GENERAL STATE SPACE MARKOV CHAINS

XINYU HUO

Abstract. This report provides a comprehensive overview of fundamental theoretical convergence results in Markov chains on general state spaces, along with a brief glimpse into their applications. Two applications of probability theory to MCMC (central limit theorems and optimal scaling problems) are also discussed

While the majority of the theorems and proofs are drawn from [RR04], I have enriched the content with additional details and examples.

Contents

xinyu.huo@mail.utoronto.ca.

1. Introduction

The subsequent content offers an overview and summary of the supervised reading course I took with Professor Jeffrey Rosenthal at the University of Toronto in Summer 2023. During my study, he motivated me to explore the areas of MCMC that I found intriguing and clarified every concept I was not familiar with. I am grateful for his thoughtful guidance and inspiring advice.

2. General State Markov Chains

In this section, we will delve into the concepts of Markov chains on general state spaces, with a focus on ergodicity and asymptotic convergence. The material presented here draws heavily from the works of [Ros06], [MT93], and [RR04].

2.1. **Fundamentals.** In this subsection, we generalize most of the notions of discrete Markov chains to general (possibly uncountable) state spaces.

Let $\mathcal{X}$ be a *general state space*, which is non-empty (possibly uncountable) set, together with a σ-algebra $\mathcal{G}$ of measurable subsets. We define *transition probabilities* $\{P(x, A)\}_{x\in\mathcal{X}, A\in\mathcal{G}}$ as follows:

(1) For each fixed $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{G})$.
(2) For each fixed $A \subset \mathcal{G}$, $P(x, A)$ is a non-negative measurable function on $\mathcal{X}$.

If $\mathcal{X}$ is countable, then $P(x, \{i\})$ corresponds to the transition probability p_{xi} of discrete Markov chains, which is the probability of moving from State x into State i in a single step. However, in the case of uncountable state spaces, we may have $P(x, \{i\}) = 0$ for all $i \in \mathcal{X}$ (continuous probability distribution). Instead, we use $P(x, A)$ to represent the probability of jumping into the subset A in a single step if the current state is x.

We first consider a finite sequence $\{X_0, X_1, X_2, \ldots, X_n\}$ of random variables on the product space $\prod_{i=0}^{n} \mathcal{X}$ (the direct product of n copies of $\mathcal{X}$), equipped with the product σ-algebra $\bigotimes_{i=0}^{n} \mathcal{G}$.

For any measurable sets $A_i \in \mathcal{X}$, fix a starting point $x \in \mathcal{X}$, we have

$$\mathbf{P}(X_1 \in A_1 | X_0 = x) = P(x, A),$$

$$\mathbf{P}(X_1 \in A_1, X_2 \in A_2 | X_0 = x) = \int_{x_1 \in A_1} P(x, dx_1) P(x_1, A_2),$$

$$\vdots$$

$$\mathbf{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n | X_0 = x) = \int_{x_1 \in A_1} P(x, dx_1) \int_{x_2 \in A_2} P(x_1, dx_2) \cdots \cdots \int_{x_{n-1} \in A_{n-1}} P(x_{n-2}, dx_{n-1}) P(x_{n-1}, A_n).$$

We also need an *initial distribution* ν for X_0, which is any probability distribution on $(\mathcal{X}, \mathcal{G})$. Then, we have

$$\mathbf{P}(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) = \int_{x_0 \in A_0} \nu(dx_0) \int_{x_1 \in A_1} P(x, dx_1) \cdots \cdots \int_{x_{n-1} \in A_{n-1}} P(x_{n-2}, dx_{n-1}) P(x_{n-1}, A_n).$$

The integrals are well-defined by the measuribility of the function $P(x, A)$ of $x \in \mathcal{X}$.

Then, we want to extend the above property to the infinite product space:

Theorem 2.1. For any initial distribution ν over $(\mathcal{X}, \mathcal{G})$ and any transition probability $\{P(x, A)\}$, there exists an stochastic process $X = \{X_0, X_1, \ldots\}$ on $\Omega := \prod_{i=0}^{\infty} \mathcal{X}$ measurable with respect to $\mathcal{F} := \bigotimes_{i=0}^{\infty} \mathcal{G}$ and a probability measure P_ν on $\mathcal{F}$ such that

(1) $P_\nu(B)$ is the probability of the event $\{X \in B\}$ for $B \in \mathcal{F}$;
(2) for any n and measurable sets $A_i \subset \mathcal{X}$, $i = 0, 1, \ldots, n$,

$$P_\nu(A_0 \times A_1 \times \cdots \times A_n) = \int_{x_0 \in A_0} \nu(dx_0) \int_{x_1 \in A_1} P(x, dx_1) \cdots \cdots \int_{x_{n-1} \in A_{n-1}} P(x_{n-2}, dx_{n-1}) P(x_{n-1}, A_n). \tag{2.1}$$

The proof can be found in [MT93] Theorem 3.4.1.

We are ready to define Markov chains on general state space:

Definition 2.2 (General State Space Markov chains)**.** The stochastic process defined in Theorem 2.1 is called a *general state space*(discrete-time, time-homogeneous) *Markov chain*, i.e. the stochastic process $X = \{X_0, X_1, \ldots\}$ on $(\Omega, \mathcal{F})$ with transition probability $\{P(x, A)\}$ and initial distribution ν satisfying Equation (2.1)

Analogous to the countable state space case (discrete Markov chains), general state space Markov chains have the property of being memoryless. This means that the future outcome of the process solely relies on the current state and is independent of the entire past history. Consequently, the transition probability after n steps is solely determined by the initial state. We can inductively establish the *n-step transition probability*, denoted as $P^n(x, A)$, as follows:

$$P^1(x, A) = P(x, A),$$

and

$$P^n(x, A) = \int_{\mathcal{X}} P(x, dz) P^{n-1}(z, A), \quad \forall n > 1.$$

Lemma 2.3. For any $1 < m < n$,

$$P^n(x, A) = \int_{\mathcal{X}} P^m(x, dy) P^{n-m}(y, A), \quad x \in \mathcal{X}, A \in \mathcal{G}. \tag{2.2}$$

Proof.

$$\begin{aligned}\int_{\mathcal{X}} P^m(x,dy)P^{n-m}(y,A) &= \int_{x_1\in\mathcal{X}} P(x,dx_1)\int_{y\in\mathcal{X}} P^{m-1}(x_1,dy)P^{n-m}(y,A)\\ &= \int_{x_1\in\mathcal{X}} P(x,dx_1)\int_{x_2\in\mathcal{X}} P(x_1,dx_2)\int_{y\in\mathcal{X}} P^{m-2}(x_2,dy)P^{n-m}(y,A)\\ &= \int_{x_1\in\mathcal{X}} P(x,dx_1)\int_{x_2\in\mathcal{X}} P(x_1,dx_2)\cdots\int_{x_m\in\mathcal{X}} P(x_{m-1},dx_m)P^{n-m}(x_m,A)\\ &= \int_{x_1\in\mathcal{X}} P(x,dx_1)P^{n-1}(z,A) = P^n(x,A).\end{aligned}$$

□

Intuitively, when X transitioning from x to A in n steps, X may takes any value $y \in \mathcal{X}$ at the intermediate state m; at the m-th step, due to the memorylessness property of Markov chains, the process continues the subsequence $n-m$ steps with respect to the transition probability $P^{n-m}(y,A)$.

Alternatively, we can write Equation (2.2) as

$$\mathbf{P}_x(X_n \in A) = \int_{\mathcal{X}} \mathbf{P}_x(X_m \in dy)\mathbf{P}_y(X_{n-m}\in A),$$

where $\mathbf{P}_x(\cdot)$ denotes the probability of an event conditioning on $X_0 = x$.

Similar to discrete Markov chains, we are interested in the *stationary distribution* of a Markov chain:

Definition 2.4 (Stationary Distribution)**.** The *stationary distribution* of a Markov chain is a probability distribution $\pi(\cdot)$ over $(\mathcal{X},\mathcal{G})$ such that

$$\pi(A) = \int_{\mathcal{X}} \pi(dx)P(x,A), \quad \forall A\in\mathcal{G},$$

or equivalently,

$$\pi(dy) = \int_{x\in\mathcal{X}} \pi(dx)P(x,dy)$$

Example 2.5. Consider the Markov chain on the real line, where $P(x,\cdot)\sim N\left(\frac{x}{2},\frac{3}{4}\right)$ for each $x\in\mathcal{X}$. Equivalently, $X_{n+1} = \frac{1}{2}X_n + U_n$, where $U_n \overset{i.i.d.}{\sim} N\left(0,\frac{3}{4}\right)$.

Consider $\pi(\cdot)\sim N(0,1)$. It suffices to prove the stationary distribution condition over $\{[a,b] : a,b\in\mathbb{R}\}$:

$$\begin{aligned}\int_{\mathbb{R}} \pi(dx)P(x,[a,b]) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\int_a^b \frac{1}{\sqrt{\frac{3}{2}\pi}}\exp\left(\frac{-2\left(y-\frac{x}{2}\right)^2}{3}\right)dydx\\ &= \int_a^b\int_{\mathbb{R}} \frac{1}{\sqrt{3\pi^2}}\exp\left(-\frac{1}{2}x^2 + \frac{-2\left(y-\frac{x}{2}\right)^2}{3}\right)dxdy\\ &= \int_a^b \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)dy = \pi([a,b]).\end{aligned}$$

Then, by countable additivity of probability measures, the stationary distribution condition holds for all $A \in \mathcal{G}$.

Moreover, notice that the stationary distribution of a Markov chain may or may not exist. A key notion related to the existence of stationary distributions is *reversibility*, as follows.

Definition 2.6 (Reversibility)**.** A Markov chain on $\mathcal{X}$ is *reversible* with respect to a probability distribution $\pi(\cdot)$ on $\mathcal{X}$ if

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx), \quad \forall x, y \in \mathcal{X},$$

or equivalently,

$$\int_{x\in A} \pi(dx)P(x, B) = \int_{y\in B} \pi(dy)P(y, A), \quad \forall A, B \in \mathcal{G}.$$

Here comes an important property of reversibility:

Proposition 2.7. If a Markov chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is stationary for the chain.

Proof. By reversibility,

$$\int_{x\in\mathcal{X}} \pi(dx)P(x, dy) = \int_{x\in\mathcal{X}} \pi(dy)P(y, dx) = \pi(dy) \int_{x\in\mathcal{X}} P(y, dx) = \pi(dy).$$

□

However, it should be noted that a Markov chain might not necessarily converge to a stationary state even if it has stationary distribution:

Example 2.8 (Reducible Markov Chain)**.** Suppose $\mathcal{X} = \{1, 2, 3\}$ and $\pi(\{1\}) = \pi(\{2\}) = \pi(\{3\}) = \frac{1}{3}$. Consider the Markov chain on $(\mathcal{X}, 2^{\mathcal{X}})$ with transition probability $P(1, \{1\}) = P(1, \{2\}) = P(2, \{1\}) = P(2, \{2\}) = \frac{1}{2}$ and $P(3, \{3\}) = 1$. Then, the Markov chain is reversible:

$$\pi(\{i\})P(i, \{j\}) = \pi(\{j\})P(j, \{i\}) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}, \quad i, j = 1, 2,$$

$$\pi(\{i\})P(i, \{j\}) = \pi(\{j\})P(j, \{i\}) = 0, \quad i = 1, 2 \text{ and } j = 3,$$

and

$$\pi(\{i\})P(i, \{j\}) = \pi(\{j\})P(j, \{i\}) = \frac{1}{3} \cdot 1 = \frac{1}{3}, \quad i, j = 3.$$

It follows that $\pi(\cdot)$ is the stationary distribution for this Markov chain. However, if we start from State 1, i.e. $X_0 = 1$, then the Markov chain will never reach State 3, i.e. $P(X_n = 3) = 0 \neq \pi(\{3\}) = \frac{1}{3}$ for all n, which means it fails to converge to the stationary distribution $\pi(\cdot)$.

Moreover, in this case, the stationary distribution is not unique; it is easy to verify that $\pi(\{1\}) = \pi(\{2\}) = \frac{1}{2}$ is another stationary distribution.

To avoid this problem, it is natural to consider the case that every state of a Markov chain is accessible from any other state. For countable state spaces, one may be familiar with the concept of *irreducibility*, which entails that for all $i, j \in \mathcal{X}$, there is $n \in \mathbb{N}$

such that $P(X_n = j|X_0 = i) > 0$. As previously mentioned, in the context of uncountable state spaces, it is possible to encounter situations where the transition probability $P(i, \{j\})$ is zero for all $i, j \in \mathcal{X}$. Consequently, $P(X_n = j|X_0 = i)$ is zero for n. Therefore, we introduce the concept *ϕ-irreducibility* for Markov chains on general state spaces, which can be viewed as a weaker version of irreducibility.

Definition 2.9 (ϕ-irreducibility)**.** A chain is *ϕ-irreducible* if there exsits a non-zero σ-finite measure ϕ on $\mathcal{X}$ such that for all $A \subset \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$.

In short, for a ϕ-irreducible Markov chain, there almost every subset $A \subset \mathcal{X}$ is accessible from any state in $\mathcal{X}$ in finite steps.

It is easy to verify that every irreducible discrete Markov chain is ϕ-irreducible with respect to any σ-finite measure ϕ.

Then, we present a running example that we will revisit multiple times throughout the report, which involves the Metropolis-Hastings algorithm introduced in Section 3.2.

Running Example. Here we present an example of a ϕ-irreducible Markov chain.

Let $\pi(\cdot)$ be a probability measure characterized by an unnormalized density function π_u with respect to d-dimensional Lebesgue measure. Consider the Metropolis-Hastings algorithm for π_u with proposal density $q(\mathbf{x}, \cdot)$ with respect to d-dimensional Lebesgue measure. We will show that the resulting Markov chain constructed by the algorithm is π-irreducible if $q(\cdot, \cdot)$ is positive and continuous on $\mathbb{R}^d \times \mathbb{R}^d$ and π_u is positive everywhere.

Fix an arbitrary subset $A \subset \mathcal{X}$ such that $\pi(A) > 0$. Then, there exists $R > 0$ such that $\pi(A_R) > 0$, where $A_R = A \cap B_R(\mathbf{0})$ ($B_R(\mathbf{0})$ represents the ball of radius R centered at $\mathbf{0}$). By continuity, for any $\mathbf{x} \in \mathbb{R}^d$, $\inf_{\mathbf{y} \in A_R} \min\{q(\mathbf{x}, \mathbf{y}), q(\mathbf{y}, \mathbf{x})\} \geq \epsilon$ for some $\epsilon > 0$. Recall that $P(\mathbf{x}, d\mathbf{y}) = q(\mathbf{x}, \mathbf{y})\alpha(\mathbf{x}, \mathbf{y})d\mathbf{y}$. We have

$$
\begin{aligned}
P(\mathbf{x}, A) \geq P(\mathbf{x}, A_R) &\geq \int_{A_R} q(\mathbf{x}, \mathbf{y}) \min\left[1, \frac{\pi_u(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi_u(\mathbf{x})q(\mathbf{x}, \mathbf{y})}\right] d\mathbf{y} \\
&= \int_{A_R} \min\left[q(\mathbf{x}, \mathbf{y}), \frac{\pi_u(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi_u(\mathbf{x})}\right] d\mathbf{y} \\
&\geq \int_{A_R} \epsilon \cdot \min\left[1, \frac{\pi_u(\mathbf{y})}{\pi_u(\mathbf{x})}\right] d\mathbf{y} \\
&= \epsilon \int_{\mathbf{y} \in A_R:\ \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})} 1 d\mathbf{y} + \epsilon \int_{\mathbf{y} \in A_R:\ \pi_u(\mathbf{x}) > \pi_u(\mathbf{y})} \frac{\pi_u(\mathbf{y})}{\pi_u(\mathbf{x})} d\mathbf{y} \\
&= \epsilon \,\mathrm{Leb}\left(\{\mathbf{y} \in A_R :\ \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})\}\right) + \frac{\epsilon K}{\pi_u(\mathbf{x})} \pi\left(\{\mathbf{y} \in A_R :\ \pi_u(\mathbf{x}) > \pi_u(\mathbf{y})\}\right),
\end{aligned}
$$

where $K = \int_{\mathcal{X}} \pi_u(\mathbf{x}) d\mathbf{x}$ is the normalizing constant for π_u.

Since $\pi(A) = \frac{\int_A \pi(\mathbf{y}) d\mathbf{y}}{K}$ for any $A \subset \mathcal{G}$, $\pi(\cdot)$ is absolutely continuous with respect to Lebesgue measure. We need to consider three cases:

- If $\mathrm{Leb}\left(\{\mathbf{y} \in A_R :\ \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})\}\right) \neq 0$ and $\pi\left(\{\mathbf{y} \in A_R :\ \pi_u(\mathbf{x}) > \pi_u(\mathbf{y})\}\right) \neq 0$, then it directly follows from the above calculation that $P(\mathbf{x}, A) > 0$.

- If $\text{Leb}(\{\mathbf{y} \in A_R : \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})\}) = 0$, then $\pi(\{\mathbf{y} \in A_R : \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})\}) = 0$. It follows that
$$\pi(\{\mathbf{y} \in A_R : \pi_u(\mathbf{x}) > \pi_u(\mathbf{y})\}) = \pi(A_R) - 0 > 0.$$
Hence, $P(\mathbf{x}, A) > 0$.
- If $\pi(\{\mathbf{y} \in A_R : \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})\}) = 0$, then $\pi(\{\mathbf{y} \in A_R : \pi_u(\mathbf{x}) > \pi_u(\mathbf{y})\}) = \pi(A_R) - 0 > 0$. By absolute continuity,
$$\text{Leb}(\{\mathbf{y} \in A_R : \pi_u(\mathbf{x}) \leq \pi_u(\mathbf{y})\}) > 0.$$
Hence, $P(\mathbf{x}, A) > 0$.

Therefore, the Markov chain is ϕ-irreducible.

In general, it is easy to verify that Markov chains generated by most algorithms are ϕ-irreducible (e.g. with respect to Lebesgue measure over an appropriate region).

However, even if a Markov chain is ϕ-irreducible, it might not converge in distribution. An example is given as follows:

Example 2.10 (Periodic Markov Chain)**.** Suppose $\mathcal{X} = \{1, 2, 3\}$ and $\pi(\{1\}) = \pi(\{2\}) = \pi(\{3\}) = \frac{1}{3}$. Consider the Markov chain on $(\mathcal{X}, 2^{\mathcal{X}})$ with transition probability $P(1, \{2\}) = P(2, \{3\}) = P(3, \{1\}) = 1$. It is easy to verify that $\pi(\cdot)$ is the stationary distribution.

Let $\phi(\cdot) = \delta_1(\cdot)$. Then, $\phi(A) > 0$ if and only if $1 \in A$. For any $A \ni 1$,
$$P^3(1, A) > P^3(1, \{1\}) = 1 > 0,$$
$$P^2(2, A) > P^2(2, \{1\}) = 1 > 0,$$
$$P^1(3, A) > P^1(3, \{1\}) = 1 > 0.$$
It follows that this Markov chain is ϕ-irreducible.

However, if $X_0 = 1$, then $X_n = 1$ if and only if n is a multiple of 3, which means $P(X_n = 1)$ oscillates between 0 and 1. Hence, $P(X_n = 1) \not\to \pi(\{3\})$.

Hence, the concept of *aperiodicity* becomes necessary. In this report, we adopt the following definition:

Definition 2.11 (Aperiodicity)**.** A Markov chain with stationary distribution $\pi(\cdot)$ is *aperiodic* if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subset \mathcal{X}$ with $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$, such that $\pi(\mathcal{X}_1) > 0$ (and hence $\pi(\mathcal{X}_i) > 0$ for all i). Otherwise, the chain is *periodic*, with *period* equal to the largest such value of d, and corresponding periodic decomposition $\mathcal{X}_1, \ldots, \mathcal{X}_d$.

Intuitively, a periodic Markov chain alternates between visiting disjoint subsets of states, following a specific pattern.

Then, we return to the running example:

Running Example. We will show that the Markov chain constructed before is aperiodic.

Suppose that it is periodic with period d and periodic decomposition $\mathcal{X}_1, \ldots, \mathcal{X}_d$. In particular, we have $P(\mathbf{x}, \mathcal{X}_2) = 1$ for all $\mathbf{x} \in \mathcal{X}_1$. Take any $\mathbf{x} \in \mathcal{X}_1$. Since the chain is π-irreducible and $\pi(\mathcal{X}_1) > 0$, we have proved previously that $P(\mathbf{x}, \mathcal{X}_1) > 0$. Then, $P(\mathbf{x}, \mathcal{X}_2) \leq 1 - P(\mathbf{x}, \mathcal{X}_1) < 1$, which contradicts to $P(\mathbf{x}, \mathcal{X}_2) = 1$.

(Alternatively, since $\pi(\mathcal{X}_1) > 0$, $\mathcal{X}_1$ has positive Lebesgue measure. It follows that

$$P(\mathbf{x}, \mathcal{X}_1) \geq \int_{\mathbf{y}\in\mathcal{X}_1} q(\mathbf{x}, \mathbf{y})\alpha(\mathbf{x}, \mathbf{y})d\mathbf{y} > 0.$$

This again leads to contradiction.)

Similar results holds true for many other MCMC algorithms, including the Gibbs sampler discussed in Section 3.4. In fact, it is quite uncommon for MCMC algorithms to be periodic.

2.2. **Total Variation Measure.** Our focus lies in studying the convergence behaviour of Markov chains. Specifically, we are interested in investigating whether $P^n(x, A)$ is "close" to $\pi(A)$ for sufficiently large values of n, as well as determining the threshold at which the value of n can be considered "large enough" for practical purposes. Hence, it becomes necessary to employ a measure that can quantify the distance between probability measures.

Definition 2.12 (Total Variation Distance)**.** For two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ defined on $(\mathcal{X}, \mathcal{G})$, the *total variation distance* between $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_{A\in\mathcal{G}} |\nu_1(A) - \nu_2(A)|.$$

We then present some simple properties of total variation distance.

Proposition 2.13. (1) $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_{f:\mathcal{X}\to[0,1]} \left|\int f d\nu_1 - \int f d\nu_2\right|$.

(2) $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \frac{1}{b-a} \sup_{f:\mathcal{X}\to[a,b]} \left|\int f d\nu_1 - \int f d\nu_2\right|$ for any $a < b$, and in particular $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \frac{1}{2} \sup_{f:\mathcal{X}\to[-1,1]} \left|\int f d\nu_1 - \int f d\nu_2\right|$.

(3) If $\pi(\cdot)$ is stationary for a Markov chain with transition probability $\{P(x, A)\}$, then $\|P^n(x, \cdot) - \pi(\cdot)\|$ is non-increasing in n, i.e. $\|P^n(x, \cdot) - \pi(\cdot)\| \leq \left\|P^{n-1}(x, \cdot) - \pi(\cdot)\right\|$ for all $n \in \mathbb{N}$.

(4) More generally, letting $(\nu_i P)(A) \coloneqq \int \nu_i(dx)P(x, A)$, we always have $\|(\nu_1 P)(\cdot) - (\nu_2 P)(\cdot)\| \leq \|\nu_1(\cdot) - \nu_2(\cdot)\|$.

(5) Let $t(n) = 2\sup_{x\in\mathcal{X}} \|P^n(x, \cdot) - \pi(\cdot)\|$, where $\pi(\cdot)$ is stationary. Then t is submultiplicative, i.e. $t(m + n) \leq t(m)t(n)$ for $m, n \in \mathbb{N}$.

(6) If $\mu(\cdot)$ and $\nu(\cdot)$ have densities g and h, respectively, with respect to some σ-finite measure $\rho(\cdot)$, and $M = \max(g, h)$ and $m = \min(g, h)$, then

$$\|\mu(\cdot) - \nu(\cdot)\| = \frac{1}{2}\int_{\mathcal{X}} (M - m)\, d\rho = 1 - \int_{\mathcal{X}} m\, d\rho.$$

(7) Given probability measures $\mu(\cdot)$ and $\nu(\cdot)$, there are jointly defined random variables X and Y such that $X \sim \nu(\cdot)$, $Y \sim \nu(\cdot)$, and $\mathbf{P}[X = Y] = 1 - \|\mu(\cdot) - \nu(\cdot)\|$.

Proof. (1) Apply (2) with $a = 0$ and $b = 1$.

(2) Let ρ be any σ-finite measure such that both ν_1 and ν_2 are absolutely continuous with respect to ρ (e.g. $\rho = \nu_1 + \nu_2$). By Radon-Nikodym Theorem, there exists a measure function $g : \mathcal{X} \to [0, \infty)$ and $h : \mathcal{X} \to [0, \infty)$ such that

$$\nu_1(A) = \int_A g\, d\rho \quad \text{and} \quad \nu_2(A) = \int_A h\, d\rho, \quad \forall A \in \mathcal{G}.$$

Then,

$$\begin{aligned}\left|\int f\,d\nu_1 - \int f\,d\nu_2\right| &= \left|\int fg\,d\rho - \int fh\,d\rho\right| \\ &= \left|\int f(g-h)\,d\rho\right| \\ &= \left|\int_{\{g\geq h\}} f(g-h)\,d\rho + \int_{\{g<h\}} f(g-h)\,d\rho\right|,\end{aligned}$$

which is maximized over $0 \leq f \leq 1$ when $f = b$ on $\{x \in \mathcal{X} : g(x) \geq h(x)\}$ and $f = a$ on $\{x \in \mathcal{X} : g(x) < h(x)\}$. Hence,

$$\begin{aligned}\sup_{f:\mathcal{X}\to[a,b]}\left|\int f\,d\nu_1 - \int f\,d\nu_2\right| &= \left|\int_{\{g\geq h\}} b(g-h)\,d\rho + \int_{\{g<h\}} a(g-h)\,d\rho\right| \\ &= \left|b\int_{\{g\geq h\}} (g-h)\,d\rho + a\int_{\{g<h\}} (g-h)\,d\rho\right| \\ &= \left|b\int_{\{g\geq h\}} g\,d\rho - b\int_{\{g\geq h\}} h\,d\rho + a\int_{\{g<h\}} g\,d\rho - a\int_{\{g<h\}} h\,d\rho\right| \\ &= |b\nu_1(\{g\geq h\}) - b\nu_2(\{g\geq h\}) + a\nu_1(\{g<h\}) - a\nu_2(\{g<h\})| \\ &= |a\nu_1(\{g\geq h\}) - a\nu_2(\{g\geq h\}) + b[1-\nu_1(\{g\geq h\})] - b[1-\nu_2(\{g\geq h\}])| \\ &= (a-b)\,|\nu_1(\{g\geq h\}) - \nu_2(\{g\geq h\})| \\ &= (a-b)\left|\int_{\{g\geq h\}} (g-h)\,d\rho\right| \qquad (2.3)\end{aligned}$$

On the other hand, for $A \in G$,

$$|\nu_1(A) - \nu_2(A)| = \left|\int_A (g-h)\,d\rho\right|,$$

which is maximized when $A = \{x : g(x) \geq h(x)\}$. Hence,

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \left|\int_{\{g\geq h\}} (g-h)\,d\rho\right|. \qquad (2.4)$$

Combing Equation (2.3) and (2.4), we get

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \frac{1}{b-a}\sup_{f:\mathcal{X}\to[a,b]}\left|\int f d\nu_1 - \int f d\nu_2\right|.$$

(3) Apply (4) with $\nu_1(\cdot) = P^{n-1}(x,\cdot)$ and $\nu_2(\cdot) = \pi(\cdot)$ since $P^n(x,A) = \int_{y\in\mathcal{X}} P^{n-1}(x,dy)P(y,A) = (P^{n-1}P)(A)$ and $\pi(A) = \int_{x\in\mathcal{X}} \pi(dx)P(x,A) = (\pi P)(A)$ for all $n \in \mathbb{N}$ and $A \in \mathcal{G}$.

(4) For all $A \in \mathcal{G}$,

$$\begin{aligned}
|(\nu_1 P)(A) - (\nu_2 P)(A)| &= \left| \int \nu_1(dx) P(x, A) - \int \nu_2(dx) P(x, A) \right| \\
(f(y) \coloneqq P(y, A)) \qquad &= \left| \int \nu_1(dx) f(y) - \int \nu_2(dx) f(y) \right| \\
(\text{By (1)}) \qquad &\leq \|\nu_1(A) - \nu_2(A)\|.
\end{aligned}$$

(5) Let $\hat{P}(x, \cdot) \coloneqq P^n(x, \cdot) - \pi(\cdot)$, $\hat{Q}(x, \cdot) \coloneqq P^m(x, \cdot) - \pi(\cdot)$, and

$$\hat{P}\hat{Q}f(x) \coloneqq \int_{y \in \mathcal{X}} f(y) \int_{z \in \mathcal{X}} \hat{P}(x, dz) \hat{Q}(z, dy).$$

We have

$$\begin{aligned}
\hat{P}\hat{Q}f(x) &= \int_{y \in \mathcal{X}} f(y) \int_{z \in \mathcal{X}} \hat{P}(x, dz) \hat{Q}(z, dy) \\
&= \int_{y \in \mathcal{X}} f(y) \int_{z \in \mathcal{X}} [P^n(x, dz) - \pi(dz)] [P^m(z, dy) - \pi(dy)] \\
&= \int_{y \in \mathcal{X}} f(y) \int_{z \in \mathcal{X}} P^n(x, dz) P^m(z, dy) - P^n(x, dz)\pi(dy) - \pi(dz) P^m(z, dy) + \pi(dz)\pi(dy) \\
&= \int_{y \in \mathcal{X}} f(y) \int_{z \in \mathcal{X}} [P^n(x, dz) P^m(z, dy) - \pi(dz) P^m(z, dy)] \\
&\quad - \int_{y \in \mathcal{X}} f(y)\pi(dy) \int_{z \in \mathcal{X}} [P^n(x, dz) - \pi(dz)] \\
&= \int_{y \in \mathcal{X}} f(y) P^{n+m}(x, dy) - \pi(dy) - \int_{y \in \mathcal{X}} \pi(dy) (\underbrace{P^n(x, \mathcal{X})}_{=1} - \underbrace{\pi(\mathcal{X})}_{=1}) \\
&= \int_{y \in \mathcal{X}} f(y) \left[P^{n+m}(x, dy) - \pi(dy) \right] \qquad (2.5)
\end{aligned}$$

Let $g(x) \coloneqq \left(\hat{Q}f\right)(x) \coloneqq \int_{y \in \mathcal{X}} \hat{Q}(x, dy) f(y)$ and $g^* \coloneqq \sup_{x \in \mathcal{X}} |g(x)|$. Consider $f : \mathcal{X} \to [0, 1]$. Then,

$$\begin{aligned}
g^* &= \sup_{x \in \mathcal{X}} \left| \int_{y \in \mathcal{X}} [P^m(x, dy) - \pi(dy)] f(y) \right| \\
&= \sup_{x \in \mathcal{X}} \sup_{f : \mathcal{X} \to [0,1]} \left| \int_{y \in \mathcal{X}} [P^m(x, dy) - \pi(dy)] f(y) \right| \\
(\text{By (1)}) \qquad &\leq \sup_{x \in \mathcal{X}} \|P^m(x, \cdot) - \pi(\cdot)\| \\
&= \frac{1}{2} t(m).
\end{aligned}$$

If $g^* = 0$, then $g(x) = \left(\hat{Q}f\right)(x) = 0$ and $\left(\hat{P}\hat{Q}f\right)(x) = 0$ for all $x \in \mathcal{X}$. If $g^* \neq 0$,

$$\begin{aligned} 2\sup_{x\in\mathcal{X}}\left|\left(\hat{P}\hat{Q}f\right)(x)\right| &= 2\sup_{x\in\mathcal{X}}\left|\left[\hat{P}\left(\hat{Q}f\right)\right](x)\right| \\ &= 2g^*\sup_{x\in\mathcal{X}}\left|\left[\hat{P}\left(\frac{g}{g^*}\right)\right](x)\right| \\ &\leq t(m)\sup_{x\in\mathcal{X}}\left|\left[\hat{P}\left(\frac{g}{g^*}\right)\right](x)\right| \end{aligned}$$

Since $g(x) < g^* \Rightarrow -1 \leq \frac{g}{g^*} \leq 1$, applying (2) with $a = -1$ and $b = -1$ gives

$$\begin{aligned} \sup_{x\in\mathcal{X}}\left|\left[\hat{P}\left(\frac{g}{g^*}\right)\right](x)\right| &\leq \sup_{x\in\mathcal{X}}\sup_{f:\mathcal{X}\to[-1,1]}\left|\int_{y\in\mathcal{X}}\hat{P}(x,dy)f(y) - 0\right| \\ &= \sup_{x\in\mathcal{X}} 2\|\hat{P}(x,\cdot)\| \\ &= 2\sup_{x\in\mathcal{X}}\|P^n(x,\cdot) - \pi(\cdot)\| \\ &= t(n). \end{aligned}$$

It follows that for all $f : \mathcal{X} \to [0,1]$,

$$2\sup_{x\in\mathcal{X}}\left|\left(\hat{P}\hat{Q}f\right)(x)\right| \leq t(m)t(n). \tag{2.6}$$

Moreover,

$$\begin{aligned} t(m+n) &= 2\sup_{x\in\mathcal{X}}\left\|P^{m+n}(x,\cdot) - \pi(\cdot)\right\| \\ &= 2\sup_{x\in\mathcal{X}}\sup_{f:\mathcal{X}\to[0,1]}\left|\int_{y\in\mathcal{X}} f(y)\,dP^{m+n}(x,dy) - \int_{y\in\mathcal{X}} f(y)\,d\pi(dy)\right| \\ \text{(By Equation (2.5))} \quad &= 2\sup_{x\in\mathcal{X}}\sup_{f:\mathcal{X}\to[0,1]}|\hat{P}\hat{Q}f(x)| \\ &= \sup_{f:\mathcal{X}\to[0,1]} 2\sup_{x\in\mathcal{X}}|\hat{P}\hat{Q}f(x)| \\ \text{(By Equation (2.6))} \quad &\leq t(m)t(n). \end{aligned}$$

(6) For the first equality: applying (2) with $a = -1$ and $b = 1$ gives

$$\begin{aligned}
\|\mu(\cdot) - \nu(\cdot)\| &= \frac{1}{2} \sup_{f:\mathcal{X}\to[-1,1]} \left| \int f \, d\mu - \int f \, d\nu \right| \\
&= \frac{1}{2} \sup_{f:\mathcal{X}\to[-1,1]} \left| \int f(g-h) d\rho \right| \\
&= \frac{1}{2} \sup_{f:\mathcal{X}\to[-1,1]} \left| \int_{g\geq h} f(g-h) d\rho + \int_{g<h} f(g-h) d\rho \right| \\
&= \frac{1}{2} \left| \int_{g\geq h} (g-h) d\rho + \int_{g<h} (h-g) d\rho \right| \\
&= \frac{1}{2} \left(\int_{g\geq h} (M-m) d\rho + \int_{g<h} (M-m) d\rho \right) \\
&= \frac{1}{2} \int (M-m) d\rho.
\end{aligned}$$

For the second equality: since $M + m = g + h$, we have

$$\begin{aligned}
\int_{\mathcal{X}} (M+m) d\rho &= \int_{\mathcal{X}} (g+h) \, d\rho \\
&= \int_{\mathcal{X}} g \, d\rho + \int_{\mathcal{X}} h \, d\rho \\
&= \mu(\mathcal{X}) + \nu(\mathcal{X}) = 2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{2} \int (M-m) d\rho &= 1 - 1 + \frac{1}{2} \int (M-m) \, d\rho \\
&= 1 - \frac{1}{2} \left(2 - \int (M-m) \, d\rho \right) \\
&= 1 - \frac{1}{2} \left(\int_{\mathcal{X}} (M+m) d\rho - \int (M-m) \, d\rho \right) \\
&= 1 - \int_{\mathcal{X}} m \, d\rho.
\end{aligned}$$

(7) Define g, h, M, n as in (6). Let $a = \int_{\mathcal{X}} m \, d\rho$, $b = \int_{\mathcal{X}} (g - m) \, d\rho$, and $c = \int_{\mathcal{X}} (h - m) \, d\rho$. If any of a, b, c equals 0, the statement is trivial. Assume $a, b, c > 0$. We then construct random variables Z, U, V, I such that Z has density $\frac{m}{a}$, U has density $\frac{g-m}{b}$, V has density $\frac{h-m}{c}$, and I is independent of Z, U, V with $\mathbf{P}[I = 1] = a$ and $\mathbf{P}[I = 0] = 1 - a$. We then let $X = Y = Z$ if $I = 1$, and

$X = U$, $Y = V$ if $I = 0$. For any $A \in \mathcal{G}$,

$$\begin{aligned}
\mathbf{P}(X \in A) &= \mathbf{P}(U \in A, I = 0) + \mathbf{P}(Z \in A, I = 1) \\
&= \mathbf{P}(U \in A)\mathbf{P}(I = 0) + \mathbf{P}(Z \in A)\mathbf{P}(I = 1) \\
&= \int_A \frac{g - m}{b}\, d\rho \cdot (1 - a) + \int_A \frac{m}{a}\, d\rho \cdot a \\
&= \frac{1 - a}{b} \int_A (g - m)\, d\rho + \int_A m\, d\rho \\
&= \frac{1 - \int_{\mathcal{X}} m\, d\rho}{\int_{\mathcal{X}} (g - m)\, d\rho} \int_A (g - m)\, d\rho + \int_A m\, d\rho \\
&= \frac{1 - \int_{\mathcal{X}} m\, d\rho}{\mu(\mathcal{X}) - \int_{\mathcal{X}} m\, d\rho} \int_A (g - m)\, d\rho + \int_A m\, d\rho \\
&= \int_A (g - m)\, d\rho + \int_A m\, d\rho \\
&= \int_A g\, d\rho = \mu(A)
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{P}(Y \in A) &= \mathbf{P}(V \in A, I = 0) + \mathbf{P}(Z \in A, I = 1) \\
&= \mathbf{P}(V \in A)\mathbf{P}(I = 0) + \mathbf{P}(Z \in A)\mathbf{P}(I = 1) \\
&= \int_A \frac{h - m}{c}\, d\rho \cdot (1 - a) + \int_A \frac{m}{a}\, d\rho \cdot a \\
&= \frac{1 - a}{c} \int_A (h - m)\, d\rho + \int_A m\, d\rho \\
&= \frac{1 - \int_{\mathcal{X}} m\, d\rho}{\int_{\mathcal{X}} (h - m)\, d\rho} \int_A (h - m)\, d\rho + \int_A m\, d\rho \\
&= \int_A h\, d\rho = \nu(A).
\end{aligned}$$

Hence, $X \sim \nu(\cdot)$ and $Y \sim \nu(\cdot)$. Moreover,

$$\begin{aligned}
\mathbf{P}(X = Y) &= \mathbf{P}(X = Y, I = 0) + \mathbf{P}(X = Y, I = 1) \\
&= \mathbf{P}(I = 1) \\
&= a \\
&= \int_{\mathcal{X}} m\, d\rho \\
&= 1 - \|\mu(\cdot) - \nu(\cdot)\|. \qquad \text{(By (6))}
\end{aligned}$$

□

2.3. **Asymptotic Convergence Theorem.** Now, we are prepared to present the main asymptotic convergence theorem, the proof of which can be found in Section 4.

Theorem 2.14. Let $X = \{X_1, \dots, \}$ be a Markov chain on a state space $\mathcal{X}$ with countably generated σ-algebra $\mathcal{G}$. If X is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$,

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0,$$

where $\|\cdot\|$ is the total variation distance.

In particular, $\lim_{n\to\infty} P^n(x, A) = \pi(A)$ for all measurable $A \subset \mathcal{X}$.

Note that the theorem applies only when the state space's σ-algebra is countably generated. Indeed, this condition is quite lenient. In fact, any countable state space is guaranteed to be countably generated. Additionally, subsets of $\mathcal{R}^d$ equipped with the standard Borel σ-algebra satisfy this condition as well. This is because the Borel σ-algebra is generated by open balls with rational centers and rational radii, which are countable.

Remark 2.15.

(1) Under the given conditions of Theorem 2.14, if $h : \mathcal{X} \to \mathbb{R}$ with $\pi(|h|) < \infty$, then the strong law of large numbers holds for h, i.e.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \pi(h) \quad \text{with probability 1.}$$

This result can be extended to positive and Harris recurrent chains (see [MT93], Theorem 17.0.1). A Markov chain is *Harris recurrent* if for all $A \subset \mathcal{X}$ with $\pi(A) > 0$ and $x \in \mathcal{X}$, the chain will eventually reach A from x with probability 1, i.e. $\mathbf{P}(\exists n \text{ s.t. } X_n \in A | X_0 = x) = 1$.

(2) To utilize Theorem 2.14, Markov chains must satisfy three conditions: ϕ-irreducibility, aperiodicity, and possess a stationary distribution. Theorem 2.14 is widely applied to MCMC algorithms because most MCMC algorithms inherently produce chains that are ϕ-irreducible and aperiodic, while also aiming to generate chains with a desired stationary distribution π.

It is important to note that the convergence mentioned in Theorem 2.14 is specifically for almost every $x \in \mathcal{X}$ with respect to the stationary distribution π. However, the chain may exhibit unpredictable behaviour on a null set of π-measure 0, leading to failure of convergence in that region. To illustrate this, consider the following simple example:

Example 2.16. Let $\mathcal{X} = \{1, 2, \dots, \}$. Let $P(1, \{1\}) = 1$, and for $x \geq 2$, $P(x, \{1\}) = \frac{1}{x^2}$ and $P(x, \{x+1\}) = 1 - \frac{1}{x^2}$. The chain has stationary distribution $\pi(\cdot) = \delta_1(\cdot)$. Indeed,

$$\int_{x\in\mathcal{X}} \pi(dx) P(x, \{1\}) = \pi(\{1\}) P(1, \{1\}) = \pi(\{1\}),$$

and

$$\int_{x\in\mathcal{X}} \pi(dx) P(x, \{i\}) = \pi(\{1\}) P(1, \{i\}) = 0 = \pi(\{i\}), \quad \forall i > 1.$$

Also, it is π-irreducible since for all A containing 1,

$$P(i, A) = 1 > 0, \quad \text{if } i = 1,$$

and

$$P(i, A) \geq \frac{1}{x^2} > 0, \quad \text{if } i > 1.$$

Moreover, it is aperiodic since a subset has positive π-measure if and only if it contains 1.

Hence, by applying Theorem 2.14, we have for π-a.e. $x \in \mathcal{X}$, the chain converges to the stationary distribution π with probability 1.

Consider $X_0 = x \geq 2$, then $\mathbf{P}[X_n = x + n \text{ for all } n] = \prod_{j=x}^{\infty} \left(1 - \frac{1}{j^2}\right) > 0$. Hence, $P^n(x, \{3, 4, \ldots\}) \not\to \pi(\{3, 4, \ldots\}) = 0$. Here Theorem 2.14 holds only for $x = 1$ which is indeed π-a.e. $x \in \mathcal{X}$, but it does not hold for $x \geq 2$.

Remark 2.17. The transient behaviour of the chain on the null set in Example 2.16 is not a random occurrence. In the scenario where the chain converges to a different stationary distribution on the null set, it will still possess a positive probability of escaping the null set due to its ϕ-irreducibility (for all states in the null set, we have $\phi(A) > 0 \Rightarrow \exists n$ s.t. $P^n(x, A) > 0$). Then, the chain would eventually exit the null set with probability 1 and thus converge to $\pi(\cdot)$ from the null set.

Under what circumstances do the conclusions of Theorem 4 hold for all $x \in \mathcal{X}$, rather than just π-almost everywhere? This occurs when the transition probability $P(x, \cdot)$ is absolutely continuous with respect to $\pi(\cdot)$ for all $x \in \mathcal{X}$. In such cases, the chain will not be able to escape the null set since $\pi(A) = 0 \Rightarrow P(x, A) = 0$. Consequently, the chain will converge to $\pi(\cdot)$ within the null set. This property also holds for any Metropolis algorithm where the proposal distributions $Q(x, \cdot)$ are absolutely continuous with respect to $\pi(\cdot)$. More generally, we can extend the same result to Harris recurrent chains, which is a stronger condition compared to ϕ-irreducibility with respect to $\pi(\cdot)$.

Lastly, we explore periodic chains, as they occasionally emerge in MCMC algorithms, and many of the theory can be applied to this case.

Corollary 2.18. If a Markov chain is ϕ-irreducible, with period $d \geq 2$, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$,

$$\lim_{n \to \infty} \left\| \frac{1}{d} \sum_{i=n}^{n+d-1} P^i(x, \cdot) - \pi(\cdot) \right\| = 0$$

and also the strong law of large numbers in Remark 2.15 (1) continues to hold without change.

Proof. Let $\mathcal{X}_1, \ldots, \mathcal{X}_d \subset \mathcal{X}$ be the periodic decomposition. Let P' be the d-step chain P^d restricted to $\mathcal{X}_1$. Clearly, P' is ϕ-irreducible and aperiodic.

Let $\pi'(\cdot)$ denote the stationary distribution of P'. Then, for

$$(\pi' P^{j-1})(A) := \int_{x \in \mathcal{X}_1} \pi'(dx) P^{j-1}(x, A),$$

we have

$$\begin{aligned}\int_{x\in\mathcal{X}_j}(\pi' P^{j-1})(dx)P^d(x,A) &= \int_{x\in\mathcal{X}_j}\int_{y\in\mathcal{X}_1}\pi'(dy)P^{j-1}(y,dx)P^d(x,A)\\ &= \int_{x\in\mathcal{X}_j}\int_{y\in\mathcal{X}_1}\int_{z\in\mathcal{X}_1}\pi'(dy)P^{j-1}(y,dx)P^{d-(j-1)}(x,dz)P^{j-1}(z,A)\\ &= \int_{y\in\mathcal{X}_1}\int_{z\in\mathcal{X}_1}\pi'(dy)P^d(y,dz)P^{j-1}(z,A)\\ &= \int_{z\in\mathcal{X}_1}\pi(dz)P^{j-1}(z,A)\\ &= (\pi' P^{j-1})(A).\end{aligned}$$

Hence, $(\pi' P^{j-1})(\cdot)$ is the stationary distribution of the d-step chain restricted to $\mathcal{X}_j$. Moreover,

$$\begin{aligned}\int_{x\in\mathcal{X}}\left(\frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(dx)\right)P(x,A) &= \frac{1}{d}\sum_{j=1}^{d}\int_{x\in\mathcal{X}_j}(\pi' P^{j-1})(dx)P(x,A)\\ &= \frac{1}{d}\sum_{j=1}^{d}\int_{x\in\mathcal{X}_j}\int_{y\in\mathcal{X}_1}\pi'(dy)P^{j-1}(y,dx)P(x,A)\\ &= \frac{1}{d}\sum_{j=1}^{d}\int_{y\in\mathcal{X}_1}\pi'(dy)P^j(y,A)\\ &= \frac{1}{d}\sum_{j=1}^{d}(\pi' P^j)(A)\\ \text{(by periodicity)}\qquad &= \frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(A),\end{aligned}$$

and

$$\frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\mathcal{X}_j) = \frac{1}{d}\sum_{j=0}^{d-1}1 = 1.$$

It follows that $\pi(\cdot) = \frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)$.

Due to periodicity, we assume WLOG that $x \in \mathcal{X}_1$. From Proposition 2.13 (4), we have

$$\|P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)\| \le \|P^{md}(x,\cdot) - \pi'(\cdot)\|, \quad \forall j \in \mathbb{N}.$$

Then,

$$\left\|\frac{1}{d}\sum_{i=md}^{md+d-1}P^i(x,\cdot)-\pi(\cdot)\right\| = \left\|\frac{1}{d}\sum_{i=0}^{d-1}P^{md+j}(x,\cdot)-\frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)\right\|$$

(triangle inequality)
$$\leq \frac{1}{d}\sum_{i=0}^{d-1}\left\|P^{md+j}(x,\cdot)-(\pi' P^j)(\cdot)\right\|$$
$$\leq \frac{1}{d}\sum_{i=0}^{d-1}\left\|P^{md}(x,\cdot)-\pi'(\cdot)\right\| = \frac{1}{d}\sum_{i=0}^{d-1}\left\|P'(x,\cdot)-\pi'(\cdot)\right\|$$

Applying Theorem 2.14 to P' gives

$$\lim_{m\to\infty}\left\|P^{md}(\cdot)-\pi'(\cdot)\right\| = 0, \quad \forall\pi\text{-a.e. } x\in\mathcal{X}.$$

Hence,

$$\lim_{m\to\infty}\left\|\frac{1}{d}\sum_{i=md}^{md+d-1}P^i(x,\cdot)-\pi(\cdot)\right\| \leq \lim_{m\to\infty}\left\|P^{md}(\cdot)-\pi'(\cdot)\right\| = 0, \quad \forall\pi\text{-a.e. } x\in\mathcal{X}.$$

By Proposition 2.13 (3), we can conclude that

$$\lim_{n\to\infty}\left\|\frac{1}{d}\sum_{i=n}^{n+d-1}P^i(x,\cdot)-\pi(\cdot)\right\| \leq \lim_{m\to\infty}\left\|\frac{1}{d}\sum_{i=md}^{md+d-1}P^i(x,\cdot)-\pi(\cdot)\right\| = 0, \quad \forall\pi\text{-a.e. } x\in\mathcal{X}.$$

To establish the strong law of large numbers, let $\overline{P}$ be the transition probability over the state space $\mathcal{X}_1\times\cdots\mathcal{X}_d$ with corresponding sequence $\{(X_{md}, X_{md+1},\ldots,X_{md+d-1})\}_{m=0}^{\infty}$. It is obvious that the chain induced by $\overline{P}$ is ϕ-irreducible and aperiodic with the stationary distribution

$$\overline{\pi} = \pi'\times(\pi' P)\times\cdots\times(\pi' P^{d-1}).$$

Let $h:\mathcal{X}\to\mathbb{R}$ with $\pi(|h|)<\infty$. Define $\overline{h}:\mathcal{X}_1\times\cdots\mathcal{X}_d\to\mathbb{R}$ by $\overline{h}(x_0,\ldots,x_{d-1}) = \frac{1}{d}\sum_{j=0}^{d-1}h(x_j)$. We have $\overline{\pi}(|h|)<\infty$ and then applying the strong law of large numbers to $\overline{P}$ gives

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\overline{h}(X_{id}, X_{id+1},\ldots,X_{id+d-1}) = \overline{\pi}(\overline{h}) \quad \text{with probability 1.}$$

It follows that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}h(X_i) = \pi(h) \quad \text{with probability 1.}$$

□

Remark 2.19. In the case of an irreducible Markov chain with a finite state space, the assumption of periodicity is not required to establish Corollary 2.18.

2.4. **Uniform Ergodicity.** Theorem 2.14 establishes the convergence of a Markov chain to its stationary distribution under certain conditions. However, in practice, we are often interested in understanding the speed at which the chain converges to the stationary distribution. Our"qualitative result" regarding the convergence rate is *uniform ergodicity*.

Definition 2.20 (Uniform Ergodicity)**.** A Markov chain with stationary distribution $\pi(\cdot)$ is *uniformly ergodic* if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M\rho^n, \quad n = 1, 2, 3, \cdots$$

for some $\rho < 1$ and $M < \infty$.

An equivalence of uniform ergodicity is stated as follows:

Proposition 2.21. A Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic if and only if $\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| < \frac{1}{2}$ for some $n \in \mathbb{N}$.

Proof. Assume the chain is uniformly ergodic. Then,

$$\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| \leq M\rho^n \Rightarrow \lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| \leq \lim_{n\to\infty} M\rho^n = 0.$$

Hence, for sufficiently large n, we have

$$\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| < \frac{1}{2}.$$

Conversely, assume $\sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\| < \frac{1}{2}$ for some $n \in \mathbb{N}$. Recall the notation introduced in Proposition 2.13 (5), where $t(n) = 2\sup_{x\in\mathcal{X}} |P^n(x,\cdot) - \pi(\cdot)|$ is defined. Let $\beta \coloneqq t(n) < 1$. Then, by the submultiplicative property,for all $j \in \mathbb{N}$,

$$t(jn) \leq (t(n))^j = \beta^j.$$

It follows from Proposition 2.13 (c) that

$$\|P^m(x,\cdot) - \pi(\cdot)\| \leq \left\|P^{\lfloor m/n\rfloor n}(x,\cdot) - \pi(\cdot)\right\| \leq \frac{1}{2} t\left(\lfloor m/n\rfloor n\right) \leq \beta^{\lfloor m/n\rfloor} \leq \beta^{-1}\left(\beta^{1/n}\right)^m.$$

Therefore, the chain is uniformly ergodic with $M = \beta^{-1}$ and $\rho = \beta^{1/n}$. □

The above proposition holds if we replace $\frac{1}{2}$ by δ for any $0 < \delta < \frac{1}{2}$. However, it is false for $\delta \geq \frac{1}{2}$. Let's see a simple example. Consider $\mathcal{X} = \{1, 2\}$ with $P(1, \{1\}) = P(2, \{2\}) = 1$, and $\pi(\cdot)$ is uniform on $\mathcal{X}$. Then, $\|P^n(x,\cdot) - \pi(\cdot)\| = \frac{1}{2}$ for all $x \in \mathcal{X}$ and $n \in \mathbb{N}$ and thus the chain is not uniformly ergodic.

Next, we explore the condition that guarantees uniform ergodicity. Prior to delving into that discussion, we require the following definition:

Definition 2.22 (Small Set)**.** A subset $C \subset \mathcal{X}$ is *small* (or, (n_0, ϵ, ν)-*small* if there exists a positive integer $n_0, \epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that the following *minorisation condition* holds:

$$P^{n_0}(x,\cdot) \geq \epsilon\nu(\cdot), \quad \forall x \in C, \tag{2.7}$$

i.e. $P^{n_0}(x, A) \geq \epsilon\nu(A)$ for all $x \in C$ and all measurable $A \subset \mathcal{X}$.

Remark 2.23. Alternative formulations of this definition may additionally require that C has positive stationary measure. However, for the sake of simplicity, we do not explicitly impose this requirement. However, $\pi(C) > 0$ follows under the additional assumption of the drift condition discussed in the subsequent section.

Intuitively, this condition means that all of the n_0-step transitions from within C have a overlapped component of size ϵ.

Example 2.24. Let $\mathcal{X}$ be a countable space. If

$$\epsilon_{n_0} \coloneqq \sum_{y\in\mathcal{X}} \inf_{x\in C} P^{n_0}(x,\{y\}) > 0,$$

then C is $(n_0, \epsilon_{n_0}, \nu)$-small where $\nu(\{y\}) = \epsilon_{n_0}^{-1} \inf_{x\in C} P^{n_0}(x,\{y\})$. Indeed,

$$P^{n_0}(x,\{y\}) \geq \epsilon_{n_0} \cdot \nu(\{y\}) = \inf_{x\in C} P^{n_0}(x,\{y\}),$$

$$\nu(\mathcal{X}) = \sum_{y\in\mathcal{X}} \epsilon_{n_0}^{-1} \inf_{x\in C} P^{n_0}(x,\{y\}) = \epsilon_{n_0}^{-1} \underbrace{\sum_{y\in\mathcal{X}} \inf_{x\in C} P^{n_0}(x,\{y\})}_{=\epsilon_{n_0}} = 1,$$

and for all $y_1 \neq y_2$,

$$\begin{aligned}
\nu(\{y_1,y_2\}) &= \epsilon_{n_0}^{-1} \inf_{x\in C} P^{n_0}(x,\{y_1,y_2\}) \\
&= \epsilon_{n_0}^{-1} \inf_{x\in C} P^{n_0}(x,\{y_1\}) + \epsilon_{n_0}^{-1} \inf_{x\in C} P^{n_0}(x,\{y_2\}) \\
&= \nu(\{y_1\}) + \nu(\{y_2\}).
\end{aligned}$$

For a finite state space, if the chain is irreducible (or just indecomposible) and aperiodic, then $\epsilon_{n_0} > 0$ for sufficiently large n_0. Indeed, for any states i, j, there is an $N_{ij} \in \mathbb{N}$ such that $P^n(i,j) > 0$ for all $n \geq N_{ij}$ due to irreducibility and aperiodicity. Then, we can take $M = \max\{N_{ij} : i \in \mathcal{X}, j \in \mathcal{X}\}$, which smaller than infinity since $\mathcal{X}$ is finite. Hence, $P^M(i,j) > 0$ for all $i, j \in \mathcal{X}$.

For a general state space, if the transition probability is absolutely continuous with respect to some measure $\eta(\cdot)$, i.e. $P^{n_0}(x, dy) = p_{n_0}(x,y)$, then we can take $\epsilon_{n_0} = \int_{y\in\mathcal{X}} (\inf_{x\in\mathcal{X}} p_{n_0}(x,y))\, \eta(dy)$.

Definition 2.25 (Pseudo-small Set)**.** A subset $C \subset \mathcal{X}$ is pseudo-small if there exists a $n_0 \in \mathbb{Z}$, $\epsilon > 0$, and a probability measure $\nu_{xy}(\cdot)$ on $\mathcal{X}$ (depending on x, y) such that the following *pseudo-minorisation condition* holds: for all $x, y \in C$,

$$P^{n_0}(x,\cdot) \geq \epsilon \nu_{xy}(\cdot),$$

and

$$P^{n_0}(y,\cdot) \geq \epsilon \nu_{xy}(\cdot).$$

The above notion of pseudo-small set is weaker than that of small set. Intuitively, this condition means for every pair $(x,y) \in C \times C$ of states, the n_0-step transitions has a overlapped component, the size of which depends on the choice of x and y.

Theorem 2.26. Consider a Markov chain with stationary probability distribution $\pi(\cdot)$. Suppose the minorisation condition (2.7) is satisfied for some $n_0 \in \mathbb{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, in the special case $C = \mathcal{X}$ (i.e. the entire state space is small). Then the chain is uniformly ergodic, and in fact $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \mathcal{X}$.

Remark 2.27.
(1) The pseudo-minorisation conditon is perfectly adequate for pairwise coupling construction, which is used to prove Theorem 2.26 in Section 4. Consequently, the minorisation condition in the above theorem can be replaced by the pseudo-minorisation condition, without affecting any bounds that rely on pairwise coupling. This includes all of the bounds explored in this section.
(2) Theorem 2.26 allows us to find a quantitative bound on the distance to stationary distribution $\|P^n(x,\cdot) - \pi(\cdot)\|$. After determining the values of ϵ and n_0, we can identify an appropriate n_* such that $\|P^{n_*}(x,\cdot) - \pi(\cdot)\| \leq c$, where the specific choice of c depends on the context. We can then say that n_* iterations "suffices for convergence" of the Markov chain to a certain standard or level of accuracy. Moreover, for a discrete state space, we can use ϵ_{n_0} specified in Example 2.16.

Running Example. Recalling our previously introduced running example, we have imposed strong conditions of strong continuity on q. Therefore, it is reasonable to conjecture that compact sets would be small. However, without additional regularity conditions, this conjecture proves to be false. Consider the following example: suppose dimension $d = 1$, $\pi_u(x) = \mathbf{1}_{0<|x|<1}|x|^{-1/2}$, and $q(x,y) \propto \exp\left\{-\frac{(x-y)^2}{2}\right\}$. Let N be any neighbourhood containing zero. We will show that N is not small. We have

$$P(x,dy) = q(x,y)\min\left\{1, \frac{\pi_\mu(y)}{\pi_\nu(x)}\right\} dy \propto \exp\left\{-\frac{(x-y)^2}{2}\right\} \min\left\{1, \frac{\mathbf{1}_{0<|y|<1}|y|^{-1/2}}{\mathbf{1}_{0<|x|<1}|x|^{-1/2}}\right\} dy.$$

Let $x \in N$. If $x \to 0$, we observe that $P(x,dy) \to 0$. Hence, the minorisation condition is not satisfied.

Return to the general setup of our running example. Let C be any compact set on which π_u is bounded by $k < \infty$. We will show that C is small. Let $\mathbf{x} \in C$ and D be any compact set of positive Lebesgue and π measure such that $\inf_{\mathbf{x},\mathbf{y}\in C\cup D} q(\mathbf{x},\mathbf{y}) = \epsilon > 0$ and $\sup_{\mathbf{x}\in C,\mathbf{y}\in D} q(\mathbf{x},\mathbf{y}) = M < \infty$ (this is possible since q is continuous). We then have for any $\mathbf{x} \in C$, $\mathbf{y} \in D$,

$$P(\mathbf{x},d\mathbf{y}) \geq q(\mathbf{x},\mathbf{y})\, d\mathbf{y} \min\left\{1, \frac{\pi_\mu(\mathbf{y})q(\mathbf{y},\mathbf{x})}{\pi_\mu(\mathbf{x})q(\mathbf{x},\mathbf{y})}\right\} \geq \epsilon\, d\mathbf{y} \min\left\{1, \frac{\epsilon\pi_\mu(\mathbf{y})}{Mk}\right\}.$$

Therefore, C is small.

We can conclude that if π_u is continuous, the state space $\mathcal{X}$ is compact, and q is continuous and positive, then $\mathcal{X}$ is small; as a result, the Markov chain is guaranteed to be uniformly ergodic.

2.5. **Geometric ergodicity.** Since only a few MCMC algorithms satisfy the requirement of uniform ergodicity, it becomes necessary to relax the minorization condition imposed on the entire state space. Our goal is to establish a more general theorem that can provide bounds on the convergence rate of MCMC algorithms. To this end, we introduce a weaker condition known as geometric ergodicity, defined as follows:

Definition 2.28. A Markov chain with stationary distribution $\pi(\cdot)$ is geometrically ergodic if

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\rho^n, \quad n = 1, 2, 3, \ldots$$

for some $\rho < 1$, where $M(x) < \infty$ for π-a.e. $x \in \mathcal{X}$.

In contrast to uniform ergodicity, geometric ergodicity allows the constant M to depend on the initial state x.

As we have previously shown that all irreducible and aperiodic Markov chains on finite state spaces are uniformly ergodic, it follows that they are also geometrically ergodic. However, in the case of an infinite state space $\mathcal{X}$, the conditions of irreducibility and aperiodicity alone are insufficient to guarantee geometric ergodicity. For example, a symmetric random-walk Metropolis algorithm is geometrically ergodic essentially if and only if $\pi(\cdot)$ has finite exponential moments. As a result, we will now delve into the conditions that establish geometric ergodicity.

Definition 2.29. Given Markov chain transition P on a state space $\mathcal{X}$, and a measurable function $f : \mathcal{X} \to \mathbb{R}$, define the function $Pf : \mathcal{X} \to \mathbb{R}$ such that $(Pf)(x)$ is the conditional expected value of $f(X_{n+1})$, given that $X_n = x$. In symbols, $(Pf)(x) = \int_{y\in\mathcal{X}} f(y)P(x, dy)$.

Definition 2.30 (Drift Condition)**.** A small set C satisfies a *drift condition* (or, *univariate geometric drift condition* if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty]$ such that

$$PV \leq \lambda V + b\mathbf{1}_C, \tag{2.8}$$

i.e. $\int_{y\in X} P(x, dy)V(y) \leq \lambda V(x) + b\mathbf{1}_C(x)$ for all $x \in \mathcal{X}$.

Th main result guaranteeing geometric ergodicity can be stated as follows:

Theorem 2.31. Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Suppose that minorisation condition 2.7 is satisfied for some $C \subset \mathcal{X}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose further that the drift condition 2.8 is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty]$ with $V(X) < \infty$ for at least one $x \in \mathcal{X}$ (and hence for π-a.e.) $x \in \mathcal{X}$. Then, the chain is geometrically ergodic.

Theorem 2.31 is proven in Section 4 by direct coupling constructions.

Example 2.32. Here we consider a simple example of geometric ergodicity of Metropolis algorithms on $\mathbb{R}$. Let $\mathcal{X} = \mathbb{R}^+$ and $\pi_u(x) = e^{-x}$. We will use a symmetric (about x) proposal distribution $q(x, y) = q(|y - x|)$ with support contained in $[x - a, x + a]$. Take

a drift function $V(x) = e^{cx}$ for some $c > 0$. Then, for $x \geq a$, we compute

$$\begin{aligned}
PV(x) &= \int_{y\in\mathcal{X}} V(y)P(x,dy) \\
&= \int_{x-a}^{x+a} V(y)P(x,dy) + V(x)\int_{x-a}^{x+a}(1-P(x,dy)) \\
&= \int_{x-a}^{x} V(y)q(x,y)\alpha(x,y)dy + \int_{x}^{x+a} V(y)q(x,y)\alpha(x,y)dy + \int_{x-a}^{x} V(x)q(x,y)\alpha(x,y)dy \\
&\quad + \int_{x}^{x+a} V(x)q(x,y)\alpha(x,y)dy \\
&= \int_{x-a}^{x} V(y)q(x,y)\cdot 1 dy + \int_{x}^{x+a} V(y)q(x,y)\frac{\pi_u(y)}{\pi_u(x)}dy + \int_{x-a}^{x} V(x)q(x,y)\cdot(1-1)dy \\
&\quad + \int_{x}^{x+a} V(x)q(x,y)\left(1-\frac{\pi_u(y)}{\pi_u(x)}\right)dy \\
&= \int_{x-a}^{x} V(y)q(x,y)dy + \int_{x}^{x+a} V(y)q(x,y)\frac{\pi_u(y)}{\pi_u(x)}dy + \int_{x}^{x+a} V(x)q(x,y)\left(1-\frac{\pi_u(y)}{\pi_u(x)}\right)dy \\
&= \int_{x}^{x+a} V(2x-y)q(x,y)dy + \int_{x}^{x+a} V(y)q(x,y)\frac{\pi_u(y)}{\pi_u(x)}dy + \int_{x}^{x+a} V(x)q(x,y)\left(1-\frac{\pi_u(y)}{\pi_u(x)}\right)dy \\
&= \int_{x}^{x+a} q(x,y)\underbrace{\left[V(2x-y) + V(y)\frac{\pi_u(y)}{\pi_u(x)} + V(x)\left(1-\frac{\pi_u(y)}{\pi_u(x)}\right)\right]}_{=:I(x,y)}dy
\end{aligned}$$

We have

$$\begin{aligned}
I(x,y) &= V(2x-y) + V(y)\frac{\pi_u(y)}{\pi_u(x)} + V(x)\left(1-\frac{\pi_u(y)}{\pi_u(x)}\right) \\
&= e^{c(2x-y)} + e^{cy}\frac{e^{-y}}{e^{-x}} + e^{cx}\left(1-\frac{e^{-y}}{e^{-x}}\right) \\
&= e^{2cx-cy} + e^{cy+x-y} + e^{cx} - e^{cx}e^{x-y} \\
&= e^{cx}\left[e^{-c(y-x)} + e^{(c-1)(y-x)} + 1 - e^{-(y-x)}\right] \\
(u \coloneqq y-x) \qquad &= e^{cx}\left[e^{-cu} + e^{(c-1)u} + 1 - e^{-u}\right] \\
&= 2e^{cx}\left[1 - \underbrace{\frac{(1-e^{(c-1)u})(1-e^{-cu})}{2}}_{=:\epsilon}\right] \\
&= 2V(x)(1-\epsilon),
\end{aligned}$$

Note that $0 < \epsilon < 1$ if $c < 1$. Then, take any $0 < \epsilon < 1$, we have

$$PV(x) \leq V(x)(1-\epsilon)\int_{x}^{x+a} 2q(x,y)dy = (1-\epsilon)V(x)$$

Similarly, one can show that $PV(x)$ is bounded on $[0, a]$. Additionally, since $[0, a]$ is compact and $\pi_u(x)$ is bounded on $[0, a]$, it is a small set. Hence, we have shown that both minorisation and drift conditions are satisfied. The resulting Markov chain is geometrically ergodic by Theorem 2.31.

Geometric ergodicity is a useful property that provides insights into the convergence behaviour of MCMC algorithms. However, it is important to note that while geometric ergodicity is desirable, it does not always guarantee the efficacy of an MCMC algorithm.

Example 2.33 (Witch's Hat)**.** Let $\mathcal{X} = [0, 1]$, $\delta = 10^{-100}$. Consider $\pi_u(\mathbf{x}) = \delta + \mathbf{1}_{[a,a+\delta]}(\mathbf{x})$, where $0 < a < 1 - \delta$. Then,

$$\pi([a, a + \delta]) = \frac{\int_{[a,a+\delta]} \pi_u(\mathbf{x})dx}{\int_{[0,1]} \pi_u(\mathbf{x})dx} = \frac{(\delta + 1)\delta}{(\delta + 1)\delta + \delta(1 - \delta)} = 0.55.$$

Let's run a Metropolis algorithm on π_u. Since the interval $[a, a + \delta]$ is very small, unless the sampler gets really lucky, the outcome will appear to converge to Uniform($[0, 1]$), which is very different from $\pi(\cdot)$. However, the algorithm is still geometrically ergodic (and even uniformly ergodic). Hence, the example illustrates that geometric ergodicity does not necessarily ensure the behaviour of a sampler.

In addition to the above example, there are numerous examples of MCMC algorithms which generate geometrically ergodic, but exhibit extremely slow convergence to the stationary distribution since Theorem 2.31 does not provides no quantitative bounds on $M(x)$ and ρ. As a result, it is preferable, whenever feasible, to obtain explicit quantitative bounds on the convergence of Markov chains.

2.6. **Quantitative Convergence Rates.** Considering the aforementioned, our objective is to establish explicit quantitative bounds on convergence rates. Specifically, we seek bounds of the form

$$P^n(x, \cdot) - \pi(\cdot) \leq g(x, n)$$

where $g(x, n)$ is an explicit function that, ideally, remains small for large n.

Our result requires the following *bivariate drift condition*:

Definition 2.34 (Bivariate Drift Condition)**.** A small set satisfies a *bivariate drift condition* if there is a constant $\alpha > 0$ and a function $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ such that

$$\overline{P}h(x, y) \leq \frac{h(x, y)}{\alpha}, \quad \forall (x, y) \notin C \times C, \tag{2.9}$$

where

$$\overline{P}h(x, y) \coloneqq \int_{\mathcal{X}} \int_{\mathcal{X}} h(z, w)P(x, dz)P(y, dw).$$

Intuitively, $\overline{P}$ represents running two independent copies of the chain. The bivariate drift condition is closely related to the univariate one, as exemplified by the following proposition:

Proposition 2.35. Suppose the univariate drift condition (2.8) is satisfied for some $V : \mathcal{X} \to [1, \infty]$, $C \subset \mathcal{X}$, $\lambda < 1$, and $b < \infty$. Let $d = \inf_{x \in C^{\complement}} V(x)$. Then, if $d > \frac{b}{1-\lambda} - 1$, then the bivariate drift condition (2.9) is satisfied for the same C, with $h(x, y) = \frac{1}{2}[V(x) + V(y)]$ and $\alpha^{-1} = \lambda + \frac{b}{d+1} < 1$.

Proof. If $(x,y) \notin C \times C$, then either $x \notin C$ or $y \notin C$ (or both). Assume WLOG that $x \notin C$, then $V(x) \geq d$ and

$$h(x,y) = \frac{1}{2}\left[V(x) + V(y)\right] \geq \frac{1}{2}(d+1) \Rightarrow \frac{h(x,y)}{\frac{1}{2}(d+1)} \geq 1.$$

Moreover, it follows from the univariate drift condition that $PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b$. Then, we compute $\overline{P}h$

$$\begin{aligned}
\overline{P}h(x,y) &= \int_{\mathcal{X}}\int_{\mathcal{X}} h(z,w)P(x,z)P(y,dw) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \frac{1}{2}\left[V(z) + V(w)\right] P(x,dz)P(y,dw) \\
&= \int_{\mathcal{X}} \frac{1}{2}V(w) + \frac{1}{2}PV(x)P(y,dw) \\
&= \frac{1}{2}\left[PV(x) + PV(y)\right] \\
&\leq \frac{1}{2}\left[\lambda V(x) + \lambda V(y) + b\right] \\
&= \lambda h(x,y) + \frac{b}{2} \\
&\leq \lambda h(x,y) + \frac{b}{2}\frac{h(x,y)}{\frac{1}{2}(d+1)} \\
&= h(x,y)\left(\lambda + \frac{b}{d+1}\right).
\end{aligned}$$

Since $d > \frac{b}{1-\lambda} - 1$ by assumption, we have

$$\begin{aligned}
d+1 > \frac{b}{1-\lambda} &\Rightarrow \frac{b}{d+1} < 1-\lambda \\
&\Rightarrow \frac{b}{d+1} + \lambda < 1.
\end{aligned}$$

Therefore, the bivariate drift is satisfied with $\alpha^{-1} = \lambda + \frac{b}{d+1}$. □

Now, take

$$B \coloneqq \max\left\{1, \alpha^{n_0}(1-\epsilon)\sup_{C\times C}\overline{R}h\right\}, \tag{2.10}$$

where for $(x,y) \in C \times C$,

$$\overline{R}h(x,y) \coloneqq \int_{\mathcal{X}}\int_{\mathcal{X}} (1-\epsilon)^{-2}h(z,w)\left(P^{n_0}(x,dz) - \epsilon\nu(dz)\right)\left(P^{n_0}(y,dw) - \epsilon\nu(dw)\right).$$

Given these assumptions, we present our result as follows:

Theorem 2.36. Consider a Markov chain on a state space $\mathcal{X}$, having transition kernel P. Suppose there is $C \subset \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1,\infty)$, a probability distribution $\nu(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, $n_0 \in \mathbb{N}$, and $\epsilon > 0$, such that the minorisation condition (2.7) and bivariate drift condition (2.9). Define B_{n_0} by (2.10). Then for any joint initial distribution $\mathcal{L}(X_0, X_0')$,

and any integers $1 \le j \le k$, if $\{X_n\}$ and $\{X'_n\}$ are two copies of the Markov chain started in the joint initial distribution $\mathcal{L}(X_0, X'_0)$, then

$$\left\|\mathcal{L}(X_k) - \mathcal{L}(X'_k)\right\| \le (1-\epsilon)^j + \alpha^{-k}(B_{n_0})^{j-1}\mathbf{E}[h(X_0, X'_0)].$$

In particular, by choosing $j = \lfloor rk \rfloor$ for sufficiently small $r > 0$, we obtain an explicit, quantitative convergence bound which goes to 0 exponentially quickly as $k \to \infty$.

Theorem 2.36 is proved in Section 4.

Remark 2.37. Although applying Theorem 2.36 to realistic MCMC algorithms can be challenging, it is feasible and often can rigorously provide a reasonably small number of iterations which is is adequate to ensure convergence.

Alternatively, in cases where applying Theorem 2.36 to complicated Markov chains proves challenging, MCMC practitioners often rely on "convergence diagnostics". These diagnostics involve conducting statistical analyses on the realized output $X_1, X_2, \ldots$ to assess whether the distributions of X_n appear to be "stable" for sufficiently large values of n. For instance, one approach is to run the Markov chain multiple times from different initial states and observe if the chains all converge to approximately the same distribution. This technique often yields satisfactory results in practice. However, it does not provide rigorous guarantees and can occasionally be misled into prematurely claiming convergence.

2.7. **More examples.** Consider a Markov chain $\{X_n\}$ on the real line, where $P(x, \cdot) = N(\frac{x}{2}, \frac{3}{4})$ for each $x \in \mathbb{R}$. Equivalently, $X_{n+1} = \frac{1}{2}X_n + U_{n+1}$, where $\{U_n\}$ are i.i.d. with $U_n = N(0, \frac{3}{4})$.

Note that for any $A \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}
\int_{\mathbb{R}} \pi(dx)P(x,A) &= \int_{x\in\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \left(\int_{y\in A} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy \right) dx \\
&= \int_{x\in\mathbb{R}} \int_{y\in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy dx \\
&= \int_{y\in A} \int_{x\in\mathbb{R}} \frac{2}{2\pi\sqrt{3}} e^{-\frac{1}{2}x^2 - \frac{2}{3}(y-\frac{x}{2})^2} dx dy \\
&= \int_{y\in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \pi(A),
\end{aligned}$$

which means the Markov chain $\{X_n\}$ is stationary with respect to $N(0,1)$.

Then, for any $A \subset \mathcal{B}(\mathbb{R})$ such that $\lambda(A) > 0$ (λ is the Lebesgue measure on $\mathbb{R}$), for all $x \in \mathbb{R}$,

$$P(x, A) = \int_{y\in A} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy > 0.$$

It follows that $\{X_n\}$ is λ-irreducible.

The next is to show that $\{X_n\}$ is aperiodic. Suppose to the contrary that $\{X_n\}$ is periodic with periodic decomposition $\mathcal{X}_1, \ldots, \mathcal{X}_d$.

Let $x \in \mathcal{X}_1$, then

$$P(x, \mathcal{X}_2) = \int_{y\in\mathcal{X}_2} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy = 1.$$

It follows that

$$\int_{y\in\mathcal{X}_2^{\complement}} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy = 0.$$

Since $0 < \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} < \infty$, we have $\lambda(\mathcal{X}_2^{\complement}) = 0$. Since $\mathcal{X}_1 \subset \mathcal{X}_2^{\complement}$, $\lambda(\mathcal{X}_1) = 0$. Hence, for $x \in \mathcal{X}_d$,

$$P(x, \mathcal{X}_1) = \int_{y\in\mathcal{X}_1} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy = 0,$$

which contradicts to periodicity. Applying Theorem 2.14 gives

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0;$$

more explicitly,

$$\frac{1}{2^{n-1}}x + \frac{1}{2^{n-2}}U_2 + \frac{1}{2^{n-3}}U_3 + \cdots + U_n \to N(0,1),$$

for any $x \in \mathbb{R}$.

We now start to build a small set C. Consider $C \coloneqq \{x : x^2 \le c\}$, where $0 < c < 1$. Then, let

$$\begin{aligned}
\epsilon &\coloneqq \int_{\mathbb{R}} \left(\inf_{x\in C} P(x, dy) \right) \\
&= \int_{\mathbb{R}} \inf_{x\in[-\sqrt{c},\sqrt{c}]} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy \\
&= \int_{-\infty}^{0} \inf_{x\in[-\sqrt{c},\sqrt{c}]} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy + \int_{0}^{\infty} \inf_{x\in[-\sqrt{c},\sqrt{c}]} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} dy \\
&= \int_{-\infty}^{0} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y+\frac{\sqrt{c}}{2})^2} dy + \int_{0}^{\infty} \inf_{x\in[-\sqrt{c},\sqrt{c}]} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{\sqrt{c}}{2})^2} dy > 0.
\end{aligned}$$

Additionally, define $\nu(A) \coloneqq \frac{\int_{y\in A} \inf_{x\in C} P(x,dy)}{\int_{y\in\mathbb{R}} \inf_{x\in C} P(x,dy)} = \frac{\int_{y\in A} \inf_{x\in C} P(x,dy)}{\epsilon}$ for all measurable A. Clearly, ν is a probability measure. Then, for $x \in C$ and all measurable subsets A,

$$P(x, A) = \int_{y\in A} P(x, dy) \ge \int_{y\in A} \left(\inf_{x\in C} P(x, dy) \right) = \epsilon\nu(A),$$

which means C is a small set.

Then, we are going to build the bivariate drift condition based on this small set C. Note that

$$\mathbb{E}[X_{n+1}^2 | X_n = x] = \int y^2 \cdot \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}(y-\frac{x}{2})^2} = \frac{x^2+3}{4}.$$

Let $h(x, y) \coloneqq 1 + x^2 + y^2$ and $\{X_n\}$, $\{X_n'\}$ be two copies of the Markov chain. We have

$$\begin{aligned}
\overline{P}h(x,y) &= \mathbb{E}[h(X_{n+1}, X_{n+1}') | X_n = x, X_n' = y] \\
&= 1 + \mathbb{E}[X_{n+1}^2 | X_n = x] + \mathbb{E}[(X_{n+1}')^2 | X_n' = y] \\
&= 1 + \frac{x^2+3}{4} + \frac{y^2+3}{4} = \frac{h(x,y)}{4} + \frac{9}{4}.
\end{aligned}$$

If $(x,y) \notin C \times C$, then $h(x,y) > 1 + 2c \Rightarrow \frac{h(x,y)}{1+2c} > 1$. It follows that

$$\overline{P}h(x,y) = \frac{h(x,y)}{4} + \frac{9}{4} \leq \frac{h(x,y)}{4} + \frac{9}{4}\frac{h(x,y)}{1+2c} = \frac{(10+2c)h(x,y)}{4+8c}.$$

Since $0 < c < 1$, we have $\alpha \coloneqq \frac{10+2c}{4+8c} > 1$. Therefore, we can apply Theorem 2.36 to get quantitative convergence rates.

However, if the transition probability $P(x,\cdot)$ is instead given by $N(x, \frac{3}{4})$, then the chain does not have a stationary distribution. Suppose to the contrary that $\pi(\cdot)$ is the stationary distribution for $\{X_n\}$ with $P(x,\cdot) = N(x, \frac{3}{4})$. Then,

$$\pi(dx) = \int \pi(dy)P(y,dx),$$

which means

$$X + Y \sim \pi(\cdot), \quad \text{where } X \sim \pi(\cdot) \text{ and } Y \sim N(0, \frac{3}{4}).$$

It follows that for $n \geq 1$,

$$Z_n \coloneqq X + Y_1 + Y_2 + \cdots + Y_n \sim \pi(\cdot), \quad \text{where } X \sim \pi(\cdot) \text{ and } Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} N(0, \frac{3}{4}).$$

Then, for any $a \in \mathbb{R}$,

$$\pi((a,\infty) = \lim_{n\to\infty} \mathbf{P}(Z_n > a) = \frac{1}{2}.$$

This leads to a contradiction since $\lim_{a\to\infty} \pi((a,\infty)) = 0$.

3. Markov Chains Monte Carlo Algorithms

MCMC algorithms are widely used in statistics to sample from complicated probability distributions in high dimensions. They not only offer practical solutions but also raise intriguing questions related to probability theory and the mathematics of Markov chains. In this section, we will explore some classic MCMC algorithms and provide illustrative examples.

3.1. **Motivation.** Let's consider a density function π_u defined over a measure space $(\mathcal{X}, \mathcal{G}, \mu)$. The density function may be unnormalized but satisfies $0 < \int_{\mathcal{X}} \pi_u d\mu < \infty$. Typically, $\mathcal{X}$ represents an open subset of $\mathbb{R}^d$, and the densities are taken with respect to Lebesgue measure, though other settings are also possible. This density function gives rise to a probability measure $\pi(\cdot)$ on $\mathcal{X}$ using the following formula:

$$\pi(A) = \frac{\int_A \pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}, \quad \forall A \in \mathcal{G}.$$

Our objective is to estimate expectations of functions $f : \mathcal{X} \to \mathbb{R}$ with respect to $\pi(\cdot)$, i.e.

$$\pi(f) = \mathbf{E}_\pi[f(X)] = \frac{\int_A f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}.$$

However, when $\mathcal{X}$ is high-dimensional, and π_u is a complicated function, computing the integrals in the above equation, either analytically or numerically, becomes extremely challenging.

The classical Monte Carlo solution to this problem is to simulate i.i.d. random variables $Z_1, Z_2, \ldots, Z_N \sim \pi(\cdot)$, and then estimate $\pi(f)$ by

$$\hat{\pi}(f) \coloneqq \frac{1}{N} \sum_{i=1}^{N} f(Z_i).$$

Note that $\hat{\pi}(f)$ is an unbiased estimator of $\pi(f)$, with standard deviation of order $O\left(\frac{1}{\sqrt{N}}\right)$. Furthermore, if $\pi(f) < \infty$, then by the classical central limit theorem, the error $\hat{\pi}(f) - \pi(f)$ will have a limiting normal distribution. However, if π_u is complicated, it is very difficult to directly simulate i.i.d random variables from $\pi(\cdot)$.

MCMC algorithms provide a solution to the drawbacks of Monte Carlo methods by constructing a Markov chain that can be efficiently run on a computer and has the desired stationary distribution $\pi(\cdot)$. Assuming that the Markov chain satisfies certain conditions (e.g., as stated in Theorem 2.36), the MCMC algorithm is guaranteed to converge after a certain number of iterations. For sufficiently large n, the distribution of X_n will be approximately stationary, i.e. $\mathcal{L}(X_n) \approx \pi(\cdot)$. We can then use $Z_1 = X_n$ as a starting point and restart the Markov chain to generate Z_2, Z_3, and so on. Then, we can use these samples to compute the unbiased estimator $\hat{\pi}(f)$ as in the classical Monte Carlo method.

Remark 3.1. In practice, instead of starting a fresh Markov chain for each new sample, we often take an entire tail of Markov chain to create an estimate such as $\frac{1}{N-B} \sum_{i=B+1}^{N} f(X_i)$, where the *burn-in value* B is chosen to ensure $\mathcal{L}(X_B) \approx \pi(\cdot)$.

However, in this case, the different $f(X_i)$ are not independent, but the estimate can be computed more efficiently. We tend to ignore this issues.

Remark 3.2. MCMC is just one method among various approaches for sampling and estimating from complicated probability distributions, such as "rejection sampling" and "importance sampling". However, other algorithms have limited applicability and are effective only in specific cases.

Next, we discuss one of the most prevalent applications of MCMC algorithms, which is Bayesian statistical inference.

Let $L(\mathbf{y}|\theta)$ be the *likelihood function* (i.e. the density function of data $\mathbf{y}$ given unknown parameters θ) of a statistical model, for $\theta \in \mathcal{X}$. Usually, $\mathcal{X} \subset \mathbb{R}^d$. Let the *prior density* of θ be $p(\theta)$. The (unnormalized) *posterior density* given $\mathbf{y}$ is

$$\pi_u(\theta) \coloneqq L(\mathbf{y}|\theta)p(\theta).$$

The *posterior mean* of any functional f is given by

$$\pi(f) = \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}.$$

MCMC algorithms have proven to be extremely helpful for such Bayesian estimates.

In the next few subsections, we will see that constructing an appropriate Markov chain with a desired stationary distribution is surprisingly straightforward.

3.2. **The Metropolis-Hastings Algorithm.** Suppose again that $\pi(\cdot)$ has a (possibly unnormalized) density π_u. Let $Q(x,\cdot)$ be any easily-simulated Markov chain, whose transition probability has a (possibly unnormalized) density with respect to Lebesgue measure, i.e. $Q(x,dy) \propto q(x,y)dy$.

The Metropolis-Hastings algorithm proceeds as follows:

1. Choose some X_0.
2. Given X_n, generate a *proposal* Y_{n+1} from $Q(X_n,\cdot)$.
3. Flip an independent coin, whose probability of heads equal to $\alpha(X_n, Y_{n+1})$, where
$$\alpha(x,y) \coloneqq \min\left\{1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}\right\}.$$
To avoid ambiguity, we set $\alpha(x,y) = 1$ whenever $\pi(x)q(x,y) = 0$.
4. If the coin is heads, accept the proposal by setting $X_{n+1} = Y_{n+1}$; if the coin is tails, then reject the proposal by setting $X_{n+1} = X_n$.
5. Replace n by $n+1$ and repeat.

To simulate the flipping of an independent coin with a desired probability α, we can achieve this by generating $U_{n+1} \sim \text{Uniform}([0,1])$ and accepting the proposal if $U_{n+1} \leq \alpha$, otherwise rejecting the proposal.

The transition probability of the resulting Markov chain can be expressed as follows:
$$P(x,A) = \int_{y\in A} \alpha(x,y)q(x,dy) + \delta_x(A)\int_{y\in\mathcal{X}} (1-\alpha(x,y)q(x,dy), \quad x\in\mathcal{X}, A\in\mathcal{G}.$$

The derivation of the Metropolis-Hasting algorithms relies on the following proposition:

Proposition 3.3. The Metropolis-Hastings algorithm (as described above) produces a Markov chain $\{X_n\}$ which is reversible with respect to $\pi(\cdot)$.

Proof. Assume $x \neq y$ and set $c = \int_{\mathcal{X}} \pi_u(x)dx$. We have
$$\begin{aligned}\pi(dx)P(x,dy) &= \left[c^{-1}\pi_u(x)dx\right]\left[q(x,y)\alpha(x,y)dy\right] \\ &= c^{-1}\pi_u(x)q(x,y)\min\left\{1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}\right\} \\ &= c^{-1}\min\left\{\pi_u(x)q(x,y), \pi_u(y)q(y,x)\right\},\end{aligned}$$
which is symmetric in x and y. Hence, we have
$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx).$$
□

If follows from Proposition 2.7 that converges to the stationary distribution $\pi(\cdot)$. However, there is no guarantee that starting from any initial state will lead to convergence to the stationary distribution (e.g. Example 2.8 and 2.10); one may additionally require ϕ-irreducibility and aperiodicity (Theorem 2.14). To determine the required number of iterations, one can refer to Section 2.

To run the Metropolis-Hastings algorithm, we only need to compute ratios of densities $\frac{\pi_u(y)}{\pi_u(x)}$ in $\alpha(x,y)$, eliminating the need to calculate the integral $\int_{\mathcal{X}} \pi_u(x)dx$. Hence, with an appropriate $Q(x,\cdot)$, running the algorithm would be quite feasible.

The choice of proposal distributions $Q(x,\cdot)$ is another critical factor to consider. There are several commonly used approaches for choosing the proposal density, including:

- Symmetric Metropolis Algorithm: $q(x,y) = q(y,x)$, and the acceptance probability simplifies to
$$\alpha(x,y) = \min\left\{1, \frac{\pi_u(y)}{\pi_u(x)}\right\}.$$
- Random walk Metropolis-Hastings: $q(x,y) = q(y-x)$. For example, $Q(x,\cdot) = N(x,\sigma^2)$ or $Q(x,\cdot) = \text{Uniform}([x-1, x+1])$.
- Independence sampler: $q(x,y) = q(y)$, i.e. $Q(x,\cdot)$ does not depend on x.
- Langevin algorithm: The proposal is generated by
$$Y_{n+1} \sim N\left(X_n + \frac{\delta}{2}\nabla \log \pi(X_n), \delta\right),$$
- for some (small) $\delta > 0$. This is motivated by a discrete approximation to a Langevin diffusion processes.

3.3. **Combining Chains.** If P_1 and P_2 are two distinct chains, each with stationary distribution $\pi(\cdot)$, then the new chain P_1P_2 also has stationary distribution $\pi(\cdot)$. Therefore, one can make new MCMC algorithms out of old ones, by specifying that new algorithm applies first the chain P_1, followed by the chain P_2, and then repeats the chain P_1, and so on. More generally, it is possible to combine many different chains in this manner. It is important to note that even if each individual chain, P_1 and P_2, is reversible, the resulting combined chain P_1P_2 may not be reversible in general.

3.4. **The Gibbs Sampler.** Suppose again that $\pi_u(\cdot)$ is d-dimensional density, with $\mathcal{X}$ an open subset of $\mathbb{R}^d$, and write $\mathbf{x} = (x_1, \ldots, x_d)$. The *$i$-th component Gibbs sampler* is defined such that P_i leaves all components besides i unchanged, and replaces the i-th component by a draw from the full conditional distribution of $\pi(\cdot)$ conditional on all the other components. More formally, let
$$S_{\mathbf{x},i,a,b} \coloneqq \{y \in \mathcal{X} : y_j = x_j \text{ for } j \neq i, \text{ and } a \leq y_i \leq b\},$$
then
$$P_i(\mathbf{x}, S_{x,i,a,b}) \coloneqq \frac{\int_a^b \pi_u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)dt}{\int_{-\infty}^{\infty} \pi_u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)dt}, \quad a \leq b.$$

Proposition 3.4. P_i is reversible with respect to $\pi(\cdot)$ for any $i \in \{1, \ldots, d\}$.

Proof. For $\mathbf{x}$ and $\mathbf{y}$ such that $y_j = x_j$ for $j \neq i$, we have
$$\begin{aligned}
\pi(d\mathbf{x})P_i(\mathbf{x}, d\mathbf{y}) &= \frac{\pi_u(d\mathbf{x})}{\int_{\mathcal{X}} \pi_u(\mathbf{x})d\mathbf{x}} \frac{\pi_u(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)dy}{\int_{-\infty}^{\infty} \pi_u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)dt} \\
&= \frac{\pi_u(d\mathbf{x})}{\int_{-\infty}^{\infty} \pi_u(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)dt} \frac{\pi_u(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)dy}{\int_{\mathcal{X}} \pi_u(\mathbf{x})d\mathbf{x}} \\
&= \frac{\pi_u(d\mathbf{x})}{\int_{-\infty}^{\infty} \pi_u(y_1, \ldots, x_{i-1}, t, y_{i+1}, \ldots, x_n)dt} \frac{\pi_u(y_1, \ldots, y_{i-1}, y, y_{i+1}, \ldots, x_n)dy}{\int_{\mathcal{X}} \pi_u(\mathbf{x})d\mathbf{x}} \\
&= P_i(\mathbf{y}, d\mathbf{x})\pi(d\mathbf{y}).
\end{aligned}$$

□

It follows that P_i has $\pi(\cdot)$ as its stationary distribution. In fact, P_i can be viewed as a special case of Metropolis-Hastings algorithm with acceptance probability of $\alpha(x, y) = 1$.

Next, we can build the full Gibbs sampler by combining different P_i chains using one of the following approaches:

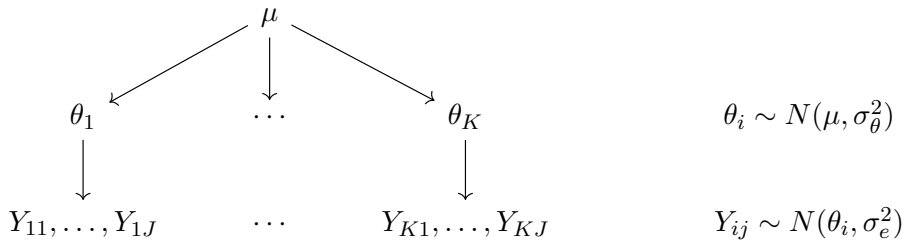
- Deterministic-scan Gibbs sampler: $P = P_1 P_2 \cdots P_d$, that is, it performs the d different Gibbs sampler components, in sequential order.
- Random-scan Gibbs sampler: $P = \frac{1}{d} \sum_{i=1}^{d} P_i$, that is, it does one of the d different Gibbs sampler components, chosen uniformly at random.

Both versions yield an MCMC algorithm with $\pi(\cdot)$ as its stationary distribution. The output of a Gibbs sampler exhibits a "zig-zag pattern," where the components are updated one at a time. Moreover, the random-scan Gibbs sampler is reversible, while the deterministic-scan Gibbs sampler is usually not.

3.5. **Detailed Bayesian Example: Variance Components Model.** The model consists of fixed constants μ_0 and positive constants a_1, b_1, a_2, b_2, and σ_0^2. There are three hyperparameters: σ_θ^2, σ_e^2, and μ, each with priors based on these constants as follows:

$$\sigma_\theta^2 \sim IG(a_1, b_1),$$
$$\sigma_e^2 \sim IG(a_2, b_2),$$
$$\mu \sim N(\mu_0, \sigma_0^2).$$

Additionally, there are K further parameters $\theta_1, \theta_2, \ldots, \theta_K$, which are conditionally independent given the hyperparameters. Specifically, $\theta_i \sim \text{Normal}(\mu, \sigma_\theta^2)$. The data Y_{ij}, where $1 \le i \le K$ and $1 \le j \le J$, is assumed to be distributed as $Y_{ij} \sim \text{Normal}(\theta_i, \sigma_e^2)$, conditionally independently given the parameters. A graphical representation of the model is shown below:



The Bayesian paradigm then involves conditioning on the values of the data $\{Y_{ij}\}$, and considering the joint distribution of all $K + 3$ parameters given this data. That is, we are interested in the distribution

$$\pi(\cdot) = \mathcal{L}(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K | \{Y_{ij}\})$$

defined on the state space $\mathcal{X} = (0, \infty)^2 \times \mathbb{R}^{K+1}$. We would like to sample from this distribution $\pi(\cdot)$. We compute that this distributions' unnormalized density is given by

$$\pi_u(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, \ldots, \theta_K) \propto e^{-\frac{b_1}{\sigma_\theta^2}} (\sigma_\theta^2)^{-a_1-1} e^{-\frac{b_2}{\sigma_e^2}} (\sigma_e^2)^{-a_2-1} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \times \prod_{i=1}^{K} \frac{e^{\frac{-(\theta_i-\mu)^2}{2\sigma_\theta^2}}}{\sigma_\theta} \times \prod_{i=1}^{K} \prod_{j=1}^{J} \frac{e^{\frac{-(Y_{ij}-\theta_i)^2}{2\sigma_e^2}}}{\sigma_e}.$$

The above expression appears to be complicated and high-dimensional. Now, let's focus on constructing MCMC algorithms to sample from the target density π_u. We will start with the Gibbs sampler. In order to run a Gibbs sampler, we need to compute the full conditional distributions by $\frac{\pi(\cdot)}{\int \pi(\cdot)dx}$ (1-dimensional integration) as follows:

$$\mathcal{L}(\sigma_\theta^2|\mu,\sigma_e^2,\theta_1,\ldots,\theta_K,Y_{ij}) = IG\left(a_1+\frac{1}{2}K, b_1+\frac{1}{2}\sum_i(\theta_i-\mu)^2\right),$$

$$\mathcal{L}(\sigma_e^2|\mu,\sigma_\theta^2,\theta_1,\ldots,\theta_K,Y_{ij}) = IG\left(a_2+\frac{1}{2}KJ, b_2+\frac{1}{2}\sum_{i,j}(Y_{ij}-\theta_i)^2\right),$$

$$\mathcal{L}(\mu|\sigma_\theta^2,\sigma_e^2,\theta_1,\ldots,\theta_K,Y_{ij}) = N\left(\frac{\sigma_\theta^2\mu_0+\sigma_0^2\sum_i\theta_i}{\sigma_\theta^2+K\sigma_0^2}, \frac{\sigma_\theta^2\sigma_0^2}{\sigma_\theta^2+K\sigma_0^2}\right),$$

$$\mathcal{L}(\theta_i|\mu,\sigma_\theta^2,\sigma_e^2,\theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_K,Y_{ij}) = N\left(\frac{J\sigma_\theta^2\overline{Y}_i+\sigma_e^2\mu}{J\sigma_\theta^2+\sigma_e^2}, \frac{\sigma_\theta^2\sigma_0^2}{J\sigma_\theta^2+\sigma_e^2}\right),$$

where $\overline{Y}_i = \frac{1}{J}\sum_{j=1}^J Y_{ij}$, and the last equation holds for $1 \le i \le K$.

The Gibbs sampler then proceeds by updating the $K+3$ variables in turn (either deterministic- or random-scan), according to the conditional distributions. This process is feasible because the conditional distributions can be easily simulated.

Alternatively, we can run a Metropolis-Hastings algorithm to sample from this model. For example, we can choose a symmetric random-walk Metropolis algorithm with proposals of the form $N(X_n, \sigma^2 I_{K+1})$ for some $\sigma^2 > 0$. This algorithm can be effective for this model as long as the value of σ^2 is appropriately selected.

4. Convergence Proofs using Coupling Constructions

In this section, we present proofs for some of the theorems mentioned earlier using the method of coupling, which appears to be particularly effective for analyzing MCMC algorithms on general state spaces. This method offers a more concise alternative to the other lengthy analytical arguments found in the literature.

4.1. **The Coupling Inequality.**

Consider two random variables, X and Y, defined jointly on a space $\mathcal{X}$. Let $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ be their respective probability distributions. Then, we have

$$\begin{aligned}
\|\mathcal{L}(X)-\mathcal{L}(Y)\| =& \sup_A |\mathbf{P}(X\in A)-\mathbf{P}(Y\in A)| \\
=& \sup_A |[\mathbf{P}(X\in A, X=Y)+\mathbf{P}(X\in A, X\neq Y)] \\
& -[\mathbf{P}(Y\in A, Y+X)+\mathbf{P}(Y\in A, Y\neq X)]| \\
=& \sup_A |\mathbf{P}(X\in A, X\neq Y)-\mathbf{P}(Y\in A, Y\neq X)| \\
\le& \mathbf{P}(X\neq Y).
\end{aligned}$$

In short, we have the relationship

$$\|\mathcal{L}(X)-\mathcal{L}(Y)\| \le \mathbf{P}(X\neq Y) \tag{4.11}$$

The fundamental concept of coupling relies on the above relationship. The goal is to construct two random variables: one updates according to the transition probability and another follows the stationary distribution. Then, we can demonstrate desired convergence results using Equation (4.11).

4.2. **The Coupling Construction.** Suppose that C is a small set. Consider the following coupling construction.

Begin with $X_0 = x$ and $X_0' \sim \pi(\cdot)$, and set $n = 0$. Then, repeat the following loop forever.

Beginning of the loop. Given X_n and X_n':

(1) If $X_n = X_n'$, choose $X_{n+1} = X_{n+1}' \sim P(X_n, \cdot)$, and replace n by $n+1$.
(2) Else, if $(X_n, X_n') \in C \times C$, then:
   (a) With probability ϵ, choose $X_{n+n_0} = X_{n+n_0}' \sim \nu(\cdot)$;
   (b) Else, with probability $1-\epsilon$, conditionally independently choose

$$X_{n+n_0} \sim \frac{1}{1-\epsilon}[P^{n_0}(X_n, \cdot) - \epsilon\nu(\cdot)]$$

$$X_{n+n_0}' \sim \frac{1}{1-\epsilon}[P^{n_0}(X_n', \cdot) - \epsilon\nu(\cdot)]$$

   In the case $n_0 > 1$, for completeness, we go back and construct $X_n + 1, \ldots, X_{n+n_0-1}$ from their correct conditional distributions given X_0 and X_{n+n_0}, and similarly (and conditionally independently) construct $X_{n+1}', \ldots, X_{n+n_0-1}'$ from their correct conditional distributions given X_n' and X_{n+n_0}'.
   Replace n by $n+n_0$.
(3) Else, conditionally independently choose $X_{n+1} \sim P(X_n, \cdot)$ and $X_{n+1}' \sim P(X_n', \cdot)$.

Then return to the beginning of the loop.

We then check the marginal distributions of X_n and X_n' are correctly updates:

Given that $X_0 = x$ and $X_0' \sim \pi(\cdot)$, it is straightforward to check that conditions (1) and (3) ensure that the two chains marginally follow the correct distributions ($P(X_n, \cdot)$ and $\pi(\cdot)$, respectively.

For condition (2), when $(X_n, X_n') \in C \times C$, we have

$$X_{n+n_0} \sim \epsilon \cdot \nu(\cdot) + (1-\epsilon) \cdot \frac{1}{1-\epsilon}[P^{n_0}(X_n, \cdot) - \epsilon\nu(\cdot)] = P^{n_0}(X_n, \cdot),$$

and

$$X_{n+n_0}' \sim \epsilon \cdot \nu(\cdot) + (1-\epsilon) \cdot \frac{1}{1-\epsilon}[P^{n_0}(X_n', \cdot) - \epsilon\nu(\cdot)] = P^{n_0}(X_n', \cdot).$$

Therefore, $P(X_n \in A) = P^n(x, A)$ and $P(X_n' \in A) = \pi(A)$ (since X_0' starts from the stationary distribution) for all n.

For the intermediate steps $X_{n+1}, \ldots, X_{n+n_0-1}$, we update inductively by

$$\mathbf{P}(X_{n+i} \in A | X_{n+i-1} = b, X_{n+n_0} = c) = \int_A P(b, dx) P^{n_0-i}(x, c) \quad \forall 1 \le i \le n_0 - 1.$$

4.3. **Proof of Theorem 2.26.**

Theorem 2.26. Consider a Markov chain with stationary probability distribution $\pi(\cdot)$. Suppose the minorisation condition (2.7) is satisfied for some $n_0 \in \mathbb{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, in the special case $C = \mathcal{X}$ (i.e. the entire state space is small). Then the chain is uniformly ergodic, and in fact $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \mathcal{X}$.

Proof. In this theorem, the small set $C = \mathcal{X}$; hence, we only need to consider the second step in the coupling construction. Obviously, every n_0 steps, we have probability at least ϵ of coupling (making $X_n = X'_n$). Consider $n = n_0 m$ where $n_0, m \in \mathbb{N}$,

$$\mathbf{P}(X_n \neq X'_n) \leq (1-\epsilon)^m.$$

It follows from the coupling inequality (4.11) that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^m = (1-\epsilon)^{\frac{n}{n_0}}.$$

Generally, by Proposition 2.13 (3), we have

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\left\lfloor \frac{n}{n_0} \right\rfloor} \quad \forall n \in \mathbb{N}.$$

□

4.4. **Proof of Theorem 2.36.** Prior to proving Theorem 2.31, let us first take a look at Theorem 2.36 since we will need this results to establish Theorem 2.31.

Theorem 2.36. Consider a Markov chain on a state space $\mathcal{X}$, having transition kernel P. Suppose there is $C \subset \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, a probability distribution $\nu(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, $n_0 \in \mathbb{N}$, and $\epsilon > 0$, such that the minorisation condition (2.7) and bivariate drift condition (2.9). Define B_{n_0} by (2.10). Then for any joint initial distribution $\mathcal{L}(X_0, X'_0)$, and any integers $1 \leq j \leq k$, if $\{X_n\}$ and $\{X'_n\}$ are two copies of the Markov chain started in the joint initial distribution $\mathcal{L}(X_0, X'_0)$, then

$$\left\|\mathcal{L}(X_k) - \mathcal{L}(X'_k)\right\| \leq (1-\epsilon)^j + \alpha^{-k}(B_{n_0})^{j-1}\mathbf{E}[h(X_0, X'_0)].$$

In particular, by choosing $j = \lfloor rk \rfloor$ for sufficiently small $r > 0$, we obtain an explicit, quantitative convergence bound which goes to 0 exponentially quickly as $k \to \infty$.

Proof of Theorem 2.36. We first consider the case $n_0 = 1$ in the minorisation condition for the small set C. In this case, we write B_{n_0} as B.

Let

$$N_k \coloneqq \#\{m : 0 \leq m \leq k, (X_m, X'_m) \in C \times C\},$$

and let $\tau_1, \tau_2, \ldots$ be the times of the successive visits of $\{(X_n, X'_n)\}$ to $C \times C$. Then for any integer j with $1 \leq j \leq k$,

$$\mathbf{P}(X_k \neq X'_k) = \mathbf{P}(X_k \neq X'_k, N_{k-1} \geq j) + \mathbf{P}(X_k \neq X'_k, N_{k-1} < j). \tag{4.12}$$

Then, we are going to bound the probability that X_n and X'_n are unequal by bounding the two terms on the right hand side of the above equation.

Notice that the event $(X_k \neq X'_k, N_{k-1} \geq j)$ implies that after at least j times entering $C \times C$ before the $k-1$-th step (including $k-1$), the Markov chains have not successfully coupled. This indicates, whenever the Markov chains enter $C \times C$ before the $k-1$-the

step, the chains are updated using (b) of condition 2 in the coupling construction. Hence, we have

$$\mathbf{P}(X_k \neq X'_k, N_{k-1} \geq j) \leq (1-\epsilon)^j.$$

For the second term in (4.12), let

$$M_k := \alpha^k B^{-N_{k-1}} h(X_k, X'_k)\mathbf{1}(X_k \neq X'_k), \quad k = 0, 1, \ldots$$

(where $N_{-1} = 0$). We want to prove $\{M_k\}$ is supermartingale.

Lemma 4.1. We have

$$\mathbf{E}[M_{k+1}|X_0, \ldots, X_k, X'_0, \ldots, X'_k] \leq M_k,$$

i.e. $\{M_k\}$ is a supermartingale.

Proof of Lemma 4.1. If $(X_k, X'_k) \notin C \times C$, then $N_k = N_{k-1}$. It follows that

$$\mathbf{E}[M_{k+1}|X_0, \ldots, X_k, X'_0, \ldots, X'_k] = \alpha^{k+1} B^{-N_{k-1}} \mathbf{E}[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_{k+1} \neq X'_{k+1})|X_k, X'_k]$$

Notice that $\mathbf{1}(X_{k+1} \neq X'_{k+1}) \leq \mathbf{1}(X_k \neq X'_k)$ since $X_{k+1} \neq X'_{k+1}$ for $(X_k, X'_k) \notin C \times C$ implies $X_k \neq X'_k$. We have

$$\begin{aligned}
\mathbf{E}[M_{k+1}|X_0, \ldots, X_k, X'_0, \ldots, X'_k] &\leq \alpha^{k+1} B^{-N_{k-1}} \mathbf{E}[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_k \neq X'_k)|X_k, X'_k] \\
&= \alpha^{k+1} B^{-N_{k-1}} \mathbf{E}[h(X_{k+1}, X'_{k+1})|X_k, X'_k]\mathbf{1}(X_k \neq X'_k) \\
&= M_k \alpha \frac{\mathbf{E}[h(X_{k+1}, X'_{k+1})|X_k, X'_k]}{h(X_k, X'_k)} \\
&\leq M_k
\end{aligned}$$

by the bivariate drift condition (2.9).

If $(X_k, X'_k) \in C \times C$, then $N_k = N_{k-1} + 1$. If $X_k = X'_k$, we have $M_{k+1} = M_k = 0$ and then the supermartingale inequality holds trivially. Assume $X_k \neq X'_k$, we have

$$\begin{aligned}
\mathbf{E}[M_{k+1}|X_0, \ldots, X_k, X'_0, \ldots, X'_k] &= \alpha^{k+1} B^{-N_{k-1}-1} \mathbf{E}[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_{k+1} \neq X'_{k+1})|X_k, X'_k] \\
&= \alpha^{k+1} B^{-N_{k-1}-1}(1-\epsilon)(\bar{R}h)(X_k, X'_k) \\
&= M_k \frac{\alpha(1-\epsilon)(\bar{R}h)(X_k, X'_k)}{h(X_k, X'_k)B}; \\
&\leq M_k
\end{aligned}$$

by the definition of B and the fact $h(x, y) \geq 1$.

Therefore, $\{M_k\}$ is a supermartingale. □

Since $B \geq 1$,

$$\begin{aligned}
\mathbf{P}(X_k \neq X'_k, N_{k-1} < j) &= \mathbf{P}(X_k \neq X'_k, N_{k-1} \leq j-1) \\
&\leq \mathbf{P}(X_k \neq X'_k, B^{-N_k-1} \geq B^{-(j-1)}) \\
&= \mathbf{P}(\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}} \geq B^{-(j-1)}) \\
\text{(by the Markov's inequality)} \quad &\leq B^{j-1}\mathbf{E}[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}}] \\
(h(x,y) \geq 1) \quad &\leq B^{j-1}\mathbf{E}[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}}h(X_k, X'_k)] \\
\text{(by defintion)} \quad &= \alpha^{-k}B^{j-1}\mathbf{E}[M_k] \\
\text{(by Lemma 4.1)} \quad &\leq \alpha^{-k}B^{j-1}\mathbf{E}[M_0] \\
\text{(by defintion)} \quad &= \alpha^{-k}B^{j-1}\mathbf{E}[h(X_0, X'_0)]
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|P^n(x,\cdot) - \pi(\cdot)\| &\leq \mathbf{P}(X_n \neq X'_n) \\
\text{(by 4.12)} \quad &= \mathbf{P}(X_n \neq X'_n, N_{n-1} \geq j) + \mathbf{P}(X_n \neq X'_n, N_{n-1} < j) \\
&\leq (1-\epsilon)^j + \alpha^{-n}B^{j-1}\mathbf{E}[h(X_0, X'_0)],
\end{aligned}$$

which prove Theorem 2.36 in the case where $n_0 = 1$.

Finally, we consider the case that $n_0 > 1$. In this case, we do not count the visits to $C \times C$ corresponding to the "filling in " times for going back and constructing $X_{n+1}, \ldots, X_{n+n_0-1}$ in condition 2 of the coupling construction. Instead, we define N_k as the number of visits to $C \times C$, and τ_i as the actual visit times, excluding any "filling in" times. Moreover, replace N_{k-1} with N_{k-n_0} in (4.12) and the definition of $\{M_k\}$. Let $t(k)$ be the latest time smaller than k which does not correspond to a "filling in" time.

Lemma 4.2. $\{M_{t(k)}\}$ is a supermartingale, where $M_{t(k)} = \alpha^{t(k)}B^{-N_{t(k)-n_0}}h(X_{t(k)}, X'_{t(k)})\mathbf{1}(X_{t(k)} \neq X'_{t(k)})$.

Proof of Lemma 4.2. If $(X_{t(k)}, X'_{t(k)}) \notin C \times C$, then $N_k = N_{k-n_0}$ and $t(k) = k$. In this case, the proof follows exactly the same steps as in Lemma 4.1.

If $(X_{t(k)}, X'_{t(k)}) \notin C \times C$, we assume $X_{t(k)} \neq X'_{t(k)}$ (otherwise it is trivial). Then $N_{t(k)} = N_{t(k)-n_0} + 1$ and the rest of the proof follows as in Lemma 4.1. □

Since $B_{n_0} \geq 1$,

$$\begin{aligned}
\mathbf{P}(X_{t(k)} \neq X'_{t(k)}, N_{t(k)-n_0} < j) &= \mathbf{P}(X_{t(k)} \neq X'_{t(k)}, N_{t(k)-n_0} \leq j-1) \\
&\leq \mathbf{P}(X_{t(k)} \neq X'_{t(k)}, B_{n_0}^{-N_{t(k)}-n_0} \geq B_{n_0}^{-(j-1)}) \\
&= \mathbf{P}(\mathbf{1}(X_{t(k)} \neq X'_{t(k)}) B_{n_0}^{-N_{t(k)}-n_0} \geq B_{n_0}^{-(j-1)}) \\
\text{(by the Markov's inequality)} \quad &\leq B_{n_0}^{j-1} \mathbf{E}[\mathbf{1}(X_{t(k)} \neq X'_{t(k)}) B_{n_0}^{-N_{t(k)}-n_0}] \\
(h(x,y) \geq 1) \quad &\leq B_{n_0}^{j-1} \mathbf{E}[\mathbf{1}(X_{t(k)} \neq X'_{t(k)}) B_{n_0}^{-N_{t(k)}-n_0} h(X_k, X'_k)] \\
\text{(by defintion)} \quad &= \alpha^{-t(k)} B_{n_0}^{j-1} \mathbf{E}[M_{t(k)}] \\
\text{(by Lemma 4.1)} \quad &\leq \alpha^{-t(k)} B_{n_0}^{j-1} \mathbf{E}[M_0] \\
\text{(by defintion)} \quad &= \alpha^{-t(k)} B_{n_0}^{j-1} \mathbf{E}[h(X_0, X'_0)]
\end{aligned}$$

Combining the above inequality with (4.12) gives Theorem 2.36. □

4.5. **Proof of Theorem 2.31.** The proof of Theorem 2.31 relies on the previous theorem (Theorem 2.36). Let's first recall 2.31:

Theorem 2.31. Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Suppose that minorisation condition 2.7 is satisfied for some $C \subset \mathcal{X}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose further that the drift condition 2.8 is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty]$ with $V(X) < \infty$ for at least one $x \in \mathcal{X}$ (and hence for π-a.e.) $x \in \mathcal{X}$. Then, the chain is geometrically ergodic.

Proof of Theorem 2.31. We may assume WLOG that

$$\sup_{x \in C} V(x) < \infty \tag{4.13}$$

due to the following lemma:

Lemma 4.3. Given a small set C and drift function V satisfying (2.7) and (2.8), we can find a small set $C_0 \subset C$ such that (2.7) and (2.8) still hold (with the same n_0, ϵ and b, but with λ replaced by some $\lambda_0 < 1$), and such that (4.13) also holds.

Proof of Lemma 4.3. Consider λ and b as defined in (2.8). Choose δ such that $0 < \delta < 1 - \lambda$. Set $\lambda_0 = 1 - \delta$ and $K = \frac{b}{1-\lambda-\delta}$, then define C_0 as follows:

$$C_0 := C \cap \{x \in \mathcal{X} : V(x) \leq K\}.$$

Since $C_0 \subset C$, (2.7) continues to hold over C_0.

Then, we are going to check if (2.8) holds with C replaced by C_0 and λ replaced by λ_0. For $x \in C_0$, since $\delta < 1 - \lambda$, we have

$$(PV)(x) \leq \lambda V(x) + b \leq (1-\delta)V(x) + b = \lambda_0 V(x) + b.$$

Similarly, for $x \notin C$, we have

$$PV(x) \leq \lambda V(x) \leq (1-\delta)V(x) = \lambda_0 V(x).$$

Finally, we consider the case $x \in C \backslash C_0$. Since $V(x) \geq K$, we have

$$\begin{aligned}(PV)(x) &\leq \lambda V(x) + b \\ &= (1-\delta)V(x) - (1-\lambda-\delta)V(x) + b \\ &\leq (1-\delta)V(x) - (1-\lambda-\delta)K + b \\ &= (1-\delta)V(x) \\ &= \lambda_0 V(x),\end{aligned}$$

which shows the drift condition (2.8) holds with C replaced by C_0 and λ replaced by λ_0. □

(4.13) together with (2.8) implies

$$\sup_{(x,y)\in C\times C} \overline{R}h(x,y) < \infty,$$

which also ensures that the quantity B_{n_0} is finite.

Let $h(x,y) = \frac{1}{2}[V(x)+V(y)]$ and $d := \inf_{x\in C^{\complement}} V(x)$. Then, if $d > \frac{b}{1-\lambda} - 1$, the bivariate drift condition will hold by Proposition 2.35. In this case, applying Theorem 2.36 proves Theorem 2.31.

If $d \leq \frac{b}{1-\lambda} - 1$, then the previous argument is not applicable. Our strategy involves enlarging the set C in a way that the updated value of d meets the condition $d > \frac{b}{1-\lambda} - 1$, and then using aperiodicity to demonstrate that this modified C remains a small set, i.e. satisfies (2.7) with possibly larger n_0 and smaller ϵ. Then, applying Theorem 2.36 again proves Theorem 2.31.

To proceed, we choose any $d' > \frac{b}{1-\lambda} - 1$. Let $S := \{x \in \mathcal{X} : V(x) \leq d'\}$ and $C' := C \cup S$. Then, $\inf_{x\in C'^{\complement}} V(x) > d' > \frac{b}{1-\lambda} - 1$. Hence, by Proposition 2.35, the bivariate drift condition hold. It remains to show that C' is a small set:

Lemma 4.4. $C'$is a small set.

To prove the above lemma, we need to introduce the concept of "petite set," as follows:

Definition 4.5 (Petite Set)**.** A subset $C \subset \mathcal{X}$ is *petite* (or (n_0, ϵ, ν)-*petite*), relative to a Markov chain P, if there exists a positive integer $n_0, \epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that

$$\sum_{i=1}^{n_0} P^i(x,\cdot) \geq \epsilon\nu(\cdot), \quad x \in C.$$

A petite set allows states in C to cover the minorisation measure $\epsilon\nu(\cdot)$ at different times i, while a small set restricts states in C to cover $\epsilon\nu(\cdot)$ at a single, specific time. Obviously, any small set is petite. However, the opposite does not hold true in general since a petite set does not rule out periodic behaviour of the chain. For example, some of the states $x \in C$ cover $\epsilon\nu(\cdot)$ only at odd times, and other only at even times. Nevertheless, we can establish the following lemma:

Lemma 4.6. For an aperiodic, ϕ-irreducible Markov chain, all petite sets are small sets

The proof is provided in Appendix A.

With the help of Theorem 4.6, we can prove Lemma 4.4.

Proof of Lemma 4.4. Choose N large enough that $r := 1 - \lambda^N d' > 0$. Define $\tau_C := \inf\{n \geq 1 : X_n \in C\}$ as the first return time to C, and let $Z_n := \lambda^{-n} V(X_n)$ and $W_n := Z_{\min(n,\tau_C)}$. Then, the univariate drift condition implies that W_n is a supermartingale; indeed, if $\tau_C \leq n$, then

$$\mathbf{E}[W_{n+1}|X_0, X_1, \ldots, X_n] = \mathbf{E}[Z_{\tau_C}|X_0, X_1, \ldots, X_n] = Z_{\tau_C} = W_n$$

while if $\tau_C > n$, then $X_n \notin C$ and the univariate drift condition gives

$$\begin{aligned}\mathbf{E}[W_{n+1}|X_0, X_1, \ldots, X_n] &= \lambda^{-(n+1)}(PV)(X_n) \\ &\leq \lambda^{-(n+1)}\lambda V(X_n) \\ &= \lambda^{-n} V(X_n) \\ &= W_n.\end{aligned}$$

Hence, for $x \in S$, using Markov's inequality and the fact that $V \geq 1$,

$$\begin{aligned}\mathbf{P}(\tau_C \geq N|X_0 = x] &= \mathbf{P}(\lambda^{-\tau_C} \geq \lambda^{-N}|X_0 = x) \\ \text{(Markov's inequality)} \quad &\leq \lambda^N \mathbf{E}[\lambda^{-\tau_C}|X_0 = x] \\ (V(x) \geq 1) \quad &\leq \lambda^N \mathbf{E}[\lambda^{-\tau_C} V(X_{\tau_C})|X_0 = x] \\ \text{(by definition)} \quad &= \lambda^N \mathbf{E}[W_{\tau_C}|X_0 = x] \\ (W_n \text{ is a supermartingale}) \quad &\leq \lambda^N \mathbf{E}[W_0|X_0 = x] \\ &= \lambda^N V(x) \\ &\leq \lambda^N d'.\end{aligned}$$

It follows that

$$\mathbf{P}(\tau_C < N|X_0 = x) \geq r.$$

On the other hand, since C is $(n_0, \epsilon, \nu(\cdot))$-small,

$$P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot), \quad \forall x \in C.$$

Then,

$$\begin{aligned}\sum_{i=1}^{N+n_0} P^i(x, \cdot) &\geq \sum_{i=1+n_0}^{N+n_0} P^i(x, \cdot) \\ &= \sum_{i=1}^{N} P^{i+n_0}(x, \cdot) \\ &\geq \sum_{i=1}^{N} \int_{y \in C} P^i(x, dy) P^{n_0}(y, \cdot) \\ &\geq \int_{y \in C} \sum_{i=1}^{N} P^i(x, dy) \epsilon\nu(\cdot) \\ &= \mathbf{P}(\tau_C \leq N|X_0 = x)\epsilon\nu(\cdot) \\ &\geq r\epsilon\nu(\cdot)\end{aligned}$$

Therefore, $S \cup C$ is petite. Then, by Lemma 4.6, $C' = S \cup C$ is small. □

Since the minorisation condition and bivariate drift condition holds for C', applying Theorem 2.36 proves Theorem 2.31.

□

5. Proof of Theorem 2.14

Theorem 2.14. Let $X = \{X_1, \dots, \}$ be a Markov chain on a state space $\mathcal{X}$ with countably generated σ-algebra $\mathcal{G}$. If X is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. $x \in \mathcal{X}$,

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0,$$

where $\|\cdot\|$ is the total variation distance.

In particular, $\lim_{n\to\infty} P^n(x, A) = \pi(A)$ for all measurable $A \subset \mathcal{X}$.

By the previous proofs, we can see that the coupling construction is particularly effective for handling small sets. However, the above theorem does not assume the existence of any small set. As a result, we need the following result about the existence of small sets:

Theorem 5.1. Every ϕ-irreducible Markov chain, on a state space with countably generated σ-algebra, contains a small set $C \subset \mathcal{X}$ with $\phi(C) > 0$. (In fact, each $B \subset \mathcal{X}$ with $\phi(B) > 0$ in turn contains a small set $C \subset B$ with $\phi(C) > 0$.) Furthermore, the minorisation measure $\nu(\cdot)$ may be taken to satisfy $\nu(C) > 0$.

The key idea in proving Theorem 2.14 is to show that (X_n, X'_n) will hit $C \times C$ infinitely often, which means they will have infinitely many opportunities to couple, with probability $\geq \epsilon > 0$ of coupling each time. As a result, they will eventually couple with probability 1, which proves Theorem 2.14.

Proof of Theorem 2.14. Let us start by presenting the following lemma about return probabilities:

Lemma 5.2. Consider a Markov chain on a state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$. Suppose that for some $A \subset \mathcal{X}$, we have $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$. Then for π-almost everywhere $x \in \mathcal{X}$, $\mathbf{P}_x(\tau_A < \infty) = 1$

Proof of Lemma 5.2. Suppose, for the sake of contradiction, that the stated conclusion does not hold, i.e.

$$\pi(\{x \in \mathcal{X} : \mathbf{P}_x(\tau_A = \infty) > 0\}) > 0. \tag{5.14}$$

Then, we present the following claims (proved later):

Claim 1. Condition (5.14) implies that there are constants $l, l_0 \in \mathbb{N}, \delta > 0$, and $B \subset \mathcal{X}$ with $\pi(B) > 0$, such that

$$\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1 : X_{kl_0} \in B\} < l) \geq \delta, \quad x \in B.$$

Proof of Claim 1. By (5.14), we can find δ_1 and a subset $B_1 \subset \mathcal{X}$ with $\pi(B_1) > 0$ such that $\mathbf{P}_x(\tau_A < \infty) \leq 1 - \delta_1$ for all $x \in B_1$ (Trivially, we can take $\{x \in \mathcal{X} : \mathbf{P}_x(\tau_A = \infty) > 0\}$ as B_1).

On the other hand, given that $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$, we can find $l_0 \in \mathbf{N}$, δ_2 and $B_2 \subset B_1$, such that $\pi(B_2) > 0$ and $P^{l_0}(x, A) \geq \delta_2$ for all $x \in B_2$.

Let $\eta \coloneqq \#\{k \geq 1 : X_{kl_0} \in B_2\}$. Recall that for any $x \in B_2$, we have $\mathbf{P}_x(\tau_A = \infty) \leq 1 - P^{l_0}(x, A) \leq 1 - \delta_2$. This means, once the Markov chain enters the states within B_2, the probability of the event $\{\tau_A = \infty\}$ occurring becomes less than $1 - \delta_2$. If the chain enters B_2 for r times, the probability of $\{\tau_A = \infty\}$ happening would then be less than $(1 - \delta_2)^r$, i.e. for any $r \in \mathbb{N}$ and $x \in \mathcal{X}$, we have

$$\mathbf{P}_x(\tau_A = \infty, \eta = r) \leq (1 - \delta_2)^r.$$

In particular,

$$\mathbf{P}_x(\tau_A = \infty, \eta = \infty) = 0.$$

Hence, for $x \in B_2$, we have

$$\mathbf{P}_x(\tau_A = \infty, \eta < \infty) = 1 - \mathbf{P}_x(\tau_A = \infty, \eta = \infty) - \mathbf{P}_x(\tau_A < \infty) \geq 1 - 0 - (1 - \delta_1) = \delta_1.$$

If $\eta < \infty$, then $\sup\{k \geq 1 : X_{kl_0} \in B_2\}$ is finite. It follows that there is $l \in \mathbb{N}, \delta > 0$, and $B \subset B_2$ with $\pi(B) > 0$ such that

$$\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1 : X_{kl_0} \in B_2\} < l) \geq \delta, \quad \forall x \in B.$$

Finally, since $B \subset B_2$, we have

$$\sup\{k \geq 1 : X_{kl_0} \in B_2\} \geq \sup\{k \geq 1 : X_{kl_0} \in B\},$$

which gives the desired result

$$\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1 : X_{kl_0} \in B\} < l) \geq \delta, \quad x \in B.$$

□

Claim 2. Let B, l, l_0, and δ be as in Claim 1. Let $L \coloneqq ll_0$, and $S \coloneqq \sup\{k \geq 1 : X_{kL} \in B\}$, using the convention that $S = -\infty$ if the set $\{k \geq 1 : X_{kL} \in B\}$ is empty. Then, for all integers $1 \leq r \leq j$,

$$\int_{x \in \mathcal{X}} \pi(dx) \mathbf{P}_x(S = r, X_{jL} \notin A) \geq \pi(B)\delta.$$

Proof of Claim 2. We have

$$\begin{aligned}
\int_{x \in \mathcal{X}} \pi(dx) \mathbf{P}_x(S = r, X_{jL} \notin A) &= \int_{x \in \mathcal{X}} \pi(dx) \int_{y \in B} P^{rL}(x, dy) \mathbf{P}_y(S = -\infty, X_{(j-r)L} \notin A) \\
&= \int_{y \in B} \int_{x \in \mathcal{X}} \pi(dx) P^{rL}(x, dy) \mathbf{P}_y(S = -\infty, X_{(j-r)L} \notin A) \\
\text{(By stationarity)} \qquad &= \int_{y \in B} \pi(dy) \mathbf{P}_y(S = -\infty, X_{(j-r)L} \notin A) \\
(\{\sup\{k \geq 1 : X_{kl_0} \in B\} < l\} \subset \{S = -\infty\}, \{\tau_A = \infty\} \subset \{X_{(j-r)L} \notin A\}) \qquad &\geq \int_{y \in B} \pi(dy) \mathbf{P}_y(\tau_A = \infty, \sup\{k \geq 1 : X_{kl_0} \in B\} < l) \\
\text{(By Claim 1)} \qquad &\geq \int_{y \in B} \pi(dy)\delta \\
&= \pi(B)\delta
\end{aligned}$$

□

With these two claims, we are ready to complete the proof of this lemma. By stationarity, for any $j \in \mathbb{N}$, we have

$$\begin{aligned}
\pi(A^{\complement}) &= \int_{x\in\mathcal{X}} \pi(dx) P^{jL}(x, A^{\complement}) \\
&= \int_{x\in\mathcal{X}} \pi(dx) \mathbf{P}_x(X_{jL} \notin A) \\
&\geq \sum_{r=1}^{j} \int_{x\in\mathcal{X}} \pi(dx) \mathbf{P}_x(S = r, X_{jL} \notin A) \\
\text{(by Claim 2)} \qquad &\geq \sum_{r=1}^{j} \pi(B)\delta \\
&= j\pi(B)\delta
\end{aligned}$$

For $j > \frac{1}{\pi(B)\delta}$, the above inequality gives $\pi(A^{\complement}) > 1$, which is impossible. This gives a contradiction and proves Lemma 5.2. □

Next, let us first take a small set C as in Theorem 5.1. Returning to the coupling construction (X_n, Y_n), consider the set $G \subset \mathcal{X} \times \mathcal{X}$, which consists of pairs (x, y) for which $\mathbf{P}_{(x,y)}(\exists n \geq 1 : X_n = Y_n)$ is satisfied. By the coupling construction, if $(X_0, X_0') \coloneqq (x, X_0') \in G$, then $\lim_{n\to\infty} \mathbf{P}(X_n = X_n') = 1$. It follows from the coupling inequality (4.11) that

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\| = 0,$$

which proves Theorem 2.14. It remains to show that

$$\mathbf{P}((x, X_0') \in G) = 1, \quad \forall \pi\text{-a.e. } x \in \mathcal{X}.$$

Let $G_x \coloneqq \{y \in \mathcal{X} : (x, y) \in G\}$ for $x \in \mathcal{X}$ and $\overline{G} \coloneqq \{x \in \mathcal{X} : \pi(G_x) = 1\}$. It suffices to prove the following lemma:

Lemma 5.3. $\pi(\overline{G}) = 1$.

Proof. To begin, we are going to show that $(\pi \times \pi)(G) = 1$. Since $\nu(C) > 0$, as proved in Theorem 5.1, it follows from the minorisation condition and Lemma A.1 that, for any $(x, y) \in \mathcal{X} \times \mathcal{X}$, the joint chain has a positive probability of eventually entering $C \times C$. According to Lemma 5.2, the joint chain will return to $C \times C$ with probability 1 for π-a.e. Once the joint chain reaches $C \times C$, then it will update from $\overline{R}$, which is absolutely continuous with respect to $\pi \times \pi$, if the joint chain is not coupled. Again, by Lemma 5.2, the chain will return to $C \times C$ with probability 1. As a result, the chain will repeatedly revisit $C \times C$ with probability 1, until such a time that $X_n = X_n'$. By the coupling construction, each time of the joint chain being in $C \times C$ yields a probability of at least ϵ for $X_n = X_n'$. Consequently, we will eventually reach $X_n = X_n'$, thereby showing that $(\pi \times \pi)(G) = 1$.

Assume that $\pi(\overline{G}) < 1$. Then, since $\pi(G_x) < 1$ on $\overline{G}^{\complement}$,

$$(\pi \times \pi)(G^{\complement}) = \int_{\mathcal{X}} \pi(dx)\pi(G_x^{\complement}) \geq \int_{\overline{G}^{\complement}} \pi(dx)(1 - \pi(G_x) > 0,$$

contradicting the fact that $(\pi \times \pi)(G) = 1$.

□

□

6. Central Limit Theorems for Markov Chains

Let's consider a Markov chain $\{X_n\}$ on a state space $\mathcal{X}$, which is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$. We start the chain from stationarity, i.e. X_0 follows the distribution $\pi(\cdot)$. Additionally, let $h : \mathcal{X} \to \mathbf{R}$ be a functional with a finite stationary mean denoted as $\pi(h) \coloneqq \int_{x\in\mathcal{X}} h(x)\pi(dx)$.

Definition 6.1 (Central Limit Theorem)**.** If there is some $\sigma^2 < \infty$ such that the normalized sum $n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X_i) - \pi(h)]$ converges weakly to a $N(0, \sigma^2)$ distribution, then h is said to satisfy a *Central Limit Theorem* (*CLT*, or $\sqrt{n}$*-CLT*).

We allow for the special case $\sigma^2 = 0$, corresponding to the constant 0.

Under the assumptions of reversibility or uniform integrability, we find that

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \mathbf{E}\left[\left(\sum_{i=1}^{n} [h(X_i) - \pi(h)]\right)^2\right], \tag{6.15}$$

and also $\sigma^2 = \tau \mathrm{Var}_\pi(h)$, where $\tau \coloneqq \sum_{k\in\mathbf{Z}} \mathrm{Corr}(h(X_0), h(X_k))$ represents the *integrated autocorrelation time*. In the reversible case, this is also related to spectral measures; however, we will not discuss this topic in this paper. It is evident that $\sigma^2 < \infty$ requires that $\mathbf{Var}_\pi(h) < \infty$, i.e. $\pi(h^2) < \infty$.

These Central Limit Theorems (CLTs) play a crucial role in understanding the errors originating from Monte Carlo estimation, making them an important topic of discussion in the MCMC literature.

6.1. A Negative Result.

In this subsection, we will illustrate that where CLTs might not hold, even when $\pi(h^2) < \infty$. For example, Metropolis-Hastings algorithms with very low acceptance probabilities can lead to $\tau = \infty$, resulting in the failure of $\sqrt{n}$-CLTs. To elaborate further, we will establish the following result:

Lemma 6.2. Consider a reversible Markov chain, beginning in its stationary distribution $\pi(\cdot)$, and let $r(x) \coloneqq \mathbf{P}(X_{n+1} = X_n | X_n = x)$. Then if

$$\lim_{n\to\infty} n\pi([h - \pi(h)]^2 r^n) = \infty, \tag{6.16}$$

then a $\sqrt{n}$-CLT does not hold for h.

Proof. We compute directly from (6.15) that

$$\begin{aligned}
\sigma^2 &= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left[\left(\sum_{i=1}^{n}[h(X_i)-\pi(h)]\right)^2\right] \\
&\geq \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left[\left(\sum_{i=1}^{n}[h(X_i)-\pi(h)]\right)^2 \mathbf{1}(X_0 = X_1 = \cdots = X_n)\right] \\
&= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left[(n[h(X_i)-\pi(h)])^2\, r(X_0)^n\right] \\
&= \lim_{n\to\infty} \frac{1}{n}\cdot n^2\mathbf{E}\left[(h(X_i)-\pi(h))^2 r(X_0)^n\right] \\
&= \lim_{n\to\infty} n\pi\left([h-\pi(h)]^2 r^n\right) \\
\text{(By (6.16))} \qquad &= \infty
\end{aligned}$$

Therefore, a $\sqrt{n}$-CLT does not exist. □

The next question to address is what conditions on the transitions of the Markov chain, as well as on the functional h, ensure a $\sqrt{n}$-CLT for h.

6.2. **Conditions Guaranteeing CLTs.** In this subsection, we introduce several positive results about the existence of CLTs and some of these results will be proved in the following subsections.

For i.i.d. samples, the classical theory ensures a CLT when the second moments are finite. For uniformly ergodic chains, an identical result exists:

Theorem 6.3. If a Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic, then a $\sqrt{n}$-CLT holds for h whenever $\pi(h^2) < \infty$.

Then, it is natural to consider the scenario where a chain is geometrically ergodic but not uniformly ergodic. Interestingly, a similar result exists:

Theorem 6.4. If a Markov chain with stationary distribution $\pi(\cdot)$ is geometrically ergodic, then a $\sqrt{n}$-CLT holds for h whenever $\pi(|h|^{2+\delta}) < \infty$ for some $\delta > 0$.

The above result can be strengthened if the chain is reversible:

Theorem 6.5. If the Markov chain is geometrically ergodic and reversible, then a $\sqrt{n}$-CLT holds for h whenever $\pi(h^2) < \infty$.

It is worth pondering over the following open issue: Consider a Markov chain that is geometrically ergodic, but not necessarily reversible. Let $h : \mathcal{X} \to \mathbb{R}$ where $\pi(h^2) < \infty$. Does a $\sqrt{n}$-CLT always hold for h in this context?

To explore possible solutions for this open question, a promising starting point could involve examining chains of the form $P = P_1P_2$, where each of P_1 and P_2 is reversible with respect to $\pi(\cdot)$, but P is not reversible. Showing that $\sqrt{n}$-CLT's must exist whenever $\pi(h^2) < \infty$ may gives interesting results.

On the other hand, demonstrating a counterexample would involve a Markov chain that is geometrically ergodic but not reversible, and a functional $h : \mathcal{X} \to \mathbb{R}$ such that $\pi(h^2) < \infty$ but $\pi(|h|^{2+\delta}) = \infty$ for all $\delta > 0$, which does not have a $\sqrt{n}$-CLT.

Olle Häggström has produced a counterexample showing that the answer is no in general; see [Hä05] for details.

If P is reversible, then as demonstrated in the following theorem, the only requirement for the establishment of a $\sqrt{n}$-CLT is the finiteness of σ^2:

Theorem 6.6. For a ϕ-irreducible and aperiodic Markov chain which is reversible, a $\sqrt{n}$-CLT holds for h whenever $\sigma^2 < \infty$, where σ^2 is given by (6.15).

In a different direction, we have the following:

Theorem 6.7. Suppose a Markov chain is geometrically ergodic, satisfying the univariate drift condition (2.8) for some $V : \mathcal{X} \to [1, \infty]$ which is finite π-a.e. Let $h : \mathcal{X} \to \mathbb{R}$ with $h^2 \leq KV$ for some $K < \infty$. Then a $\sqrt{n}$-CLT holds for h.

Before we proceed with the proofs of the previously mentioned results, let us consider the following propositions, which could have practical significance.

Proposition 6.8. The above CLT results (Theorem 6.3 - Theorem6.7) all remain true if instead of beginning with $X_0 \sim \pi(\cdot)$, as above, we begin with $X_0 = x$, for π-a.e. $x \in \mathcal{X}$.

Proof. The assumptions of the above CLT results all indicate that the chain is ϕ-irreducible and aperiodic, with stationary distribution $\pi(\cdot)$. Hence, by Theorem 2.14, the chain converges to $\pi(\cdot)$ from π-a.e. $x \in \mathcal{X}$. Fix such an $x \in \mathcal{X}$ and an arbitrary $\epsilon > 0$. There exists some $m \in \mathbb{N}$ such that

$$\|P^m(x, \cdot) - \pi(\cdot)\| \leq \epsilon, \quad \forall n \geq m.$$

Then, by Proposition 2.13 (7), we can jointly construct copies $\{X_n\}$ and $\{X'_n\}$ of the Markov chain, starting from $X_0 = x$ and $X'_0 \sim \pi(\cdot)$, such that

$$\mathbf{P}(X_n = X'_n, \forall\, n \geq m) \geq 1 - \|P^m(x, \cdot) - \pi(\cdot)\| \geq 1 - \epsilon.$$

Therefore, for any $A \subset \mathcal{X}$,

$$\begin{aligned}
&\limsup_{n\to\infty} \left| \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X_i) - \pi(h)] \in A \right) - \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X'_i) - \pi(h)] \in A \right) \right| \\
=&\limsup_{n\to\infty} \left| \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X_i) - \pi(h)] \in A, X_n = X'_n \right) + \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X_i) - \pi(h)] \in A, X_n \neq X'_n \right) \right. \\
&\left. - \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X'_i) - \pi(h)] \in A, X_n = X'_n \right) - \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X'_i) - \pi(h)] \in A, X_n \neq X'_n \right) \right| \\
\leq&\limsup_{n\to\infty} \left| \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X_i) - \pi(h)] \in A, X_n \neq X'_n \right) - \mathbf{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} [h(X'_i) - \pi(h)] \in A, X_n \neq X'_n \right) \right| \\
\leq&1 - \mathbf{P}(X_n = X'_n, \forall\, n \geq m) \\
\leq&1 - (1 - \epsilon) \\
=&\epsilon
\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, and $\mathbf{P}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}[h(X_i') - \pi(h)]\right)$ converges weakly to $N(0,\sigma^2)$ by the previous CLT results, we have $\mathbf{P}\left(n^{-\frac{1}{2}}\sum_{i=1}^{n}[h(X_i) - \pi(h)]\right)$, which starts from $x \in \mathcal{X}$, also follows $\sqrt{n}$-CLTs. □

Proposition 6.9. Theorem 6.3 and Theorem 6.4 remain true if the chain is periodic of period $d \geq 2$, provided that the d-step chain $P' = P^d|_{\mathcal{X}_1}$ (as in the proof of Corollary 2.18) has all the other properties required of P in the original result (i.e. ϕ-irreducibility, and uniform or geometric ergodicity), and that the function h still satisfies the same moment condition.

Proof. Recall the proof of Corollary 2.18, let $\overline{P}$ be the d-step chain defined on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ and $\overline{h}(x_0,\ldots,x_{d-1}) \coloneqq h(x_0)+\cdots+h(x_{d-1})$. Clearly, As in the proof of Corollary 2.18, let $\overline{P}$ be the d-step chain defined on $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ and $\overline{h}(x_0,\ldots,x_{d-1}) \coloneqq h(x_0)+\cdots+h(x_{d-1})$. Then $\overline{P}$ inherits the irreducibility and ergodicity properties of P':

- Irreducibility: Since P' is irreducible, there exists a σ-finite measure ϕ on $\mathcal{X}$ such that for all $A \subset \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x,A)$ such that $(P')^n(x,A) > 0$. Then, consider the measure $\overline{\phi} \coloneqq \phi \times (\phi P) \times \cdots \times (\phi P^{d-1})$. If $A_1 \times \cdots \times A_d \subset \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ $\overline{\phi}(A_1 \times \cdots \times A_d) > 0$, then $\phi(A_1) > 0, (\phi P)(A_2) > 0, \ldots, (\phi P^{d-1})(A_d) > 0$. It follows from the ϕ-irreducibility of P' that there exists $\overline{n} \in \mathbb{N}$ such that
$$(\overline{P})^{\overline{n}}((x_1,\ldots,x_d), A_1 \times \cdots \times A_d) > 0.$$
- Ergodicity: By [RR01b] Theorem 1, since P' is de-initializing for $\overline{P}$, we have
$$\|\overline{P}((x_1,\ldots,x_d),\cdot) - \overline{\pi}(\cdot)\| \leq \|P'(x,\cdot),\pi(\cdot)\|.$$
Then, ergodicity follows.

Then, Applying Theorem 6.3 or 6.4 establishes a CLT for $\overline{P}$ and $\overline{h}$, which implies a CLT for P and h.

□

Remark 6.10. In particular, for any irreducible (or indecomposible) Markov chain on a finite space (then it is uniformly ergodic), we can deduce from Theorem 6.3 and Proposition 6.9 that a $\sqrt{n}$-CLT is always valid since $\pi(h^2)$ is always finite.

6.3. **CLT Proofs using the Poisson Equation.** In this subsection, we will prove some of the previously mentioned results using the Poisson euqation. We will start by introducing a version of the martingale central limit theorem, a theorem widely covered in standard textbooks.

Theorem 6.11. Let $\{Z_n\}$ be a stationary ergodic sequence, with $\mathbf{E}[Z_n|Z_1,\ldots,Z_{n-1}] = 0$ and $\mathbf{E}[(Z_n)^2] < \infty$. Then, $n^{-\frac{1}{2}}\sum_{i=1}^{n} Z_i$ converges weakly to $N(0,\sigma^2)$ distribution for some $\sigma^2 < \infty$.

To make use of Theorem 6.11, consider the Poisson equation $h - \pi(h) = g - Pg$. A useful result is the following:

Theorem 6.12. Let P be a transition kernel for an aperiodic, ϕ-irreducible Markov chain on a state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$, with $X_0 \sim \pi(\cdot)$. Let

$h : \mathcal{X} \to \mathbb{R}$ with $\pi(h^2) < \infty$, and suppose there exists $g : \mathcal{X} \to \mathbb{R}$ with $\pi(g^2) < \infty$ which solves the Poisson equation, i.e. such that $h - \pi(h) = g - Pg$. Then h satisfies a $\sqrt{n}$-CLT.

Proof. Let $Z_n \coloneqq g(X_n) - Pg(X_{n-1})$. Then, since the Markov chain starts from the stationary distribution, we have $X_n \sim \pi(\cdot)$ and $\{Z_n\}$ is stationary. Additionally, $\{Z_\}n$ inherits irreducibility and aperiodicity from $\{X_n\}$. It follows from Theorem 2.14 that the Markov chain converges asymptotically, which implies $\{Z_n\}$ is ergodic. We notice that

$$\begin{aligned}\mathbf{E}[g(X_n) - Pg(X_{n-1})|X_0, \ldots, X_{n-1}] &= \mathbf{E}[g(X_n)|X_{n-1}] - Pg(X_{n-1}) \\ &= Pg(X_{n-1}) - Pg(X_{n-1}) \\ &= 0.\end{aligned}$$

Since $Z_1, \ldots, Z_{n-1} \in \sigma(X_0, \ldots, X_{n-1})$, we have

$$\mathbf{E}_\pi[Z_n|Z_1, \ldots, Z_{n-1})] = \mathbf{E}[g(X_n) - Pg(X_{n-1})|X_0, \ldots, X_{n-1}] = 0.$$

Then, by Theorem 6.11, $n^{-\frac{1}{2}} \sum_{i=1}^n Z_i$ converges to $N(0, \sigma^2)$. Moreover,

(by the Poisson equation)

$$\begin{aligned}n^{-\frac{1}{2}} \sum_{i=1}^n [h(X_i) - \pi(h)] &= n^{-\frac{1}{2}} \sum_{i=1}^n [g(X_i) - Pg(X_i)] \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n [g(X_i) - Pg(X_{i-1})] + n^{-\frac{1}{2}} Pg(X_0) - n^{-\frac{1}{2}} Pg(X_n) \\ &= n^{-\frac{1}{2}} \sum_{i=1}^n Z_i + n^{-\frac{1}{2}} Pg(X_0) - n^{-\frac{1}{2}} Pg(X_n).\end{aligned}$$

Since $n^{-\frac{1}{2}} Pg(X_0)$ and $n^{-\frac{1}{2}} Pg(X_n)$ converge to zero in probability as $n \to \infty$ and $n^{-\frac{1}{2}} \sum_{i=1}^n Z_i$ converges weakly to $N(0, \sigma^2)$, we have built an $\sqrt{n}$-CLT for h, i.e. $n^{-\frac{1}{2}} \sum_{i=1}^n [h(X_i) - \pi(h)]$ converges weakly to $N(0, \sigma^2)$. □

Corollary 6.13. Let P be a transition kernel for an aperiodic, ϕ-irreducible Markov chain on a state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$, with $X_0 \sim \pi(\cdot)$. Let $h : \mathcal{X} \to \mathbb{R}$ with $\pi(h^2) < \infty$. If $\sum_{k=0}^\infty \sqrt{\pi\left((P^k[h - \pi(h)])^2\right)} < \infty$, then h satisfies a $\sqrt{n}$-CLT.

Proof. Let $g(x) \coloneqq \sum_{k=0}^\infty g_k(x)$, and

$$g_k(x) \coloneqq P^k h(x) - \pi(h) = P^k[h - \pi(h)](x),$$

where by convention $P^0h(x) = h(x)$. By direct computation, we have

$$\begin{aligned}(g - Pg)(x) &= \sum_{k=0}^{\infty} g_k(x) - \sum_{k=0}^{\infty} Pg_k(x) \\ &= \sum_{k=0}^{\infty} g_k(x) - \sum_{k=0}^{\infty} P^{k+1}h(x) - \underbrace{P\pi(h)}_{=\pi(h)} \\ &= \sum_{k=0}^{\infty} g_k(x) - \sum_{k=0}^{\infty} g_{k+1}(x) \\ &= g_0(x) \\ &= P^0h(x) - \pi(h) \\ &= h(x) - \pi(h),\end{aligned}$$

and the Poisson equation is satisfied. Then, we are going to show that $\pi(g^2) < \infty$. Since the $L^2(\pi)$ norm satisfies the triangle inequality,

$$\sqrt{\pi(g^2)} = \sqrt{\pi\left[\left(\sum_{k=0}^{\infty} g_k\right)^2\right]} \leq \sum_{k=0}^{\infty} \sqrt{\pi(g_k^2)} = \sum_{k=0}^{\infty} \sqrt{\pi\left((P^k[h - \pi(h)])^2\right)} < \infty.$$

It follows from Theorem 6.12 that h satisfies a $\sqrt{n}$-CLT. □

In the rest of this subsection, we will provide the proofs for Theorem 6.5 and Theorem 6.7.

Theorem 6.5. If the Markov chain is geometrically ergodic and reversible, then a $\sqrt{n}$-CLT holds for h whenever $\pi(h^2) < \infty$.

Proof of Theorem 6.5. Consider the usual $L^2(\pi)$ operator norm for P, which is

$$\|P\|_{L^2(\pi)} = \sup_{\substack{\pi(f)=0 \\ \pi(f^2)=1}} \pi\left((Pf)^2\right) = \sup_{\substack{\pi(f)=0 \\ \pi(f^2)=1}} \int_{x\in\mathcal{X}} \left(\int_{y\in\mathcal{X}} f(y)P(x, dy)\right)^2 \pi(dx).$$

It is shown in Theorem 2 of [RR97] that reversible chains are geometrically ergodic if and only if they satisfy

$$\|P\|_{L^2(\pi)} < 1.$$

It follows that there is $\beta < 1$ such that

$$\pi((Pf)^2) \leq \beta^2\pi(f^2), \quad \forall\, f \text{ with } \pi(f) = 0 \text{ and } \pi(f^2) < \infty.$$

Furthermore, since reversibility implies self-adjointness of P in $L^2(\pi)$, we have

$$\|P^k\|_{L^2(\pi)} = \|P\|^k_{L^2(\pi)}.$$

By the above inequality and equality, we have

$$\pi((P^kf)^2) \leq \beta^{2k}\pi(f^2).$$

Let $g_k = P^k h - \pi(h)$ as in the proof of Corollary 6.13. Then, since $\pi(g) = 0$ and $\pi(g^2) < \infty$, we have

$$\pi((g_k)^2) \leq \beta^{2k} \pi((h - \pi(h))^2).$$

It follows that

$$\sum_{k=0}^{\infty} \sqrt{\pi(g_k^2)} \leq \sum_{k=0}^{\infty} \sqrt{\beta^{2k} \pi((h - \pi(h))^2)} = \sqrt{\pi((h - \pi(h))^2)} \sum_{k=0}^{\infty} \beta^k = \frac{\sqrt{\pi((h - \pi(h))^2)}}{1 - \beta} < \infty.$$

Therefore, the result follows from Corollary 2.18. □

Theorem 6.7. Suppose a Markov chain is geometrically ergodic, satisfying the univariate drift condition (2.8) for some $V : \mathcal{X} \to [1, \infty]$ which is finite π-a.e. Let $h : \mathcal{X} \to \mathbb{R}$ with $h^2 \leq KV$ for some $K < \infty$. Then a $\sqrt{n}$-CLT holds for h.

Proof of Theorem 6.7. Proposition 1 in [RR97] builds the equivalence between geometric ergodicity and V-uniformly ergodic. This equivalence implies that there is $C < \infty$ and $\rho < 1$ such that for $x \in \mathcal{X}$ and $|f| \leq V$,

$$|P^n f(x) - \pi(f)| \leq CV(x)\rho^n.$$

Let $g_k = P^k[h - \pi(h)]$ as in the proof of Corollary 6.13. By the Cauchy-Schwartz inequality,

$$(g_k)^2 = (P^k[h - \pi(h)])^2 \leq P^k([h - \pi(h)]^2).$$

□

6.4. **Proof of Theorem 6.4.**

Theorem 6.4. If a Markov chain with stationary distribution $\pi(\cdot)$ is geometrically ergodic, then a $\sqrt{n}$-CLT holds for h whenever $\pi(|h|^{2+\delta}) < \infty$ for some $\delta > 0$.

In this subsection, we will use regeneration theory to to provide a relatively straightforward proof of Theorem 6.4.

Th regeneration construction is very similar to the coupling construction, except now just for a single chain $\{X_n\}$. A small set is still crucial, as in the coupling construction. An important fact we will leverage is the equivalence of the minorization condition (2.7) and the univariate drift condition (2.8) with geometric ergodicity (also equivalent to V-uniform ergodicity); this fact follows form Theorem 15.0.1, Theorem 16.0.1, and Theorem 14.3.7 of [MT93], and Proposition 1 of [RR04].

The regeneration construction is given as follows:

Begin with $X_0 = x$ where $x \in \mathcal{X}$, and set $n = 0$. Then, repeat the following loop forever.

Beginning of the loop. Given X_n

(1) If $X_n \in C$, then:
   (a) With probability ϵ, choose $X_{n+n_0} \sim \nu(\cdot)$;
   (b) Else, with probability $1 - \epsilon$, choose

$$X_{n+n_0} \sim R(X_n, \cdot).$$

In the case $n_0 > 1$, for completeness, we go back and construct $X_n + 1, \ldots, X_{n+n_0-1}$ from their correct conditional distributions given X_0 and X_{n+n_0}.

Replace n by $n + n_0$.

(2) Else, choose $X_{n+1} \sim P(X_n, \cdot)$.

Then return to the beginning of the loop.

Consider the *regeneration times* $T_1, T_2, \ldots$, which are the moments when $X_{T_i} \sim \nu(\cdot)$ as in Condition 1 of the regeneration construction.

Thus, the regeneration times occur with probability ϵ precisely n_0 iterations after each time the chain enters C (not counting those entries of C which are within n_0 of a previous regeneration attempt). Obviously, provided the chain enters C, the regeneration times occur with probability ϵ precisely n_0 iterations (excluding those entries of C which are within the "filling in" times for going back and constructing $X_{n+1}, \ldots, X_{n+n_0}$).

We can break sums $\sum_{i=0}^{n}[h(X_i) - \pi(h)]$ into sums of independent and identically distributed (i.i.d.) "tours". Indeed, take the random variables $Y_j \coloneqq (X_{T_j}, X_{T_j+1}, \ldots, X_{T_{j+1}-1})$ as a complete tour; it is not hard to see that $Y_1, Y_2, \ldots$ begin from the same fixed distribution $\nu(\cdot)$ and are i.i.d. Let $T_0 = 0$, and let $r(n) \coloneqq \sup\{i \geq 0 : T_i \leq n\}$. We can break the sum as

$$\sum_{i=1}^{n}[h(X_i) - \pi(h)] = \sum_{j=1}^{r(n)} \sum_{i=T_j}^{T_{j+1}-1} [h(X_i) - \pi(h)] + E(n), \tag{6.17}$$

where $E(n) \coloneqq X_0 + \cdots + X_{T_1-1} + X_{T_{r(n)+1}} + \cdots + X_n$ is an error term which collects the terms corresponding to the first tour $X_0, \ldots, X_{T_1-1}$ and the incomplete final tour $X_{T_{r(n)}+1}, \ldots, X_n$.

By the elementary renewal theory,

$$\frac{r(n)}{n} \to \epsilon\pi(C) \quad \text{in probability.}$$

If each sum has finite second moment and the error term is bounded in probability, we can apply classic central limit theorem and then prove Theorem 6.4.

Lemma 6.14. $E(n)$ is $O_p(1)$ as $n \to \infty$.

Proof. Geometric ergodicity implies (as in the proof of Lemma 4.6) exponential tails on the return times to C, i.e. there exists $N \in \mathbb{N}$ such that

$$\mathbf{P}(\tau_C \geq n | X_0 = x) \leq \lambda^n V(x), \quad \forall n \geq N.$$

It then follows that

$$\mathbf{E}_\pi[\beta^{T_1}] < \infty, \quad \text{and} \quad \mathbf{E}[\beta^{T_{j+1}-T_j}] < \infty. \tag{6.18}$$

By the above inequalities, standard renewal theory ensures that $E(n)$ has limiting distribution as $n \to \infty$. Hence, $E(n)$ is $O_p(1)$ as $n \to \infty$ and can be neglected when multiplied by $n^{-\frac{1}{2}}$. □

Since each tour starts from the distribution $\nu(\cdot)$, the finiteness of second moments of each term in (6.17) can be concluded by the following lemma:

Lemma 6.15. $\int_{x\in\mathcal{X}} \nu(dx)\mathbf{E}\left[\left(\sum_{i=0}^{T_1-1}[h(X_i - \pi(h)]\right)^2 \middle| X_0 = x\right] < \infty.$

Proof. Note that

$$\begin{aligned}
&\text{(By stationarity)} && \pi(\cdot) = \int_{x\in\mathcal{X}} \pi(dx)P(x,\cdot)\\
&(C \subset \mathcal{X}) && \geq \int_{x\in C} \pi(dx)P(x,\cdot)\\
&(C \text{ is small}) && \geq \pi(C)\epsilon\nu(\cdot).
\end{aligned}$$

It follows that

$$\nu(dx) \leq \frac{\pi(dx)}{\nu(C)\epsilon}$$

and then

$$\begin{aligned}
&\int_{x\in\mathcal{X}} \pi(dx)\mathbf{E}\left[\left(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\right)^2 \Big| X_0 = x\right]\\
\leq &\int_{x\in\mathcal{X}} \frac{\pi(dx)}{\nu(C)\epsilon}\mathbf{E}\left[\left(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\right)^2 \Big| X_0 = x\right].
\end{aligned}$$

It suffices to prove that the right hand side is finite, which is equivalent to prove

$$\int_{x\in\mathcal{X}} \mathbf{E}\left[\left(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\right)^2 \Big| X_0 = x\right]\pi(dx) < \infty.$$

For notational simplicity, set $H_i = h(X_i) - \pi(h)$. We have

$$\left(\sum_{i=1}^{T_1-1}[h(X_i)-\pi(h)]\right)^2 = \left(\sum_{i=0}^{\infty}\mathbf{1}_{i<T_1}H_i\right)^2.$$

Then, by Cauchy-Schwartz inequality $\mathbf{E}[AB] \leq \sqrt{\mathbf{E}[A^2]\mathbf{E}[B^2]}$,

$$\begin{aligned}
\mathbf{E}_\pi\left[\left(\sum_{i=1}^{T_1-1}[h(X_i)-\pi(h)]\right)^2\right] &= \mathbf{E}_\pi\left[\left(\sum_{i=0}^{\infty}\mathbf{1}_{i<T_1}H_i\right)^2\right]\\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\mathbf{E}_\pi\left[(\mathbf{1}_{i<T_1}H_i)(\mathbf{1}_{j<T_1}H_j)\right]\\
&\leq \sum_{i=0}^{\infty}\sum_{i=0}^{\infty}\sqrt{\mathbf{E}_\pi\left[(\mathbf{1}_{i<T_1}H_i)^2\right]\mathbf{E}_\pi\left[(\mathbf{1}_{j<T_1}H_j)^2\right]}\\
&\leq \left(\sum_{i=1}^{\infty}\sqrt{\mathbf{E}_\pi\left[\mathbf{1}_{i<T_1}H_i^2\right]}\right)^2 \qquad (6.19)
\end{aligned}$$

□

Let $p = 1 + \frac{2}{\delta}$ and $q = 1 + \frac{\delta}{2}$. We have $pq = 2 + \frac{2}{\delta} + \frac{\delta}{2} = p + q \Rightarrow \frac{1}{p} + \frac{1}{q} = 1$. Then, by Holder's inequality,

$$\mathbf{E}_\pi[\mathbf{1}_{i<T_1} H_i^2] \leq \mathbf{E}_\pi[\mathbf{1}_{i<T_1}]^{\frac{1}{p}} \mathbf{E}_\pi[|H_i|^{2q}]^{\frac{1}{q}}. \tag{6.20}$$

Since $X_0 \sim \pi(\cdot)$, $\mathbf{E}_\pi[|H_i|^{2q}]$ is a constant, independent of i, say $K \coloneqq \mathbf{E}_\pi[|H_i|^{2q}]$. Additionally, $K < \infty$ since $\pi(|h|^{2+\delta}) < \infty$.

Then, we again take a look at (6.18). By Markov's inequality,

$$\mathbf{E}_\pi[\mathbf{1}_{0\leq i<T_1}] \leq \mathbf{E}_\pi[\mathbf{1}_{\beta^{T_1}>\beta^i}] \leq \beta^{-i}\mathbf{E}_\pi[\beta^{T_1}]. \tag{6.21}$$

Combining (6.19) and (6.20) gives

$$\begin{aligned}
\mathbf{E}_\pi\left[\left(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\right)^2\right] &\leq \left(\sum_{i=1}^{\infty}\sqrt{\mathbf{E}_\pi[\mathbf{1}_{i<T_1}]^{\frac{1}{p}}\mathbf{E}_\pi[|H_i|^{2q}]^{\frac{1}{q}}}\right)^2 \\
&= \left(K^{\frac{1}{2q}}\sum_{i=0}^{\infty}\sqrt{\mathbf{E}_\pi[\mathbf{1}_{i<T_1}]^{\frac{1}{p}}}\right)^2 \\
\text{(by (6.21))} \quad &\leq \left(K^{\frac{1}{2q}}\sum_{i=0}^{\infty}\sqrt{(\beta^{-i}\mathbf{E}_\pi[\beta^{T_1}])^{\frac{1}{p}}}\right)^2 \\
&= \left(K^{\frac{1}{2q}}\mathbf{E}_\pi[\beta^{T_1}]^{\frac{1}{2p}}\sum_{i=0}^{\infty}\beta^{-\frac{i}{2p}}\right)^2 \\
\text{(since } \beta^{\frac{1}{2p}} > 1\text{)} \quad &= \left(\frac{K^{\frac{1}{2q}}\mathbf{E}_\pi[\beta^{T_1}]^{\frac{1}{2p}}}{1-\beta^{-\frac{1}{2p}}}\right)^2 \\
\text{(by (6.18) and finiteness of } K\text{)} \quad &< \infty
\end{aligned}$$

7. Optimal Scaling and Weak Convergence

In this section, we will provide a brief overview of another application of probability theory to MCMC, known as the optimal scaling problem. However, we will only give an introduction and not delve into extensive details here.

Sometimes, the Metropolis-Hastings algorithms might be very inefficient, i.e. it will take too many iterations to reach the target distribution.

Let $\pi_u : \mathbb{R}^d \to [0, \infty)$ be a continuous d-dimensional density (d large). Consider running a random walk Metropolis algorithm for π_u, with proposal distribution given by $Q(x, \cdot) = N(x, \sigma^2 I_d)$. The acceptance probability simplifies to

$$\alpha(x, y) = \min\left\{1, \frac{\pi_u(y)}{\pi_u(x)}\right\}.$$

We are interested in the following question: How do we determine the appropriate value for σ?

If σ^2 is chosen to be too small, then the proposal Y_{n+1} generated from X_n will be very close to X_n. Due to continuity, $\frac{\pi_u(y)}{\pi_u(x)}$ will be relatively large, resulting in a high acceptance probability. Consequently, the next state X_{n+1} will either be very closed to

X_n (accept $X_{n+1} = Y_{n+1}$), or stay at X_n (reject, and set $X_{n+1} = X_n$). Therefore, the chain will moves very slowly, leading to very poor performance.

On the other hand, if σ^2 is chosen to be too large, then the generated proposal Y_{n+1} will often be far from the current state X_n. This can be advantageous since accepting such a big jump would pushes the Markov chain forward. However, again by continuity, the acceptance probability will be quite low. Hence, unless the chain happens to be very "lucky", most of these large-step proposals will be rejected, causing the chain becoming "trapped" in the same state for long periods of time. This scenario would also result in poor performance.

Therefore, we aim to pick values that satisfy a Goldilocks Principle: σ should be "just right", neither too small nor too large.

To prove theorems about this, assume for now that

$$\pi_u(\mathbf{x}) = \prod_{i=1}^{d} f(x_i), \tag{7.22}$$

i.e. that the density π_u can be factored into i.i.d. components, each with (smooth) density f. Although this assumption is quite restrictive and not practically useful, as it would allow each coordinate to be simulated independently, it does provide a framework for developing interesting theoretical insights. Also, assume that chain begins in stationarity, i.e. that $X_0 \sim \pi(\cdot)$.

7.1. **The Random Walk Metropolis (RWM) Case.** Define $I := \mathbf{E}[((\log f(Z))')^2]$ where $Z \sim f(z)dz$. In this subsection, we will provide a brief overview of how to show that under the assumption (7.22), as the dimension $d \to \infty$, choosing $\sigma^2 \approx \frac{(2.38)^2}{Id}$ becomes optimal. This choice results in an approximate asymptotic acceptance rate of 0.234.

We set $\sigma_d^2 = \frac{l^2}{d}$, where $l > 0$ is to be determined later. Let $\{X_n\}$ denote the Markov chain generated by the random walk Metropolis algorithm for $\pi(\cdot)$ on $\mathbb{R}^d$ with proposal distribution $Q(x, \cdot) = N(x, \sigma_d^2 I_d)$. Additionally, let $\{N(t)\}_{t\geq 0}$ be an Poisson process with rate d which is independent of $\{X_n\}$. Finally, let

$$Z^d := X^{(1)}_{N(t)}, \quad t \geq 0,$$

where $X^{(1)}$ denotes the first component of a multidimensional random variable. Obviously, $\{Z_t^d\}_{t\geq 0}$ follows the first component of $\{X_n\}$, with time speeded up by the Poisson process $\{N(t)\}_{t\geq 0}$.

It is proved in [GGR97] that as $d \to \infty$, the process $\{Z_t^d\}_{t\geq 0}$ converges weakly to a diffusion process $\{Z_t\}_{t\geq 0}$, which satisfies the following stochastic differential equation:

$$dZ_t = h(l)^{\frac{1}{2}} dB_t + \frac{1}{2} h(l) \nabla \log \pi_u(Z_t) dt,$$

where

$$h(l) = 2l^2 \Phi\left(-\frac{\sqrt{I}l}{2}\right)$$

corresponds to the speed of the limiting diffusion, and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{s^2}{2}} ds$ is the cumulative distribution function of $N(0,1)$.

In the paper [GGR97], the authors argue that as $d \to \infty$, the optimal choice for l is the value that maximizes the speed function $h(l)$. Through straightforward calculations, we find that the derivative of $h(l)$ with respect to l is given by:

$$\frac{d}{dl}h(l) = \frac{4l}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{\sqrt{I}l}{2}} e^{-\frac{s^2}{2}} ds - \frac{l^2}{\sqrt{2\pi}} \sqrt{I} e^{-\frac{Il^2}{8}}.$$

Setting this expression equal to zero yields two solutions: $l = 0$ and $l \approx \frac{2.381}{\sqrt{I}}$. It turns out the value $l \approx \frac{2.381}{\sqrt{I}}$ that maximizes the aforementioned speed function and results in optimally fast mixing. It is also proved in [GGR97] that the expected acceptance rate

$$A_d(l) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \pi_d(\mathbf{x}) \alpha(\mathbf{x}, \mathbf{y}) q(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$$

of the random walk Metropolis algorithm in d dimensions converges to $A(l) = 2\Phi\left(-\frac{\sqrt{I}l}{2}\right)$. Plug $l = \frac{2.38}{\sqrt{I}}$ into $A(l)$, we have $A(l) \approx 0.234$, which gives the optimal asymptotic acceptance rate.

7.2. **The Langevin Algorithm Case.** Let $J \coloneqq \mathbf{E}\left[\frac{5((\log f(Z))'')^2 - 3((\log f(Z))'')^3}{48}\right]$ where again $Z \sim f(z)dz$. In this subsection, we will provide a brief overview of how to show that under the assumption (7.22), as the dimension $d \to \infty$, choosing $\sigma^2 \approx \frac{(0.825)^2}{J^{\frac{1}{2}} d^{\frac{1}{3}}}$ becomes optimal. This choice results in an approximate asymptotic acceptance rate of 0.574.

We set $\sigma_d^2 = \frac{l^2}{d^{\frac{1}{3}}}$, where $l > 0$ is to be determined later. Let $\{X_n\}$ denote the Markov chain generated by the Langevin Algorithm for $\pi(\cdot)$ on $\mathbb{R}^d$ with proposal distribution $Q(x, \cdot) = N(x + \frac{\sigma^2}{2} \nabla \log \pi_u(x), \sigma^2 I_d)$. Additionally, let $\{N(t)\}_{t \geq 0}$ be an Poisson process with rate $d^{\frac{1}{3}}$ which is independent of $\{X_n\}$. Finally, let

$$Z^d \coloneqq X^{(1)}_{N(t)}, \quad t \geq 0,$$

where $X^{(1)}$ denotes the first component of a multidimensional random variable. Obviously, $\{Z_t^d\}_{t \geq 0}$ follows the first component of $\{X_n\}$, with time speeded up by the Poisson process $\{N(t)\}_{t \geq 0}$.

It is proved in [RR02] that as $d \to \infty$, the process $\{Z_t^d\}_{t \geq 0}$ converges weakly to a diffusion process $\{Z_t\}_{t \geq 0}$, which satisfies the following stochastic differential equation:

$$dZ_t = g(l)^{\frac{1}{2}} dB_t + \frac{1}{2} g(l) \nabla \log \pi_u(Z_t) dt,$$

where

$$g(l) = 2l^2 \Phi(-Jl^3)$$

corresponds to the speed of the limiting diffusion.

It can be shown that the value of l which maximizes the speed function $g(l)$ gives an optimal choice for l in the Langevin Algorithm. By direct computation, we find that the derivative of $g(l)$ with respect to l is given by:

$$\frac{d}{dl}g(l) = \frac{4l}{\sqrt{2\pi}}\int_{-\infty}^{-Jl^3} e^{-\frac{s^2}{2}}ds - \frac{6Jl^4}{\sqrt{2\pi}}e^{-\frac{1}{2}J^2l^6}.$$

Setting this expression equal to zero yields two solutions: $l = 0$ and $l \approx \frac{0.825}{\sqrt[3]{J}}$. It turns out the value $l \approx \frac{0.825}{\sqrt[3]{J}}$ that maximizes the aforementioned speed function and results in optimally fast mixing. It is also proved in [RR02] that the expected acceptance rate of the Langevin algorithm in d dimensions converges to $A(l) = 2\Phi\left(-Jl^3\right)$. Plug $l = \frac{0.825}{\sqrt[3]{J}}$ into $A(l)$, we have $A(l) \approx 0.574$, which gives the optimal asymptotic acceptance rate.

7.3. **Discussion of Optimal Scaling.** The above result offers a straightforward guideline for tuning the RWM and the Langevin algorithm under the assumption (7.22): adjust the proposal scaling to achieve an acceptance rate close to the optimal asymptotic rate (0.234 for RWM, 0.574 for Langevin). Applying these results in practice is actually quite simple, as computers can easily track the acceptance rate the the algorithm, allowing users to adjust σ^2 accordingly to achieve the desired acceptance rates. Furthermore, adaptive MCMC algorithms, which speed up the efficiency, have been widely used in recent times. That is, at each iteration, we allow MCMC algorithms to update the proposal distribution $Q(x, \cdot)$ according to specific rules, so that the optimal algorithm can be learnt. However, since adaptive MCMC algorithms violate the Markov property, they in general do not converge to the target distribution. Hence, adaptive MCMC algorithms should be carefully implemented to ensure stationarity; see for example [GRS98].

The above results also shed light on the computational complexity of these algorithms. To elaborate, we have $\sigma^2 = \frac{l^2}{d}$ and thus the efficiency of RWM algorithms scales like d^{-1}; it follows that its computational complexity is $O(d)$. Similarly, for the Langevin algorithms, we have $\sigma^2 = \frac{l^2}{d^{\frac{1}{3}}}$ and thus its computational complexity is $O(d^{\frac{1}{3}})$, implying that the Langevin algorithms are more efficient than RWM algorithms in terms of computational complexity.

It is worth highlighting that achieving an acceptance rate of exactly 0.234 (or 0.574) isn't essential for achieving good efficiency; a fairly close value is enough. Additionally, in practice, we don't need a very large dimensions in order to approach the asymptotic behaviour; in fact, even in dimensions as small as 5 or 10, the value of 0.234 (or 0.574) is close to being optimal. One can refer to the review article [RR01a] for further details.

The results presented above are established under the strong assumption 7.22. Numerous researchers have attempted to relax and generalize this assumption. For example, the optimal-scaling results are extended to inhomogeneously-scaled components of the form $\pi_u(\mathbf{x}) = \prod_{i=1}^{d} C_i f(C_i x_i)$, [Rob98] in [RR01a], to discrete hypercubes in [Rob98], to finite-range homogeneous Markov random fields in [BR00]; in particular, the optimal acceptance rate remains 0.234 (under appropriate assumptions) in all of these three scenarios. On the other hand, if the chain starts far out in the tails of the stationary distribution $\pi(\cdot)$, instead of $X_0 \sim \pi(\cdot)$, we will encounter some surprising behaviours; see

[CRR05] for details. The true level of generality of these optimal scaling results remains an open problem.

Appendix A. Proof of Lemma 4.6

Lemma 4.6. For an aperiodic, ϕ-irreducible Markov chain, all petite sets are small.

To prove this, we require a lemma related to aperiodicity

Lemma A.1. Consider an aperiodic Markov chain on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Let $\nu(\cdot)$ be any probability measure on $\mathcal{X}$. Assume that $\nu(\cdot) \ll \pi(\cdot)$, and that for all $x \in \mathcal{X}$, there is $n = n(x) \in \mathbb{N}$ and $\delta = \delta(x) > 0$ such that $P^n(x, \cdot) \geq \delta\nu(\cdot)$ (for example, this always holds if $\nu(\cdot)$ is a minorisation measure for a small or petite set which is reachable from all states). Let $T \coloneqq \{n \geq 1 : \exists \delta_n > 0 \text{ s.t. } \int \nu(x)P^n(x, \cdot) \geq \delta_n\nu(\cdot)\}$, and assume that T is non-empty. Then there is $n_* \in \mathbb{N}$ with $T \supset \{n_*, n_* + 1, n_* + 2, \ldots\}$.

Proof. Since $P^{n(x)}(x, \cdot) \geq \delta(x)\nu(\cdot)$ for all $x \in \mathcal{X}$, T is non-empty. If $n, m \in T$, we have

$$\begin{aligned}
\int_{x\in\mathcal{X}} \nu(dx)P^{n+m}(x, \cdot) &= \int_{x\in\mathcal{X}} \int_{y\in\mathcal{X}} \nu(dx)P^n(x, dy)P^m(y, \cdot) \\
\text{(since } n \in T\text{)} \qquad &\geq \int_{y\in\mathcal{X}} \delta_n\nu(dy)P^m(y, \cdot) \\
\text{(since } m \in T\text{)} \qquad &\geq \delta_n\delta_m\nu(\cdot).
\end{aligned}$$

Therefore, if $n, m \in T$, $n + m \in T$. Then, we are going to show that $\gcd(T) = 1$ by contradiction. Suppose to the contrary that $\gcd(T) = d > 1$. For $1 \leq i \leq d$, define

$$\mathcal{X}_i \coloneqq \left\{x \in \mathcal{X} : \exists l \in \mathbb{N} \text{ and } \delta > 0 \text{ s.t. } P^{ld-i} \geq \delta\nu(\cdot)\right\}.$$

Since $P^n(x, \cdot) \geq \delta\nu(\cdot)$ holds for any $x \in \mathcal{X}$, it follows from the assumption that $\bigcup_{i=1}^d \mathcal{X}_i = \mathcal{X}$. Set

$$S \coloneqq \bigcup_{i\neq j} (\mathcal{X}_i \cap \mathcal{X}_j)$$

and

$$\overline{S} \coloneqq S \cup \{x \in \mathcal{X} : \exists m \in \mathbb{N} \text{ s.t. } P^m(x, S) > 0\}.$$

We can see that S is the union of common areas shared by at least two X_i's and $\overline{S}$ is the set of all elements in $\mathcal{X}$ which can reach S. Let $\mathcal{X}_i' \coloneqq \mathcal{X}_i \backslash \overline{S}$. $\mathcal{X}_i'$'s are by definition disjoint. We note that for $x \in \mathcal{X}_i'$, $P(x, \overline{S}) = 0$; hence, $P(x, \bigcup_i \mathcal{X}_i') = 1 - P(x, \overline{S}) = 1$. In fact, for $x \in \mathcal{X}_i'$,

$$P(x, \mathcal{X}_{i+1}') = 1, \quad \text{if } i < d,$$

and

$$P(x, \mathcal{X}_1') = 1, \quad \text{if } i = d.$$

Indeed, suppose $x \in \mathcal{X}_i'$ for $i < d$, and $P(x, \mathcal{X}_j') > 0$ for some $j \neq i + 1$. Then, by definition, there exists $l \in \mathbb{N}$ and $\delta > 0$ such that

$$P^{ld-j}(x, \cdot) \geq \delta\nu(\cdot).$$

It follows that

$$\begin{aligned} P^{ld-(j-1)}(x,\cdot) &\geq \int_{y\in\mathcal{X}'_j} P(x,dy)P^{ld-j}(y,\cdot) \\ &\geq \int_{y\in\mathcal{X}_j} P(x,dy)\delta\nu(\cdot) \\ &= P(x,\mathcal{X}'_j)\delta\nu(\cdot) \\ \text{(Let } \delta' = P(x,\mathcal{X}'_j)\delta > 0) \qquad &= \delta'\nu(\cdot). \end{aligned}$$

Hence, $x \in \mathcal{X}'_{j-1}$. However, X'_{j-1} and X'_i are disjoint, contradiction.

We claim that for all $m \geq 0$, $\nu P^m(\mathcal{X}_i \cap \mathcal{X}_j) \coloneqq \int_{\mathcal{X}} \nu(dx)P^m(x,\mathcal{X}_i \cap \mathcal{X}_j) = 0$ whenever $i \neq j$. Indeed, if $\nu P^m(\mathcal{X}_i \cap \mathcal{X}_j) > 0$, then there would be $S' \subset \mathcal{X}$, $l_1, l_2 \in \mathbb{N}$ and $\delta > 0$ such that for all $x \in S'$,

$$P^{l_1 d-i}(x,\cdot) \geq \delta\nu(\cdot)$$

and

$$P^{l_2 d-i}(x,\cdot) \geq \delta\nu(\cdot).$$

This implies $l_1 d - i + m \in T$ and $l_2 d - j + m \in T$, contradicting the fact that $\gcd(T) = d$. Setting $m = 0$ gives $\nu(\mathcal{X}_i \cap \mathcal{X}_j) = 0$ for $i \neq j$ and setting $m > 0$ gives $\nu(\{x \in \mathcal{X} : m \in \mathbb{N} \text{ s.t. } P^m(x,S) \geq \delta\nu(\cdot)\}) = 0$. Then, by subadditivity of measures, we have

$$\nu(\overline{S}) \leq \sum_{i\neq j} \nu(\mathcal{X}_i \cap \mathcal{X}_j) + \nu(\{x \in \mathcal{X} : m \in \mathbb{N} \text{ s.t. } P^m(x,S) \geq \delta\nu(\cdot)\}) = 0.$$

It follows that

$$\nu\left(\bigcup_{i=1}^{d} \mathcal{X}'_i\right) = \nu\left(\bigcup_{i=1}^{d} \mathcal{X}_i\right) - \nu(\overline{S}) = \nu(\mathcal{X}) = 1.$$

Additionally, since $\nu \ll \pi$, we have $\pi\left(\bigcup_{i=1}^{d} \mathcal{X}'_i\right) > 0$. Since $X'_1, \ldots, X'_d$ have positive π-measure, the Markov chain is periodic with periodic decomposition $X'_1, \ldots, X'_d$. This contradicts the the assumption of aperiodicity. Therefore, $\gcd(T) = 1$. By [Bil95] p.541 (A numer theoretic fact), we can conclude that there is $n_* \in \mathbb{N}$ such that $T \supset \{n_*, n_* + 1, n_* + 2, \ldots\}$ as desired. □

Then, we are ready for the proof of Lemma 4.6.

Proof. Suppose that R is a $(n_0, \epsilon, \nu(\cdot))$-petite set. Then,

$$\sum_{i=1}^{n_0} P^i(x,\cdot) \geq \epsilon\nu(\cdot), \quad \forall x \in R.$$

Let T be as in Lemma A.1. Since

$$\sum_{i=1}^{n_0} \int_{x\in\mathcal{X}} \nu(dx)P^i(x,\cdot) \geq \epsilon\nu(\cdot), \quad \forall x \in R.$$

Hence, there is at least one $1 \leq i \leq n_0$ satisfying

$$\int_{x\in\mathcal{X}} \nu(dx)P^i(x,\cdot) \geq \epsilon\nu(\cdot), \quad \forall x \in R,$$

which implies T is not empty. By Lemma A.1, there is $n_* \in \mathbb{N}$ such that for all $n \geq n_*$, there exists $\delta_n > 0$ satisfying

$$\int_{x\in\mathcal{X}} \nu(dx)P^n(x,\cdot) > \delta_n \nu(\cdot).$$

Let $r \coloneqq \min\{\delta_n : n_* \leq n \leq n_* + n_0 - 1\}$, and set $N = n_* + n_0$. Then, for $x \in R$,

$$\begin{aligned} P^N(x,\cdot) &\geq \sum_{i=1}^{n_0} \int_{y\in\mathcal{X}} P^{N-i}(x,dy)P^i(y,\cdot) \\ &\geq \sum_{i=1}^{n_0} \int_{y\in R} r\nu(dy)P^i(y,\cdot) \\ &\geq \int_{y\in R} r\nu(dy)\epsilon\nu(\cdot) = r\epsilon\nu(\cdot). \end{aligned}$$

Therefore, R is $(N, r\epsilon, \nu(\cdot)$-small. □

References

[Bil95] P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995.

[BR00] L.A. Breyer and G.O. Roberts. From metropolis to diffusions: Gibbs states and optimal scaling. *Stochastic Processes and their Applications*, 90(2):181–206, 2000.

[CRR05] Ole F. Christensen, Gareth O. Roberts, and Jeffrey S. Rosenthal. Scaling limits for the transient phase of local metropolis-hastings algorithms. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):253–268, 2005.

[GGR97] A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110 – 120, 1997.

[GRS98] Walter R. Gilks, Gareth O. Roberts, and Sujit K. Sahu. Adaptive markov chain monte carlo through regeneration. *Journal of the American Statistical Association*, 93(443):1045–1054, 1998.

[Hä05] Olle Häggström. On the central limit theorem for geometrically ergodic markov chains. *Probability Theory and Related Fields - PROBAB THEORY RELAT FIELD*, 132:74–82, 05 2005.

[MT93] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[Rob98] Gareth O. Roberts. Optimal metropolis algorithms for product measures on the vertices of a hypercube. *Stochastics and Stochastics Reports*, 62:275–283, 1998.

[Ros06] Jeffrey S. Rosenthal. *A first look at rigorous probability theory*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition, 2006.

[RR97] Gareth Roberts and Jeffrey Rosenthal. Geometric Ergodicity and Hybrid Markov Chains. *Electronic Communications in Probability*, 2(none):13 – 25, 1997.

[RR01a] Gareth Roberts and Jeffrey Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16, 11 2001.

[RR01b] Gareth O. Roberts and Jeffrey S. Rosenthal. Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28(3):489–504, 2001.

[RR02] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal Scaling of Discrete Approximations to Langevin Diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1):255–268, 01 2002.

[RR04] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space markov chains and MCMC algorithms. *Probability Surveys*, 1(none), jan 2004.