and the total probability of error \( P_e \) of the ML decision rule is given by:

\[
P_e = \frac{1}{2}\mu(S^c) + \frac{1}{2}\nu(S) = \frac{1}{2}(1 - (\mu(S) - \nu(S))) = \frac{1}{2}(1 - \|\mu - \nu\|_{TV}). \tag{5}
\]

The \( \ell^1 \)-norm characterization of total variation distance illustrates that it is a valid distance (or metric) between distributions that is symmetric and satisfies the triangle inequality. To interpret the third characterization, recall from subsection 2.1 that if \( \mu = \nu \), then we can define a coupling with \( x = y \). The optimal coupling representation portrays that the closest a coupling can get to having \( x \) identical to \( y \) is the coupling corresponding to total variation distance. Furthermore, couplings that achieve total variation distance and maximize \( \mathbb{P}(x = y) \) are known as maximal couplings. Finally, for those familiar with the Monge-Kantorovich problem from transportation theory, we remark that the optimal coupling representation of total variation distance shows that it is a Wasserstein distance of order 1 with respect to the Hamming metric.

## 3 Convergence and Ergodic Theorems

In this section, we present two fundamental results from the basic theory of MCs. The first is an analog of the strong law of large numbers (SLLN) for irreducible MCs.

**Theorem 3** (Ergodic Theorem). Given a function \( f : \mathcal{X} \to \mathbb{R} \) and an irreducible MC \( \{x_n : n \geq 0\} \) with stationary distribution \( \pi \in \mathcal{P} \), for any initial distribution \( p_{x_0} \in \mathcal{P} \), we have:

\[
\mathbb{P} \left( \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(x_k) = \mathbb{E}_{\pi} [f(x)] \right) = 1
\]

where \( \mathbb{E}_{\pi} [f(x)] = \sum_{x \in \mathcal{X}} \pi(x)f(x) \).
One way to prove this result is to segment the MC into blocks using carefully chosen stopping times and then employing the SLLN. We omit this proof since we do not assume a thorough understanding of such topics. However, it is worth mentioning certain special cases of this result. If the irreducible MC is actually an i.i.d. process (which means the stochastic transition probability matrix $W$ has all rows equal to $\pi$), then Theorem 3 reduces to the SLLN for the sequence of i.i.d. random variables \{f(x_n) : n \geq 1\}. If the function $f(y) = 1_{y=x}$ for some $x \in X$, then Theorem 3 states that:

$$\mathbb{P} \left( \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_{x_k = x} = \pi(x) \right) = 1. \quad (6)$$

This illustrates that an irreducible MC asymptotically spends roughly $\pi(x)$ fraction of its time in state $x \in X$.

Even when (6) holds, we may not have convergence to the stationary distribution due to periodicity of the MC under consideration. So, we will present a second result that guarantees convergence to stationary distributions for irreducible and aperiodic MCs. Before stating the result, we formalize the “distance from stationarity” using total variation distance. For an irreducible MC with stochastic transition probability matrix $W$ and stationary distribution $\pi \in \mathcal{P}$, we define:

$$\forall n \geq 0, \ d(n) \triangleq \max_{x \in X} \|W^n(x, \cdot) - \pi\|_{TV} = \max_{\mu \in \mathcal{P}} \|\mu W^n - \pi\|_{TV} \quad (7)$$

which represents the “distance from stationarity” at time $n$. The maximum in the rightmost extremal problem in (7) can indeed be achieved due to the extreme value theorem. To prove the second equality in (7), notice that $\max_{x \in X} \|W^n(x, \cdot) - \pi\|_{TV} \leq \max_{\mu \in \mathcal{P}} \|\mu W^n - \pi\|_{TV}$ is clearly true, and:

$$\max_{\mu \in \mathcal{P}} \|\mu W^n - \pi\|_{TV} \leq \max_{\mu \in \mathcal{P}} \sum_{x \in X} \mu(x) \|W^n(x, \cdot) - \pi\|_{TV} \leq \max_{x \in X} \|W^n(x, \cdot) - \pi\|_{TV}$$

where the first inequality follows from the triangle inequality. We also note that $d(n)$ is non-increasing in $n$ (you can try to prove this by establishing a data processing inequality for total variation distance). The next result presents the convergence theorem for irreducible and aperiodic MCs.

**Theorem 4 (Convergence Theorem).** Given an irreducible and aperiodic MC with stochastic transition probability matrix $W$ and stationary distribution $\pi \in \mathcal{P}$, there exist constants $\lambda \in (0, 1)$ and $C > 0$ such that:

$$\forall n \geq 0, \ d(n) \leq C \lambda^n.$$ 

**Proof.** First observe from part 3 of Theorem 1 that since $W$ is irreducible and aperiodic, there exists some $m \geq 0$ such that $P = W^m$ is entry-wise strictly positive. This
means that there exists some $\delta \in (0, 1)$ such that $P$ satisfies the Doeblin minorization condition:
\[ \forall x, y \in \mathcal{X}, \ P(x, y) \geq (1 - \delta)\pi(y). \]

Let $1$ denote the $|\mathcal{X}| \times 1$ column vector with all entries equal to 1. Then, we can decompose the MC $P$ into a mixture of independent sampling from $\pi$ and another MC $Q$:
\[ P = (1 - \delta)1\pi + \delta Q \tag{8} \]
where $1\pi$ is a unit rank stochastic matrix with all rows equal to $\pi$, and $Q \triangleq \frac{1}{\delta}(P - (1 - \delta)1\pi)$ is a valid stochastic matrix due to the Doeblin minorization condition. We claim that in fact:
\[ \forall n \geq 1, \ P^n = (1 - \delta^n)1\pi + \delta^n Q^n. \tag{9} \]

For $n = 1$, this is simply (8). Suppose (9) holds for some $n = k \geq 1$: $P^k = (1 - \delta^k)1\pi + \delta^k Q^k$. Then, using this and (8) we have:
\[
\begin{align*}
P^{k+1} &= P^k P = ((1 - \delta^k)1\pi + \delta^k Q^k)(1 - \delta)1\pi + \delta Q) \\
&= (1 - \delta)(1 - \delta^k)1\pi1\pi + \delta(1 - \delta^k)1\pi Q + (1 - \delta)\delta^k Q^k 1\pi + \delta^{k+1} Q^{k+1} \\
&= ((1 - \delta)(1 - \delta^k) + \delta(1 - \delta^k) + (1 - \delta)\delta^k)1\pi + \delta^{k+1} Q^{k+1} \\
&= (1 - \delta^{k+1})1\pi + \delta^{k+1} Q^{k+1}
\end{align*}
\]
where the third line holds because $\pi 1 = 1$ ($\pi$ is a distribution that sums to 1), $\pi Q = \frac{1}{\delta}(\pi P - (1 - \delta)1\pi) = \pi$, and $Q^k 1 = 1$ (rows of a stochastic matrix sum to 1).

By induction, this implies that (9) is true for all $n \geq 1$.

Now observe from (9) that:
\[
\begin{align*}
\forall n \geq 0, \ P^n - 1\pi &= \delta^n(Q^n - 1\pi) \\
\forall n \geq 0, 0 \leq r < m, \ W^{mn+r} - 1\pi &= \delta^n(Q^n W^r - 1\pi) \\
\forall n \geq 0, 0 \leq r < m, \ \max_{x \in \mathcal{X}} \|W^{mn+r}(x, \cdot) - \pi\|_{TV} &= \delta^n \max_{x \in \mathcal{X}} \|Q^n W^r(x, \cdot) - \pi\|_{TV} \\
\forall k \geq 0, \ \max_{x \in \mathcal{X}} \|W^k(x, \cdot) - \pi\|_{TV} &\leq \delta^{\lfloor k/m \rfloor} \\
\forall k \geq 0, \ d(k) &\leq \frac{1}{\delta}(\delta^{1/m})^k
\end{align*}
\]
where the second line follows from substituting $P = W^m$ and then multiplying both sides by $W^r$, the third line follows from equating the $\ell^1$-norms of the rows on both sides and using the $\ell^1$-norm characterization of total variation distance in Theorem 2, and the fourth line holds because total variation distance is always bounded by 1 and $mn + r$ (for fixed $m$ and varying $n \geq 0, 0 \leq r < m$) runs over all $k \geq 0$. This completes the proof. \qed

We remark that another well-known method of proving this result uses coupling ideas, but we omit this alternative proof for brevity. An immediate corollary of this
result is that for an irreducible and aperiodic MC $W$:

$$\forall x, y \in X, \lim_{n \to \infty} W^n(x, y) = \pi(y)$$

which conveys that the MC converges to its stationary distribution regardless of its initial distribution. Moreover, the convergence theorem shows that irreducible and aperiodic MCs converge exponentially fast to their stationary distributions. Unfortunately, it does not provide explicit estimates of the constants $C$ and $\lambda$. So, it does not directly address the question of how many time steps we need to wait in order to guarantee we are close to the stationary distribution in total variation distance (i.e. it does not give us explicit bounds on mixing times). It is worth mentioning that for reversible MCs, the asymptotic rate of convergence to stationarity can be easily shown (via the Perron-Frobenius theorem) to be the second largest eigenvalue modulus (SLEM) of $W$. However, since $C$ is still unknown and could be very large, such SLEM estimates are still not very useful as we only run MCs for finitely many time steps in practice. In the next section, we present a brief introduction to the use of couplings to find explicit upper bounds on mixing times of MCs.

### 4 Upper Bounds on Mixing Times

We first define the notion of a mixing time, which formally captures the minimum amount of time needed for the distance $d(n)$ to be less than some prescribed constant.

**Definition 5 (Mixing Time).** Given an irreducible MC with stochastic transition probability matrix $W$ and stationary distribution $\pi \in \mathcal{P}$, we define the $\epsilon$-mixing time of this MC for any $\epsilon \in (0, 1)$ as:

$$t_{\text{mix}}(\epsilon) \triangleq \min \{ n \geq 0 : d(n) \leq \epsilon \} .$$

Furthermore, we refer to $t_{\text{mix}} \triangleq t_{\text{mix}}(1/4)$ as the mixing time of the MC.

The ensuing subsections illustrate a simple technique to upper bound the mixing times of irreducible and aperiodic MCs.

#### 4.1 Markovian Coupling

The upper bounding technique relies on the idea of couplings. Recall that a coupling of two distributions $\mu, \nu \in \mathcal{P}$ is a pair of jointly distributed random variables $(x, y)$ with joint distribution $p_{x,y}$ on $X \times X$ such that $p_x = \mu$ and $p_y = \nu$. We now define Markovian couplings, which are couplings between MCs.

**Definition 6 (Markovian Coupling).** Suppose $\{x_n : n \geq 0\}$ and $\{y_n : n \geq 0\}$ are two MCs on the state space $X$ with stochastic transition probability matrices $W_1$ and $W_2$, respectively.
respectively. Then, a Markovian coupling of these MCs is the MC \( \{z_n = (x_n, y_n) : n \geq 0\} \) on the state space \( X \times X \) with stochastic transition probability matrix \( P \) that satisfies:

\[
\forall x, x', y \in X, \sum_{y' \in X} P((x, y), (x', y')) = W_1(x, x'),
\]

\[
\forall x, y, y' \in X, \sum_{x' \in X} P((x, y), (x', y')) = W_2(y, y').
\]

Therefore, the Markovian coupling \( \{z_n = (x_n, y_n) : n \geq 0\} \) is a “joint” MC whose “marginals” are themselves the original MCs \( \{x_n : n \geq 0\} \) and \( \{y_n : n \geq 0\} \). Note that as before, there are two trivial Markovian coupling examples. If \( W_1 = W_2 \) and \( p_{x_0} = p_{y_0} \), then we can simply let \( x_n = y_n \) for all \( n \geq 0 \) to obtain the “identical” Markovian coupling \( \{z_n : n \geq 0\} \) with stochastic transition probability matrix:

\[
\forall x, x', y \in X, P((x, y), (x', y')) = W_1(x, x')\mathbb{1}_{x' = y'}
\]

and initial distribution: \( \forall x, y \in X, \ p_{z_0}(x, y) = p_{x_0}(x)\mathbb{1}_{x = y} \). Alternatively, we can run the MCs \( \{x_n : n \geq 0\} \) and \( \{y_n : n \geq 0\} \) independently and obtain the Markovian coupling \( \{z_n : n \geq 0\} \) with stochastic transition probability matrix:

\[
\forall x, x', y, y' \in X, P((x, y), (x', y')) = W_1(x, x')W_2(y, y')
\]

and initial distribution: \( \forall x, y \in X, \ p_{z_0}(x, y) = p_{x_0}(x)p_{y_0}(y) \).

When \( W = W_1 = W_2 \) (but \( p_{x_0} \) is not necessarily equal to \( p_{y_0} \)), a particularly useful fact is that any Markovian coupling can be modified so that the two “marginal” MCs run together after the first time they meet. Formally, given a Markovian coupling with initial distribution \( p_{z_0} \) and stochastic transition probability matrix \( P \), this modified Markovian coupling has the same initial distribution and stochastic transition probability matrix \( Q \) given by:

\[
Q((x, y), (x', y')) = \begin{cases} P((x, y), (x', y')) & , \quad x \neq y \\ W(x, x') & , \quad x = y \text{ and } x' = y' \\ 0 & , \quad x = y \text{ and } x' \neq y' \end{cases}
\]

for every \( x, x', y, y' \in X \). We use couplings of this kind in the next two subsections.

### 4.2 Upper Bounds via Coupling

We now prove an upper bound on the “distance from stationarity” \( d(n) \) using Markovian couplings.

**Theorem 5** (Coupling Upper Bound). Let \( \{x_n : n \geq 0\} \) be an irreducible MC with stochastic transition probability matrix \( W \) and stationary distribution \( \pi \in \mathcal{P} \). For
each pair of states $x, y \in X$, suppose $\{(x_n, y_n) : n \geq 0\}$ is a Markovian coupling of $\{x_n : n \geq 0\}$ with itself, that has been modified to satisfy (13), and starts at the state $(x_0, y_0) = (x, y) \in X \times X$ (i.e. has initial distribution $p_{x_0, y_0}(x, y) = 1$). Let $\mathbb{P}_{x, y}(\cdot)$ be the probability distribution of $\{(x_n, y_n) : n \geq 0\}$ with $(x_0, y_0) = (x, y)$, and define $\tau_{\text{coup}} \triangleq \min\{n \geq 0 : x_n = y_n\}$ to be the first time the “marginal” MCs meet for this Markovian coupling. Then, we have:

$$\forall n \geq 1, \quad d(n) \leq \max_{x, y \in X} \mathbb{P}_{x, y}(\tau_{\text{coup}} > n).$$

**Proof.** First fix any two states $x, y \in X$, and consider the Markovian coupling $\{(x_n, y_n) : n \geq 0\}$ that starts at $(x_0, y_0) = (x, y)$ and runs the two “marginal” MCs together for all $n \geq \tau_{\text{coup}}$. Since $W^n(x, x') = \mathbb{P}_{x, y}(x_n = x')$ and $W^n(y, y') = \mathbb{P}_{x, y}(y_n = y')$ for every $x', y' \in X$ and any fixed $n \geq 1$, we see that $(x_n, y_n)$ is a coupling of the distributions $W^n(x, \cdot) \in \mathcal{P}$ and $W^n(y, \cdot) \in \mathcal{P}$. The optimal coupling characterization of total variation distance in Theorem 2 allows us to upper bound the total variation distance between $W^n(x, \cdot)$ and $W^n(y, \cdot)$:

$$\|W^n(x, \cdot) - W^n(y, \cdot)\|_{\text{TV}} \leq \mathbb{P}_{x, y}(x_n \neq y_n) = \mathbb{P}_{x, y}(\tau_{\text{coup}} > n)$$

where the equality holds because our Markovian coupling runs the two “marginal” MCs together after they meet. This implies that:

$$\forall n \geq 1, \quad \max_{x, y \in X} \|W^n(x, \cdot) - W^n(y, \cdot)\|_{\text{TV}} \leq \max_{x, y \in X} \mathbb{P}_{x, y}(\tau_{\text{coup}} > n).$$

So, it suffices to prove that:

$$\forall n \geq 0, \quad d(n) = \max_{x \in X} \|W^n(x, \cdot) - \pi\|_{\text{TV}} \leq \max_{x, y \in X} \|W^n(x, \cdot) - W^n(y, \cdot)\|_{\text{TV}}.$$  

This holds due to the following sequence of inequalities:

$$\|W^n(x, \cdot) - \pi\|_{\text{TV}} \triangleq \max_{A \subseteq X} |W^n(x, A) - \pi(A)|$$

$$= \max_{A \subseteq X} \left| \sum_{y \in X} \pi(y)(W^n(x, A) - W^n(y, A)) \right|$$

$$\leq \max_{A \subseteq X} \sum_{y \in X} \pi(y)|W^n(x, A) - W^n(y, A)|$$

$$\leq \sum_{y \in X} \pi(y) \max_{A \subseteq X} |W^n(x, A) - W^n(y, A)|$$

$$= \sum_{y \in X} \pi(y) \|W^n(x, \cdot) - W^n(y, \cdot)\|_{\text{TV}}$$

$$\leq \max_{y \in X} \|W^n(x, \cdot) - W^n(y, \cdot)\|_{\text{TV}}$$

11
where the second line holds because \( \pi(A) = \sum_{y \in X} \pi(y)W^n(y, A) \) (since \( \pi \) is the stationary distribution), the third line uses the triangle inequality, the fourth line holds because the maximum of a sum is always upper bounded by the sum over the maximum, and the final line holds because a weighted average is always upper bounded by the maximum element. Taking the maximum over all \( x \in X \) in the final inequality completes the proof.

This result can be used to find explicit upper bounds on mixing times of irreducible and aperiodic MCs as the next example illustrates. We note that there are many other techniques to upper bound mixing times, as well as to lower bound mixing times, but a discussion of these techniques is beyond our scope.

4.3 Example: Lazy Random Walk on the \( k \)-Cycle

As an example, consider the random walk on the \( k \)-cycle for some \( k \geq 2 \). This is an MC with state space \( X = \mathbb{Z}_k = \{0, \ldots, k-1\} \) (which is the finite additive cyclic group of integers modulo \( k \)) and stochastic transition probability matrix given by:

\[
\forall i, j \in X, \quad W(i, j) = \begin{cases} \frac{1}{2}, & j = i + 1 \pmod{k} \text{ or } j = i - 1 \pmod{k} \\ 0, & \text{otherwise} \end{cases} \tag{14}
\]

Equivalently, we can think of the states as vertices of an undirected cycle graph \( C_k \), where at each time step, the MC randomly and uniformly chooses an adjacent vertex and moves to it (i.e. at each step, it moves clockwise or anti-clockwise with probability \( \frac{1}{2} \) each). The random walk on the \( k \)-cycle is an irreducible MC that is aperiodic if \( k \) is odd, and periodic with all states having period 2 if \( k \) is even (check this!).

One way to make this MC aperiodic for all \( k \geq 2 \) is to construct the lazy random walk on the \( k \)-cycle. This MC has stochastic transition probability matrix given by:

\[
\forall i, j \in X, \quad W_{\text{lazy}}(i, j) = \begin{cases} \frac{1}{2}, & j = i \pmod{k} \\ \frac{1}{4}, & j = i + 1 \pmod{k} \text{ or } j = i - 1 \pmod{k} \\ 0, & \text{otherwise} \end{cases} \tag{15}
\]

and is irreducible and aperiodic for all \( k \geq 2 \) (check this!). As before, we can perceive this lazy random walk as an MC on the cycle graph \( C_k \) where at each time step, the walk moves clockwise or anti-clockwise with probability \( \frac{1}{4} \) each, and does not move at all with probability \( \frac{1}{2} \).

We will upper bound the mixing time of the lazy random walk on the \( k \)-cycle. Let \( \{x_n : n \geq 0\} \) denote the MC corresponding to the lazy random walk. For any two states \( x, y \in X \), let \( \{(x_n, y_n) : n \geq 0\} \) denote a Markovian coupling of \( \{x_n : n \geq 0\} \) with itself that starts at \( (x_0, y_0) = (x, y) \). This Markovian coupling is governed by the following dynamics at each time step \( 0 \leq n < t_{\text{coup}} \) (before the “marginal” MCs meet):

- flip an unbiased coin (independent of all other coin tosses),
• if we get heads, then let \( y_{n+1} = y_n \) and generate \( x_{n+1} \) from \( x_n \) according to \( W \) (i.e. move clockwise or anti-clockwise with probability \( \frac{1}{2} \) each),

• if we get tails, then let \( x_{n+1} = x_n \) and generate \( y_{n+1} \) from \( y_n \) according to \( W \) (i.e. move clockwise or anti-clockwise with probability \( \frac{1}{2} \) each).

Furthermore, at each time step \( n \geq t_{\text{coup}} \), \( x_n = y_n \) and transitions occur according to \( W \) lazy (i.e. after meeting, the “marginal” MCs run together). It is straightforward to verify that this describes a valid Markovian coupling of \( \{x_n : n \geq 0\} \) with itself starting at \((x_0, y_0) = (x, y)\) for every \( x, y \in X \). Using Theorem 5, we get that the “distance from stationarity” of the lazy random walk on the \( k \)-cycle is upper bounded by:

\[
\forall n \geq 1, \ d(n) \leq \max_{x,y \in X} \mathbb{P}_{x,y}[t_{\text{coup}} > n] \leq \frac{\max_{x,y \in X} \mathbb{E}_{x,y}[t_{\text{coup}}]}{n} \tag{16}
\]

where the second inequality follows from Markov’s inequality, and \( \mathbb{E}_{x,y}[\cdot] \) denotes the expectation with respect to \( \mathbb{P}_{x,y}(\cdot) \).

It is a classical exercise in probability theory when analyzing the gambler’s ruin model to establish that \( \mathbb{E}_{x,y}[t_{\text{coup}}] = b(x, y)(k - b(x, y)) \), where \( b(x, y) \) denotes the “clockwise distance” between \( x \) and \( y \). We omit a proof of this result for brevity, but using it, we obtain the bound:

\[
\forall n \geq 1, \ d(n) \leq \frac{\max_{x,y \in X} b(x, y)(k - b(x, y))}{n} \leq \frac{k^2}{4n} \tag{17}
\]

where the second inequality holds because \( r(1 - r) \leq \frac{1}{4} \) for all \( r \in [0, 1] \). For any \( \epsilon \in (0, 1) \), we can let \( \frac{k^2}{4n} \leq \epsilon \), and see that \( d(n) \leq \epsilon \) if \( n \geq \frac{k^2}{4\epsilon} \). This produces the following upper bound on the \( \epsilon \)-mixing time:

\[
t_{\text{mix}}(\epsilon) \leq \frac{k^2}{4\epsilon} \tag{18}
\]

which we can specialize (by setting \( \epsilon = \frac{1}{4} \)) to get the following upper bound on the mixing time:

\[
t_{\text{mix}} \leq k^2. \tag{19}
\]

We note that this result is tight in the sense that \( t_{\text{mix}} \geq Ck^2 \) for some constant \( C > 0 \), which can also be proved using fairly simple techniques.

The upper bound in (18) guarantees that after \( \frac{k^2}{4\epsilon} \) time steps, the distribution over the states of the lazy random walk on the \( k \)-cycle is \( \epsilon \)-close to its stationary distribution in total variation distance, regardless of the choice of initial distribution. Hence, such upper bounds on \( \epsilon \)-mixing times indeed address our motivating question from the introduction. Moreover, (18) conveys (as we would expect) that using smaller \( \epsilon \) or larger \( k \) (state space size) increases the \( \epsilon \)-mixing time.
References

The material presented here is largely based on the text: