Ergodicity of Adaptive MCMC and its Applications

by

Chao Yang

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Statistics
University of Toronto

Abstract

Ergodicity of Adaptive MCMC and its Applications

Chao Yang

Doctor of Philosophy

Graduate Department of Statistics

University of Toronto

2008

Markov chain Monte Carlo algorithms (MCMC) and Adaptive Markov chain Monte Carlo algorithms (AMCMC) are most important methods of approximately sampling from complicated probability distributions and are widely used in statistics, computer science, chemistry, physics, etc. The core problem to use these algorithms is to build up asymptotic theories for them.

In this thesis, we show the Central Limit Theorem (CLT) for the uniformly ergodic Markov chain using the regeneration method. We exploit the weakest uniform drift conditions to ensure the ergodicity and WLLN of AMCMC. Further we answer the open problem 21 in Roberts and Rosenthal [48] through constructing a counter example and finding out some stronger condition which indicates the ergodic property of AMCMC.

We find that the conditions (a) and (b) in [48] are not sufficient for WLLN holds when the functional is unbounded. We also prove the WLLN for unbounded functions with some stronger conditions.

Finally we consider the practical aspects of adaptive MCMC (AMCMC). We try some toy examples to explain that the general adaptive random walk Metropolis is not efficient for sampling from multi-model targets. Therefore we discuss the mixed regional adaptation (MRAPT) on the compact state space and the modified mixed regional adaptation on the general state space in which the regional proposal distributions are optimal and the switches between different models are very efficient. The theoretical proof is to show

that the algorithms proposed here fall within the scope of general theorems that are used to validate AMCMC. As an application of our theoretical results, we analyze the real data about the "Loss of Heterozygosity" (LOH) using MRAPT.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Jeffrey Rosenthal, who led me into the research of probability theory and guided me through the whole period of my Ph.D. study. His knowledge, inspiration, energy and patience will always serve to me as an example of a good researcher.

Many thanks are due to my co-supervisor Professor Radu Craiu for his guidance in this project and for teaching me so much. Prof. Craiu's patience, encouragement and enthusiasm are warmly appreciated.

I wish to thank Dr. Ajay Jasra for sharing his ideas, discussing with me in great detail. I would aslo like to thank my Ph.D. committee members Professor Jeremy Quastel and Professor Balint Virag for their many contributions to the improvement of this thesis.

Thank you to Ida Bulat, Laura Kerr and Andrea Carter who have always been very nice to help me sort out my administrative problems.

I would like to thank my dear colleagues in the Department of Mathematics and Department of Statistics for sharing their knowledge and for discussing with me.

I also enjoyed the friendship of many students at the University of Toronto and friends in my life for all the fun and happiness. I deeply cherish my experience in Toronto during the past five years.

Last but not the least, I am deeply grateful to my parents for their endless love, concern and support, and to my wife, Yuntian Fan for her much patience, support and understanding. Thank you!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to the Problems and the Conclusions of Thesis

MCMC algorithms are extremely widely used in statistical inference to sample from complicated high-dimensional distributions. The algorithms were first used in statistical physics and later in spatial statistics. For more history, one can see [30]. However it is very difficult to find the most efficient MCMC algorithm with respect to any target distribution. Adaptive MCMC algorithm is one direction developed recently to deal with this problem by tuning the associated parameters such as proposal variances through automatically "learning" from the history simulations. The most important issue before using both the MCMC algorithms and the adaptive MCMC algorithms is to prove the asymptotic theory of them. Another critical issue is to design efficient and reliable adaptive samplers for broad classes of problems.

This thesis consists of four results. We present the first main result (which is published as A. Jasra and C. Yang [33]) in chapter 3, which is to prove the open problem 3 in Roberts and Rosenthal [46]. In [46], the authors have proved that a central limit theorem (CLT) holds for h whenever $\pi(|h|^{2+\delta}) < \infty$ and $\delta > 0$ if the Markov chain is

geometrically ergodic using the regeneration methods. And they also proposed an open problem: to provide a regeneration proof of the CLT for h whenever $\pi(|h|^2) < \infty$. In Chapter 3, we deal with this open problem.

The second main result (see C. Yang [57]) is about the ergodicity of adaptive MCMC and presented in chapter 5. We study the relationship between the recurrence concept and the ergodicity of AMCMC. Through constructing counter examples and applying the splitting chain technique to the kernel family, we show the ergodic property of AMCMC under the uniform minimal drift conditions. Actually we partially tackle the open problem 20 in Roberts and Rosenthal [48]. The problem is stated as below:

Open Problem 20: Consider an adaptive MCMC algorithm with Diminishing Adaption, such that there is $C \in \mathcal{F}$, $V : \mathcal{X} \to [1, \infty)$,$\delta > 0$, and $b < \infty$, with $\sup_C V = \nu < \infty$, and:

(i) for each $\gamma \in \mathcal{Y}$, there exists a probability measure $\nu_\gamma(\cdot)$ on C with $P_\gamma(x, \cdot) \geq \delta\nu_\gamma(\cdot)$ for all $x \in C$; and

(ii)$P_\gamma V \leq V - 1 + b\mathbb{I}_C$ for each γ;

Suppose further that the sequence $\{V(X_n)\}_{n=0}^{\infty}$ is bounded in probability, given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$. Does the adaptive MCMC algorithm converge to the target distribution?

So far we can only prove the above conclusion with some additional conditions besides conditions (i) and (ii). Furthermore, we construct another counterexample to show that $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}$ being bounded in probability given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$ is not a necessary condition of ergodicity under the diminishing adaption assumption although it is sufficient. Following this conclusion, it seems that we should have a positive answer to the open problem 21 stated as below in Roberts and Rosenthal [48].

Open Problem 21: Consider an adaptive MCMC algorithm with Diminishing Adaptation such that for all $\epsilon > 0$, there is $m \in \mathbb{N}$ such that $P[M_\epsilon(X_n, \Gamma_n) < m \text{ i.o.}|X_0 = x_*, \Gamma_0 = \gamma_*] = 1$ where $M_\epsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \epsilon\}$. Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{Y}$.Does the adaptive MCMC algorithm converge to the target distribution?

However a negative answer to this problem is given by constructing a complicated counterexample. We also explore some stronger conditions than those in the open problem 21 which can ensure the ergodicity of AMCMC.

The third main result is on the WLLN of adaptive MCMC (see C. Yang [56]) and presented in chapter 6. We construct a counter example to show that Simultaneous Uniform Ergodicity Conditions and Diminishing Adaption Conditions are not enough to have WLLN hold for unbounded functions. However we can prove the WLLN for unbounded functions under the conditions of corollary 11 in Roberts and Rosenthal [48]. Further we extend the WLLN for HST algorithm from bounded functions to unbounded functions as an application.

The fourth result (see R.V. Craiu, J.S.Rosenthal and C. Yang [14] and [58]) is concerned with the practical aspects of adaptive MCMC, particularly related to sampling from multi-model distributions. Since the random walk Metropolis is one of the mostly used algorithms in practice it is the aim for most of our theoretical results. The regional adaptation algorithms proposed in chapter 7 and chapter 8 are discussed in the context of two separate regions. We conduct some real data analysis using our mixed regional adaptive MCMC algorithm and compare the efficiency of different adaptive MCMC algorithms by simulating some toy examples.

Chapter 9 concludes the thesis and summarizes some future work directions.

In chapter 2 we introduce the MCMC algorithm and some relevant Markov Chain theories.

In chapter 4 we outline the constructions, notations and ergodicity theories of adaptive MCMC algorithms.

Chapter 2

Markov Chain and MCMC Algorithms

2.1 Why we need MCMC

Most applications of MCMC ([36], [29])are applied to the Bayesian Statistics Computations. From a Bayesian point of view, observables and parameters of a statistical model are all considered random quantities. Suppose D denotes the observations, and θ denotes model parameters and missing data. The joint distribution $P(D, \theta)$ consists of a *prior* distribution $P(\theta)$ and a *likelihood* $P(D|\theta)$ as

$$P(D, \theta) = P(D|\theta)P(\theta).$$

Having observed D, we have

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$

as the distribution of θ conditional on D, which is the *posterior* distribution of θ and is the object of all Bayesian inference. Any features of the posterior distribution are legitimate for Bayesian inference: moments, quantiles, highest posterior density regions, etc. All these quantities can be expressed in terms of posterior expectations of functions

of θ. The posterior expectation of a function $f(\theta)$ is

$$E[f(\theta)|D] = \frac{\int f(\theta)P(\theta)P(D|\theta)d\theta}{\int P(\theta)P(D|\theta)d\theta}.$$

The integrations in this expression have brought difficulties in the applications of Bayesian inference, especially in high dimensional cases. Analytic method to do direct integration $E[f(\theta)|D]$ is infeasible. Numerical evaluation of $E[f(\theta)|D]$ as an alternative method is difficult and inaccurate when the dimension is greater than about twenty. Therefore good estimates of expectations allow Bayesian inference to be used to estimate a variety of parameters, probabilities, means, etc. Monte Carlo integration evalutes $E[f(X)]$ by simulating $i.i.d$ random variables $\{X_i,\ i = 1, \cdots, n\}$ from $\pi(\cdot)$, then

$$E[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

So we use the sample mean to esitimate the mean of $f(X)$. When the samples $\{X_i\}$ are independent, if we increase the sample size n, the approximation will tend to be more accurate according to the laws of large numbers. However, drawing samples $\{X_i\}$ independently from $\pi(\cdot)$ is not feasible generally. Since $\{X_i\}$ do not necessarily need to be independent, one method of generating the samples is through a Markov chain having $\pi(\cdot)$ as its stationary distribution. This method is called *Markov chain Monte Carlo(MCMC).* MCMC has been proven to be an extremely helpful method of approximately sampling from distribution $\pi(\cdot)$ on the state space $\mathcal{X}$, especially when $\pi(\cdot)$ is very high-dimensional or too complicated to do the direct sampling. Actually the existence of MCMC algorithms has transformed Bayesian inference by allowing practitioners to sample from some simple distributions of complicated statistical models (see [53], [51],[55],[43]).

Suppose we want to sample from some complicated distribution π. The main idea of general MCMC algorithm is to construct a Makov chain $\{X_i\}_{i=1}^n$ using some simple proposal distribution Q such that $\mathcal{L}(X_n) \approx \pi(\cdot)$ when n is large enough. In fact it is very straightforward to realize such an idea. For more precise descriptions, see section 2.7. Then we can estimate the integral $\int f(x)\pi(dx)$ using $\frac{1}{n}\sum_{i=1}^n f(X_i)$. We note that when

we use the MCMC algorithm, we only need to generate samples from the much simpler distribution Q, rather than from the complicated distribution π. This idea makes the numerical computation of $E[f(\theta)|D]$ much easier and more efficient. Such good estimates make Bayesian inference much more widely applicable.

Furthermore, a wide variety of the Markov Chain's asymptotic theories are developed to prove the validity of the MCMC algorithms and to estimate the errors of them. We will introduce these theories in later sections.

In practice, to remove the impact of starting values, we usually use $\frac{1}{n-N}\sum_{i=N+1}^{n} f(X_i)$ as the estimate of $\int f(x)\pi(dx)$ for some $0 < N < n$ and N being large enough.

2.2 Definition of Markov Chain

The application of MCMC algorithms raise numerous questions related to the mathematical theory of Markov chain. Now let us recall the definition of *Transition Probability Kernels*(see [37]), $\mathcal{B}(\mathcal{X})$ will be taken as the Borel $\sigma-$field.

Definition 2.1. *If $P = \{P(x, A),\ x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$ is such that:*
(i) *for each $A \in \mathcal{B}(\mathcal{X})$, $P(\cdot, A)$ is a non-negative function on $\mathcal{X}$;*
(ii) *for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $\mathcal{B}(\mathcal{X})$,*
then we call P a Transition Probability Kernels or Markov transition function.

Definition 2.2. *A Markov chain $\boldsymbol{X} = \{X_0, X_1, \cdots\}$ is a particular type of stochastic process taking, at times $n \in \mathbb{Z}_+$, initial distribution μ and transition probability $P(x, A)$ such that $X_0 \sim \mu(\cdot)$ and*

$$P_\mu(X_{n+1} \in A|X_n = x, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0) = P(X_{n+1} \in A|X_n = x) = P(x, A).$$

We will use $P^n(x, A)$ represents the probability of jumping from x to somewhere in A after n iterations. Obviously we have:

$$P^n(x, A) = \int_{\mathcal{X}} P(y, A)P^{n-1}(x, dy).$$

In this chapter we will summarize some basis definitions and related theoretical results of Markov chain which we will use in the next chapters. And in the last section we will introduce the Metopolis-Hasting algorithms. All the results in this chapter can be found in Meyn and Tweedie [37], Roberts and Rosenthal [46].

2.3 Irreducible, Atom, Minorization Condition and Small Set

Much general Markov chain theory can be developed in complete analogy with the situation when $\mathcal{X}$ contains an atom for the $\varphi-$irreducible chain X(see [37]). Let us recall the definition of *Return time to* A first, for any $A \in \mathcal{B}^+(\mathcal{X})$,

$$\tau_A := \min\{n \geq 1 : X_n \in A\}.$$

Then we can give the definition of $\varphi-$*irreducible* chain.

Definition 2.3. *We call* $X = \{X_n\}$ $\varphi-$*irreducible if there exists a measure* φ *on* $\mathcal{B}(\mathcal{X})$ *such that, whenever* $\varphi(A) > 0$*, we have* $P(\tau_A < \infty | X_0 = x) > 0$ *for all* $x \in \mathcal{X}$*.*

Next we introduce the definition of *atom*:

Definition 2.4. *A set* $\alpha \in \mathcal{B}(\mathcal{X})$ *is called an atom for* X *if there exists a measure* μ *on* $\mathcal{B}(\mathcal{X})$ *such that:*

$$P(x, A) = \mu(A), \ x \in \alpha.$$

If X *is* $\phi-$*irreducible and* $\phi(\alpha) > 0$*, then* α *is called an accessible atom.*

Obviously each point in $\mathcal{X}$ is an atom. However we also to find some conditions under which we can construct an artificial atom. Actually we need the *Minorization Condition* as below:

Minorization Condition For some $\delta > 0$, some $C \in \mathcal{B}(\mathcal{X})$ and some probability

measure ν with $\nu(C^c) = 0$ and $\nu(C) = 1$, $P(x, A) \geq \delta \mathbb{I}_C(x)\nu(A)$.

Then we can split any Markov chain with the Minorization Condition. We first split the space $\mathcal{X}$ itself by writing $\check{\mathcal{X}} = \mathcal{X} \times \{0, 1\}$, where $\mathcal{X}_0 = \mathcal{X} \times \{0\}$ and $\mathcal{X}_1 = \mathcal{X} \times \{1\}$ are thought of as copies $\mathcal{X}$ equipped with copies $\mathcal{B}(\mathcal{X}_0)$, $\mathcal{B}(\mathcal{X}_1)$ of the $\sigma-$ field $\mathcal{B}(\mathcal{X})$. We also let $\mathcal{B}(\check{\mathcal{X}})$ be the $\sigma-$field of $\check{\mathcal{X}}$ generated by $\mathcal{B}(\mathcal{X}_0)$, $\mathcal{B}(\mathcal{X}_1)$: that is $\mathcal{B}(\check{\mathcal{X}})$ is the smallest $\sigma-$field containing sets of the form $A_0 := A \times \{0\}$, $A_1 := A \times \{1\}$, $A \in \mathcal{B}(\mathcal{X})$.

We will write x_i, $i = 0, 1$ for elements of $\check{\mathcal{X}}$, with x_0 denoting members of the upper level $\mathcal{X}_0$ and x_1 denoting members of the lower level $\mathcal{X}_1$.

If λ is any measure on $\mathcal{B}(\mathcal{X})$, then the next step in the construction is to split the measure λ into two measures on each of $\mathcal{X}_0$ and $\mathcal{X}_1$ by defining the measure λ^* on $\mathcal{B}(\check{\mathcal{X}})$ through

$$\begin{aligned} \lambda^*(A_0) &= \lambda(A \cap C)[1 - \delta] + \lambda(A \cap C^c), \\ \lambda^*(A_1) &= \lambda(A \cap C)\delta, \end{aligned}$$

where C, δ and ν are the set, the constant and the measure in the Minorization Condition. Note that the splitting is dependent on the choice of the set C, and although in general the set chosen is not relevant. We can observe the λ is the marginal measure induced by λ^*, in the sense that for any A in $\mathcal{B}(\mathcal{X})$ we have:

$$\lambda^*(A_0 \cup A_1) = \lambda(A).$$

Now we can step in the construction to the split the chain $\{X_n\}$ to the form a chain $\{\check{X}_n\}$ which lives on $(\check{\mathcal{X}}, \mathcal{B}(\check{X})$. Define the split kernel $\check{P}(x_i, A)$ for $x_i \in \check{\mathcal{X}}$ and $A \in \mathcal{B}(\check{X})$ by:

$$\begin{aligned} \check{P}(x_0, \cdot) &= P(x, \cdot)^*, \quad x_0 \in \mathcal{X}_0 - C_0; \\ \check{P}(x_0, \cdot) &= [1 - \delta]^{-1}[P(x, \cdot)^* - \delta\nu^*(\cdot)], \quad x_0 \in C_0; \\ \check{P}(x_1, \cdot) &= \nu^*(\cdot), \quad x_1 \in \mathcal{X}_1. \end{aligned}$$

We can see that outside C the chain $\{\check{X}_n\}$ behaves like $\{X_n\}$, moving on the "top" half $\mathcal{X}_0$ of the split space. Each time it arrives in C, it is "split"; with probability $1 - \delta$ it

remains in C_0, with probability δ it drops to C_1.

It is critical to note that the bottom level $\mathcal{X}_1$ is an atom with $\psi^*(X_1) == \delta\psi(C) > 0$ whenever the original chain is $\psi-$irreducible. We also have $\check{P}^n(x_i, \mathcal{X}_\infty - C_1) = 0$ for all $n \geq 1$ and all $x_i \in \check{\mathcal{X}}$, so that the atom $C_1 \subseteq \mathcal{X}_1$ is the only part of the bottom level which is reached with positive probability. We will use the notation $\check{\alpha} := C_1$ when we wish to emphasize the fact that all transitions out of C_1 are identical, so that C_1 is an atom in $\check{\mathcal{X}}$. Following Meyn and Tweedie [37] we have the following theorem:

Theorem 2.1. ***(i)*** *The chain X is the marginal chain of $\{\check{X}\}$: that is, for any initial distribution λ on $\mathcal{B}(\mathcal{X})$ and any $A \in \mathcal{B}(\mathcal{X})$,*

$$\int_{\mathcal{X}} \lambda(dx)P^k(x, A) = \int_{\check{\mathcal{X}}} \lambda^*(dy_i)\check{P}^k(y_i, A_0 \cup A_1)$$

(ii) *The chain X is $\varphi-$irreducible if $\{\check{X}\}$ is $\check{\varphi}-$irreducible; and if X is is $\varphi-$irreducible and $\varphi(C) > 0$ then $\{\check{X}\}$ is ν^*-irreducible, and $\check{\alpha}$ is an accessible atom for the split chain.*

Finally we will introduce the definition of *Small Sets*

Definition 2.5. *A set $C \in \mathcal{B}(\mathcal{X})$ is called a Small Sets if there exists an $m > 0$, and a non-trivial measure ν_m on $\mathcal{B}(\mathcal{X})$, such that for all $x \in C$, $B \in \mathcal{B}(\mathcal{X})$,*

$$P^m(x, B) \geq \nu_m(B).$$

Then we say that C is ν_m-small.

In fact, for a $\psi-$irreducible chain, every set $A \in \mathcal{B}^+(\mathcal{X})$ contains a small set in $\mathcal{B}^+(\mathcal{X})$. As a consequence, every $\psi-$irreducible chain admits some $m-$skeleton which can be split, and for which the atomic structure of the split chain can be exploited. We will use this idea to a family of Markov chain in the chapter 4, so that we can use the common atomic structure to prove the ergodicity of Adaptive Monte Carlo Markov chain algorithms(AMCMC).

Finally we introduce a generalization of small sets, *petite sets*. Let $a = \{a(n)\}$ be a

distribution, or probability measure, on $\mathbb{Z}_+$, and consider the Markov chain X_a with probability transition kernel

$$K_a(x, A) := \sum_{n=0}^{\infty} P^n(x, A)a(n),\ x \in A,\ A \in \mathcal{B}(\mathcal{X}).$$

Definition 2.6. *We call a set $C \in \mathcal{B}(\mathcal{X})$ ν_a-petite if the sampled chain satisfies the bound*

$$K_a(x, B) \geq \nu_a(B),$$

for all $x \in C$, $B \in \mathcal{B}(X)$, where ν_a is non-trivial measure on $\mathcal{B}(X)$.

2.4 Recurrence, Transience and Drift Conditions

In this section we will introduce the definition of *recurrence* and *transience* which are used to describe type of weak forms of stability. What we concern is actually the behavior of the occupation time random variable

$$\eta_A := \sum_{n=1}^{\infty} \mathbb{I}\{X_n \in A\},$$

which counts the number of visits to a set A. In terms of η_A we can study a chain through the transience and recurrence of its sets.

Definition 2.7. *The set A is called uniformly transient if for there exists $M < \infty$ such that $E_x[\eta_A] \leq M$ for all $x \in A$.*
The set A is called recurrent if $E_x[\eta_A] = \infty$ for all $x \in A$.

Using the definition of *uniformly transient* and *recurrent* of the sets we can define *recurrent chain* and *transient chain* and have the following theorem (see [37]):

Theorem 2.2. *Suppose that X is $\psi-$irreducible Markov chain. Then either*
(i) *every set in $\mathcal{B}^+(\mathcal{X})$ is recurrent, in which case we call X recurrent; or*
(ii) *there is a countable cover of X with uniformly transient sets, in which case we call X transient; and every petite set is uniformly transient.*

We can check the transience and recurrence through computing the expected drift defined by the one-step transition function P. The *Drift Markov Chains* is defined as:

Definition 2.8. *The drift operator Δ is defined for any non-negative measurable function V by*

$$\Delta V(x) := \int P(x, dy)V(y) - V(x), \quad x \in \mathcal{X}.$$

Based on the drift function, we can develop the criteria for both transience and recurrence (see [37]).

Theorem 2.3. *Suppose X is a $\psi-$irreducible chain.*
(i) *The chain X is transient if and only if there exists a bounded non-negative function V and a set $C \in \mathcal{B}^+(\mathcal{X})$ such that for any $x \in C^c$,*

$$\Delta V(x) \geq 0$$

and

$$D = \{V(x) > \sup_{y \in C} V(y)\} \in \mathcal{B}^+(\mathcal{X}).$$

(ii) *The chain is recurrent if there exists a petite set $C \subset \mathcal{X}$, and a function V which is unbounded off petite sets in the sense that $C_V(n) := \{y : V(y) \leq n\}$ is petite for all n, such that*

$$\Delta V(x) \leq 0, \quad x \in C^c.$$

2.5 Invariant Measure and Ergodicity

For many purposes, we might require that the distribution of X_n does not change as n takes on different values. Based on the Markov property it follows that the finite dimensional distributions of X are invariant under translation in time. Therefore we will consider the definition of *Invariant Measure*.

Definition 2.9. *A $\sigma-$finite measure $\pi(\cdot)$ on $\mathcal{B}(\mathcal{X})$ with the property*

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

will be called invariant.

Regarding the construction of invariant measure, we have the following theorem (see [37]):

Theorem 2.4. *If the chain X is recurrent then it admits a unique (up to constant multiples) invariant measure π, and the measure π has the representation, for any $A \in \mathcal{B}^+(\mathcal{X})$*

$$\pi(B) = \int_A \pi(d\omega)E_\omega[\sum_{n=1}^{\tau_A} \mathbb{I}\{X_n \in B\}], \quad B \in \mathcal{B}(\mathcal{X}).$$

The invariant measure π is finite if there exists a petite set C such that

$$\sup_{x \in C} E_x[\tau_C] < \infty.$$

Following these results above we have the definition of *Positive and Null Chains*

Definition 2.10. *Suppose that X is $\psi-$irreducible, and admits an invariant probability measure π. Then X is called a positive chain.*
If X does not admit such a measure, then we call X null.

Before we introduce the main theorem, we need to define some notations:

Definition 2.11. *Given Markov chain transition probabilities P on a state space $\mathcal{X}$, and a measurable function $f : \mathcal{X} \to R$, define the function $Pf : \mathcal{X} \to R$ such that $(Pf)(x)$ is the conditional expected value of $f(X_{n+1})$, given that $X_n = x$. In symbols,*

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

Now we can introduce the *Aperiodic Ergodic Theorem*(see [37]):

Theorem 2.5. *Suppose that X is an aperiodic Harris recurrent chain, with invariant measure π. The following are equivalent:*

(i) *The chain is positive Harris: that is, the unique invariant measure π is finite.*

(ii) *There exists some $\nu-$small set $C \in \mathcal{B}^+(\mathcal{X})$ and some $P^\infty(C) > 0$ such that as $n \to \infty$, for all $x \in C$,*

$$P^n(x, C) \to P^\infty(C).$$

(iii) *There exists some regular set in $\mathcal{B}^+(\mathcal{X})$: equivalently, there is a petite set $C \in \mathcal{B}(\mathcal{X})$ such that*

$$\sup_{x \in C} E_x[\tau_C] < \infty.$$

(iv) *There exist some petite set C, some $b < \infty$ and a non-negative function V finite at some one $x_0 \in \mathcal{X}$, satisfying*

$$\Delta V(x) := PV(x) - V(x) \leq -1 + b\mathbb{I}_C(x), \quad x \in \mathcal{X}.$$

Any of these conditions is equivalent to the existence of a unique invariant probability measure π such that for every initial condition $x \in \mathcal{X}$,

$$\sup_{A \in \mathcal{B}(\mathcal{X})} |P^n(x, A) - \pi(A)| \to 0$$

as $n \to \infty$, and moreover for any regular initial distribution λ, μ,

$$\sum_{n=1}^{\infty} \int \int \lambda(dx)\mu(dx) \sup_{A \in \mathcal{B}(\mathcal{X})} |P^n(x, A) - \pi(A)| < \infty.$$

We also describe the above convergence in terms of the *total variation norm between two probability measures* (see [46]).

Definition 2.12. *The total variation norm between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is*

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

Next we will list some simple properties of total variation distance (see [46], [52]) we will use in the further chapters.

Proposition 2.1. ***(a)*** *If $\pi(\cdot)$ is stationary for a Markov chain kernel P, then $\|P^n(x,\cdot)-\pi(\cdot)\|$ is non-increasing in n, i.e. $\|P^n(x,\cdot)-\pi(\cdot)\| \leq \|P^{n-1}(x,\cdot)-\pi(\cdot)\|$ for $n \in \mathbb{N}$.*

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

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

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

(d) *Given probability measure $\mu(\cdot)$ and $\nu(\cdot)$, there are jointly defined random variable X and Y such that $X \sim \mu(\cdot)$, $Y \sim \nu(\cdot)$, and $P[X=Y] = 1 - \|\mu(\cdot)-\nu(\cdot)\|$.*

2.6 Geometrically Ergordic And Uniformly Ergodic

In lots of situations, what we concern is the convergence speed of P^n as $n \to \infty$. One typical convergence rate property is *geometrically ergodic*.

Definition 2.13. *A Markov chain with stationary distribution $\pi(\cdot)$ is geometrically ergodic if*

$$\|P^n(x,\cdot)-\pi(\cdot)\| \leq M(x)\rho^n,$$

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

Next we discuss conditions which ensure geometric ergodicity, first let us consider another *drift condition*

Definition 2.14. *A Markov chain satisfies a drift condition II if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1,\infty)$ such that*

$$(PV)(x) \leq \lambda V(x) + b\mathbb{I}_C(x),$$

for all $x \in \mathcal{X}$.

We have the following *Geometric Ergodic Theorem*(see [37])

Theorem 2.6. ***Geometric Ergodic Theorem*** *Suppose that the chain X is $\psi-$irreducible and aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Suppose $C \subset \mathcal{X}$ is $(n_0, \epsilon, \nu)-$small set. Suppose further that the drift condition II 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 $\pi - a.e\ x \in \mathcal{X}$. Then the chain is geometrically ergodic.*

Another "qualitative" convergence rate property is *uniform ergodicity*:

Definition 2.15. *A Markov chain having 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 $M < \infty$.

The equivalences of uniform ergodicity are as the following theorem:

Theorem 2.7. *For any Markov chain X the following are equivalent:*

(i) X is uniformly ergodic.

(ii) For some $n \in \mathbb{Z}_+$,

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

(iii) The chain is aperiodic and Doeblin's Condition holds: that is, there is a probability measure ϕ on $\mathcal{B}(\mathcal{X})$ and $\epsilon < 1$, $\delta > 0$, $m \in \mathbb{Z}_+$ such that whenever $\phi(A) > \epsilon$,

$$\inf_{x \in \mathcal{X}} P^m(x, A) > \delta.$$

(iv) The state space $\mathcal{X}$ is μ_m-small for some m.

(v) The chain is aperiodic and there is a petite set C with

$$\sup_{x \in \mathcal{X}} E_x[\tau_C] < \infty,$$

in which case every $A \in \mathcal{B}^+(\mathcal{X})$,

$$\sup_{x \in \mathcal{X}} E_x[\tau_A] < \infty.$$

(vi) The chain is aperiodic and there is a petite set C *and a* $\kappa > 1$ *with*

$$\sup_{x \in \mathcal{X}} E_x[\kappa^{\tau_C}] < \infty,$$

in which case for every $A \in \mathcal{B}^+(\mathcal{X})$ *we have for some* $\kappa_A > 1$,

$$\sup_{x \in \mathcal{X}} E_x[\kappa_A^{\tau_A}] < \infty.$$

(vii) The chain is aperiodic and there is a bounded solution $V \geq 1$ *to*

$$\Delta V(x) \leq -\beta V(x) + b\mathbb{I}_C(x), \ x \in \mathcal{X},$$

for some $\beta > 0$, $b < \infty$, *and some petite set* C.

Under (iv), *we have in particular that for any* x,

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq \rho^{n/m},$$

where $\rho = 1 - \nu_m(\mathcal{X})$.

2.7 Metropolis-Hasting Algorithm

The Metropolis-Hastings algorithm([36], [29]) is an extremely important MCMC algorithm to sample from complicated probability distribution. Before we introduce how to construct the Markov chain using this algorithm, let us learn the definition of *reversible* first.

Definition 2.16. *A Markov chain on a state space* $\mathcal{X}$ *is reversible with respect to a probability distribution* $\pi(\cdot)$ *on* $\mathcal{X}$ *if*

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

It is easy to prove the following proposition of reversibility (see [46]).

Proposition 2.2. *If Markov chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is stationary for the chain.*

From the above proposition, we only need to create a Markov chain which is easily run, and which is reversible with respect to $\pi(\cdot)$. The simplest way to do this is to use the Metropolis-Hastings algorithm. Suppose that $\pi(\cdot)$ has a density π_λ, and $Q(x, dy)$ is any transition kernel of some Markov chain such that $Q(x, dy) \propto q(x, y)dy$. Then the Metropolis-Hastings algorithm proceeds as below:

(i) Choose some initial value X_0;

(ii) Given $X_n = x_n$, generate a *proposal* Y_{n+1} following the distribution $Q(x_n, \cdot)$. That is $Y_{n+1} \sim Q(x_n, \cdot)$;

(iii) Compute the acceptance rate $\alpha(X_n, Y_{n+1})$ as

$$\alpha(x, y) = \min[1, \frac{\pi_\lambda(y)q(y, x)}{\pi_\lambda(x)q(x, y)}].$$

(iv) We will accept the proposal by setting $X_{n+1} = Y_{n+1}$ with probability $\alpha(X_n, Y_{n+1})$; otherwise reject the proposal by setting $X_{n+1} = X_n$ with probability $1 - \alpha(X_n, Y_{n+1})$.

Proposition 2.3. *The Metropolis-Hastings algorithm (as described above) produces a Markov chain $\{X_n\}$ which is reversible with respect to $\pi(\cdot)$.*

Chapter 3

Central Limit Theorems for Markov Chains

3.1 Introduction

Let $\{X_n\}$ be a Markov chain on measurable space $(\mathcal{X}, \mathcal{E})$ with unique stationary distribution π. Let $h : \mathcal{X} \to \mathbb{R}$ be a measurable function with finite stationary mean $\pi(h) := \int_{\mathcal{X}} h(x)\pi(dx)$. Ibragimov and Linnik [1](1971) proved that if $\{X_n\}$ is geometrically ergodic, then a central limit theorem (CLT) holds for h whenever $\pi(|h|^{2+\delta}) < \infty$, $\delta > 0$. Cogburn [12](1972) proved that if a Markov chain is uniformly ergodic, with $\pi(h^2) < \infty$ then a CLT holds for h. The first result was re-proved in Roberts and Rosenthal [46](2004) using a regeneration approach; thus removing many of the technicalities of the original proof. This raised an open problem: to provide a proof of the second result using a regeneration approach. In this chapter we will provide a solution to this problem after we discuss the some results on CLT for Markov Chains.

3.2 Some Discussions

Let $\{X_n\}$ be a Markov chain with transition kernel $P : \mathcal{X} \times \mathcal{E} \to [0,1]$ and a unique stationary distribution π. Let $h : \mathcal{X} \to \mathbb{R}$ be a real-valued measurable function. We say that h satisfies a Central Limit Theorem (or $\sqrt{n}-$CLT) 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, where $N(0, \sigma^2)$ is a Gaussian distribution with zero mean and variance σ^2 (we allow that $\sigma^2 = 0$), and (e.g. Chan and Geyer [11](1994), see also Bradley [9](1985) and Chen [11](1999))

$$\sigma^2 \quad = \quad \pi(h^2) + 2 \int_E \sum_{n=1}^{\infty} h(x) P^n(h)(x) \pi(dx).$$

When the Markov chain is uniformly ergodic, we have the following theorem:

Theorem 3.1 (Cogburn [12], 1972)**.** *If a Markov chain with stationary distribution π is uniformly ergodic, then a $\sqrt{n}-$CLT holds for h whenever $\pi(h^2) < \infty$.*

Ibragimov and Linnik [1](1971) proved a CLT for h when the chain is geometrically ergodic and, for some $\delta > 0$, $\pi(|h|^{2+\delta}) < \infty$. Roberts and Rosenthal [46] (2004) provided a simpler proof using regeneration arguments. In addition, Roberts and Rosenthal [46](2004) left an open problem: To provide a proof of Theorem 3.1 (originally proved by Cogburn [12](1972)) using regeneration.

Many of the recent developments of CLTs for Markov chains are related to the evolution of stochastic simulation algorithms such as Markov chain Monte Carlo (MCMC). For example, Roberts and Rosenthal (2004) posed many open problems, including that considered here, for CLTs; see Häggström [28](2005) for a solution to another open problem. Additionally, Jones (2004) discusses the link between mixing processes and CLTs, with MCMC algorithms a particular consideration. For an up-to-date review of CLTs for Markov chains see: Bradley [9](1985), Chen [11](1999) and Jones [34](2004).

The proof of Theorem 3.1, using regeneration theory, provides an elegant framework for the proof of CLTs for Markov chains. The approach may also be useful for alternative

proofs of CLTs for chains with different ergodicity properties; e.g. polynomial ergodicity (see Jarner and Roberts [31] (2002)).

Remark: Actually the CLT may hold for some Markov chain without ergodic property with respect to its stationary distribution. We can consider an example such that the state space $\mathcal{X} = \{1, 2, 3, 4\}$ with the stationary distribution $\pi(1) = \pi(2) = \pi(3) = \pi(4) = \frac{1}{4}$ and the transition matrix $P = \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{pmatrix}$ Then we can prove that P stated above is reversible and π stated above is the unique stationary distribution to P. And for every $h : \mathcal{X} \to R$ with $\pi(h^2) < \infty$ satisfies a CLT for the P as stated above. On the other hand, since the eigenvalue of P is $1, -1, 0, 0$, we have $\lim_{n\to\infty} P^n$ does NOT exist. Therefore P is NOT ergodic.

The structure of this chapter is as below. In Section 3.3 we provide some background knowledge the regeneration construction, we also detail some technical results. In Section 3.4 we use the results of the previous Section to provide a proof of Theorem 3.1 using regenerations.

3.3 Regeneration Construction and Some related Technical Results

Now we consider the regeneration construction for the proof. Since $\mathcal{X}$ is small we use the split chain construction (Nummelin, 1984), for any $x \in \mathcal{X}$, $A \in \mathcal{E}$

$$P^m(x, A) \;=\; (1 - \epsilon)R(x, A) + \epsilon\nu(A),$$

where $R(x, A) = (1 - \epsilon)^{-1}[P^m(x, A) - \epsilon\nu(A)]$. That is, for a single chain (X_n), with probability ϵ we choose $X_{n+m} \sim \nu$, while with probability $1 - \epsilon$ we choose $X_{n+m} \sim$

$R(X_n, \cdot)$, if $m > 1$, we fill in the missing values as X_{n+1} using the appropriate Markov kernel and conditionals.

We let $T_1, T_2, \ldots$ be the regeneration times, i.e. the times such that $X_{T_i} \sim \nu$, clearly $T_i = im$. Let $T_0 = 0$ and $r(n) = \sup\{i \geq 0 : T_i \leq n\}$, using the regeneration time, we can break up the sum $\sum_{i=0}^{n}[h(X_i) - \pi(h)]$ into sums over tours as follows:

$$\sum_{i=0}^{n}[h(X_i) - \pi(h)] = \sum_{j=1}^{r(n)} \sum_{i=T_j}^{T_{j+1}-1} [h(X_i) - \pi(h)] + Q(n),$$

where

$$Q(n) = \sum_{j=0}^{T_1-1}[h(X_j) - \pi(h)] + \sum_{T_{r(n)+1}}^{n} [h(X_j) - \pi(h)].$$

We begin our construction, by noting the following result.

Lemma 3.1. *Under the formulation above, we have that:*

$$\frac{Q(n)}{n^{1/2}} \longrightarrow_p 0. \tag{3.1}$$

Proof. Let

$$\begin{aligned} Q_1^+(n) &= \sum_{j=0}^{T_1-1}[h(X_j) - \pi(h)]^+, \\ Q_1^-(n) &= \sum_{j=0}^{T_1-1}[h(X_j) - \pi(h)]^- \end{aligned}$$

and

$$\begin{aligned} Q_2^+(n) &= \sum_{T_{r(n)}+1}^{n} [h(X_j) - \pi(h)]^+, \\ Q_2^-(n) &= \sum_{T_{r(n)}+1}^{n} [h(X_j) - \pi(h)]^-, \end{aligned}$$

where $[h(X_j) - \pi(h)]^+ = \max\{h(X_j) - \pi(h), 0\}$ and $[h(X_j) - \pi(h)]^- = \max\{-[h(X_j) - \pi(h)], 0\}$.

The strategy of the proof is to show that $Q_i^{\pm}(n)/n^{1/2} \to_p 0$ as $n \to \infty$. Consider $Q_1^+(n)$,

$$Q_1^+(n) = \sum_{j=0}^{sm-1} [h(X_j) - \pi(h)]^+ \quad \text{w.p} \quad \epsilon(1-\epsilon)^{(s-1)}, \tag{3.2}$$

where $s \in \mathbb{N}$. If $Q_1^+(n)/n^{1/2} \to_p 0$, i.e. $\mathbb{P}(\exists \epsilon, Q_1^+(n) > \epsilon n^{1/2}, \text{i.o.}) = 1$ for all n, which means that $\mathbb{P}(Q_1^+(n) = \infty, \text{i.o.}) = 1$, which is impossible from (3.2). So $Q_i^+(n)/n^{1/2} \to_p 0$ as $n \to \infty$. Similarly $Q_i^-(n)/n^{1/2} \to_p 0$ as $n \to \infty$.

For Q_2 we have $Q_2^+(n) \leq \sum_{j=r_n+1}^{l_n} [h(X_j) - \pi(h)]^+ = \widetilde{Q}_2^+(n)$, where $l(n) = \inf\{i \geq 0 : T_i \geq n\}$. We know that $\widetilde{Q}_2^+(n)$ has the same distribution with $Q_2^+(n)$, so $\widetilde{Q}_i^+(n)/n^{1/2} \to_p 0$ as $n \to \infty$ and therefore, $Q_2^+(n)/n^{1/2} \to_p 0$ as $n \to \infty$. Similarly $Q_2^-(n)/n^{1/2} \to_p 0$ as $n \to \infty$. From the above discussion, we conclude that $Q(n)/n^{1/2} \to_p 0$. □

The above lemma indicates that our objective is to find the asymptotic distribution of $\sum_{j=1}^{r(n)} \sum_{i=T_j}^{T_{j+1}-1} [h(X_i) - \pi(h)]$. Given the definition of T_i, each random variable $s_j = \sum_{i=T_j}^{T_{j+1}-1} [h(X_i) - \pi(h)]$ has same distribution. However, we know that T_j depends on $X_{T_{j-1}+1}, \cdots, X_{T_{j-1}-1}$, but does not depend on the value of $X_{T_{j-1}}$. That is, we have the following lemma:

Lemma 3.2. *For any $0 \leq i < \infty$, s_i and s_{i+1} are not independent, but the two collections of random variables: $\{s_i : 0 \leq i \leq m-2\}$ and $\{s_i : i \geq m\}$ are independent for any $m \geq 2$.Therefore the random variable sequence $\{s_i\}_{i=0}^{\infty}$ is a one-dependent stationary stochastic processes.*

Proof. Clearly s_{i+1} depends on the distribution T_{i+1}, thus:

$$\begin{aligned} & \mathbb{P}\bigg(X_{T_i+1} \in dx_1, \cdots, X_{T_i+m} \in dy | X_{T_i} x, T_{i+1} - T_i > m\bigg) \\ = \quad & \frac{(1-\epsilon)R(x,dy)}{P^m(x,dy)} P(x,dx_1) \cdots P(x_{m-1}, dy) \end{aligned}$$

and

$$\mathbb{P}\bigg(X_{T_i+1} \in dx_1, \cdots, X_{T_i+m} \in dy | X_{T_i} = x, T_{i+1} - T_i = m\bigg) = \frac{\epsilon \nu(dy)}{P^m(x,dy)} P(x,dx_1) \cdots P(x_{m-1}, dy).$$

Note s_i depends on T_{i+1}. Therefore s_i and s_{i+1} are not independent. However, for any $0 \leq i \leq m-2 < m \leq j < \infty$, since $X_{T_i} \sim \nu(\cdot)$ and X_{T_j} depends $X_{T_{j-1}+1}, \cdots, X_{T_j-1}$, but is independent of all the $\{X_k : k \leq T_j\}$. Thus, we have the result. □

3.4 Proof of Theorem 3.1

To prove the Theorem 3.1 we follow the strategy:
Step 1: Prove that $I = E_\nu\left(\sum_{i=0}^{T_1-1}[h(X_i) - \pi(h)] \right) = 0$.
Step 2: Prove that $J = \int_{\mathcal{X}} \nu(dx)\mathbb{E}\left[\left(\sum_{i=0}^{T_1-1}[h(X_i) - \pi(h)] \right)^2 \middle| X_0 = x \right] < \infty$.
Step 3: Prove that a $\sqrt{n}-$CLT holds for a stationary, one-step dependent stochastic processes.

Lemma 3.3. $I = E_\nu\left(\sum_{i=0}^{T_1-1}[h(X_i) - \pi(h)] \right) = 0.$

Proof. Denote $T_1 = \tau m$ and $H_k = \sum_{i=km}^{(k+1)m-1)}[h(X_i) - \pi(h)]$, then we have:

$$I = \mathbb{E}_\nu[\sum_{k=0}^{\infty} H_k \mathbb{I}(k < \tau)].$$

Consider the splitting $m-$skeleton chain $\{\check{X}_{nm}\}$ as in section 5.1.1 of Meyn and Tweedie [37](2003), we know that $\check{\alpha} = \mathcal{X}_1$ is an accessible atom. Then we can apply theorem 10.0.1 of Meyn and Tweedie [37] (2003) to this splitting chain. That is:

$$\begin{aligned} \pi(B) = \check{\pi}(B_0 \cup B_1) &= \int_{\check{\alpha}} \check{\pi}(dw) E_w[\sum_{k=1}^{\check{\tau}_{\check{\alpha}}} \mathbb{I}\{\check{X}_{km} \in \check{B}\}] \\ &= \epsilon \int_{\mathcal{X}_1} \pi(dw) E_w[\sum_{k=1}^{\check{\tau}_{\check{\alpha}}} \mathbb{I}\{\check{X}_{km} \in \check{B}\}]. \end{aligned}$$

We can define $\check{\tau}_{\check{\alpha}} = \min\{n \geq 1 : \check{X}_{nm} \in \check{\alpha}\}$. Since for any $w \in \check{\alpha}$, $\check{P}^m(w, \cdot) \sim \nu(\cdot)$, we have $\check{\tau}_{\check{\alpha}} = \tau$. Following the Theorem 5.1.3 in Meyn and Tweedie [37] (2003), we also have $P^{km}(x, B) = \check{P}^{km}(x, \check{B})$ for any $B \in \mathcal{B}(\mathcal{X})$. Therefore we have:

$$\pi(B) = \epsilon E_\nu[\sum_{k=1}^{\tau_1} \mathbb{I}\{X_{km} \in B\}] = \epsilon E_\nu[\sum_{k=1}^{\infty} \mathbb{I}\{X_{km} \in B\}\mathbb{I}\{\tau > k\}].$$

So we have:

$$\begin{aligned} I &= \mathbb{E}_\nu[E\Big(\sum_{k=0}^{\infty} H_k \mathbb{I}(k<\tau)|X_{km}\Big)] \\ &= \sum_{k=0}^{\infty} \mathbb{E}_\nu[E\Big(H_k \mathbb{I}(k<\tau)|X_{km}\Big)] \\ &= \sum_{k=0}^{\infty} \mathbb{E}_\nu[E\Big(H_k|X_{km}\Big)\mathbb{I}(k<\tau)]. \end{aligned}$$

The last equation comes from the fact that random variables $\mathbb{I}\{\tau > k\}$ and X_{km} are independent. And we know that given $\tau_1 > k$ and X_{km}, the distribution of H_k is equal to H_0 given X_0. Therefore we have:

$$\begin{aligned} I &= \sum_{k=0}^{\infty} \mathbb{E}_\nu[E\Big(H_0|X_0\Big)\mathbb{I}(k<\tau)] \\ &= E_\pi E\Big(H_0|X_0\Big) \\ &= E_\pi(H_0) \\ &= 0 \quad . \end{aligned}$$

□

Lemma 3.4. *We have:*

$$J = \mathbb{E}_\nu\Big[\Big(\sum_{i=0}^{T_1-1}[h(X_i)-\pi(h)]\Big)^2\Big] < \infty. \tag{3.3}$$

Proof.

$$\begin{aligned}
J &= \mathbb{E}_\nu\left[\left(\sum_{k=0}^{\tau-1}\sum_{i=km}^{(k+1)m-1)}[h(X_i)-\pi(h)]\right)^2\right] \\
&\le \mathbb{E}_\nu\left[\left(\sum_{k=0}^{\infty}\mathbb{I}\{k<\tau\}|H_k|\right)^2\right] \\
&= \mathbb{E}_\nu\left[\sum_{k=0}^{\infty}|H_k|^2\mathbb{I}\{k<\tau\}+2\sum_{k=0}^{\infty}\left(|H_k|\sum_{j=k+1}^{\infty}|H_j|\mathbb{I}\{j<\tau\}\right)\{k<\tau\}\right] \\
&= \mathbb{E}_\nu\left[\sum_{k=0}^{\infty}\left(|H_k|^2+2H_k\sum_{j=i+1}^{\infty}|H_j|\mathbb{I}\{j<\tau\}\right)\mathbb{I}\{k<\tau\}\right] \\
&= \mathbb{E}_\nu\left[\sum_{k=0}^{\infty}E\left(|H_k|^2+2|H_k|\sum_{j=k+1}^{\infty}|H_j|\mathbb{I}\{j<\tau\}]\mathbb{I}\{k<\tau\}|X_{km},\mathbb{I}\{k<\tau\}\right)\right] \\
&= \mathbb{E}_\nu\left[\sum_{k=0}^{\infty}E\left(|H_k|^2+2|H_k|\sum_{j=k+1}^{\infty}|H_j|\mathbb{I}\{j<\tau\}|X_{km}\right)\mathbb{I}\{k<\tau\}\right].
\end{aligned}$$

In the last equation, we have used the fact that random variables $\mathbb{I}\{\tau>k\}$ and X_{km} are independent. Since

$$\mathbb{E}\left(|H_i|^2+2|H_i|\sum_{j=1}^{\infty}|H_j|\{j<\tau\}|X_{im}=x\right)=\mathbb{E}\left(|H_0|^2+2|H_0|\sum_{j=1}^{\infty}|H_j|\{j<\tau\}|X_0=x\right),$$

if we denoet $f(x)=\mathbb{E}\left(|H_0|^2+2|H_0|\sum_{j=1}^{\infty}|H_j|\mathbb{I}\{j<\tau\}|X_0=x\right)$, then we have:

$$\begin{aligned}
J &\le \mathbb{E}_\nu\left[\sum_{k=0}^{\infty}f(X_0)\mathbb{I}\{k<\tau\}\right] \\
&= \mathbb{E}_\nu\left[f(X_0)\mathbb{I}\{0<\tau\}\right]+\mathbb{E}_\nu\left[\sum_{k=1}^{\infty}f(X_0)\mathbb{I}\{k<\tau\}\right] \\
&\le \mathbb{E}_\nu\left[f(X_0)\right]+\mathbb{E}_\nu\left[f(X_0)\right]\sum_{k=1}^{\infty}\mathbb{E}_\nu\left[\mathbb{I}\{k<\tau\}\right].
\end{aligned}$$

The last inequality is follows since:

1. $f(X_0)\mathbb{I}\{k<\tau\}\le f(X_0)$;

2. When $k\ge 1$, $\mathbb{I}\{\tau>k\}$ is independent with X_0.

Note

$$\mathbb{E}_\nu\left[\mathbb{I}\{k<\tau\right] = \mathbb{P}_\nu(k<\tau)\le(1-\epsilon)^k$$

and

$$\begin{aligned}\pi(dy) &= \int_E P^m(x,dy)\pi(dx) \\ &\geq \epsilon\nu(dy),\end{aligned}$$

therefore we have $J \leq \frac{1}{\epsilon}\mathbb{E}_\nu[f(X_0)] \leq \frac{1}{\epsilon^2}\mathbb{E}_\pi[f(X_0)$ and

$$\begin{aligned}\mathbb{E}_\pi[f(X_0)] &\leq \mathbb{E}_\pi[\sum_{i=0}^{m-1}|h(X_i)-\pi(h)|^2] \\ &\leq m(\pi(h^2)-\pi(h)^2) < \infty.\end{aligned}$$

From the above arguments we conclude that $J < \infty$. □

Finally, we prove the Theorem 3.1:

Proof of Theorem 3.1. Following the Lemma 3.1, we can obtain:

$$\lim_{n\to\infty}\frac{\sum_{i=0}^{n}[h(X_i)-\pi(h)]}{n^{1/2}} = \lim_{n\to\infty}\frac{\sum_{j=1}^{r(n)}\sum_{i=T_j}^{T_{j+1}-1}[h(X_i)-\pi(h)]}{n^{1/2}}. \tag{3.4}$$

Define $h_i = h(X_i) - \pi(h)$, $s_j = \sum_{i=T_j+1}^{T_{j+1}} h_i$ and $\eta_j = s_{jm+1} + \cdots + s_{(j+1)m-1}$ for an integer $m \geq 2$. Following the Lemma 3.2 we know that two collections of random variables: $\{s_i : 0 \leq j \leq m-2\}$ and $\{s_i : i \geq m\}$ are independent for any $m \geq 2$; thus

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}s_j = \frac{1}{\sqrt{n}}\sum_{j=0}^{[n/m]-1}\eta_j + \frac{1}{\sqrt{n}}\sum_{j=0}^{[n/m]-1}s_{mj} + \frac{1}{\sqrt{n}}\sum_{m[n/m]}^{n}s_j.$$

It should be noted that if $j-i > m$, then X_i and X_j are independent, η_j are i.i.d random variables and s_{mj} are i.i.d. so we have:

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{j=0}^{[n/m]-1}\eta_j &\to_d N(0,\frac{\sigma_m^2}{m}), \\ \frac{1}{\sqrt{n}}\sum_{j=0}^{[n/m]}s_{mj} &\to_d N(0,\frac{\sigma_s^2}{m}),\end{aligned}$$

where $\sigma_m^2 = (m-1)\mathbb{E}(s_1^2) + 2(m-2)\mathbb{E}(s_1s_2)$ and $\sigma_s^2 = \mathbb{E}[s_1^2]$, letting $m \to \infty$, we have $\frac{\sigma_m^2}{m} \to \mathbb{E}(s_1^2) + 2\mathbb{E}(s_1s_2)$ and $m^{-1}\sigma_s^2 \to 0$, so the CLT holds.

Let

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

then

$$\begin{aligned}\sigma^2 &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}[h(X_i)-\pi(h)]\right)^2\right] \\ &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[(\sum_{j=1}^{r(n)} s_j)^2\right] \\ &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[r(n)s_1^2 + 2(r(n)-2)s_1s_2\right].\end{aligned}$$

By the elementary renewal theorem (e.g. Feller [17](1968)), $\lim_{n\to\infty}\frac{r_n}{n} = \mathbb{E}(T_2 - T_1)$. Since $\mathbb{P}[T_2 - T_1 = n_0 s] = \varepsilon(1-\varepsilon)^{(s-1)}$, $\mathbb{E}(T_2 - T_1) = \sum_{s=1}^{\infty}[n_0 s\varepsilon(1-\varepsilon)^{(s-1)}] = \frac{n_0}{\varepsilon} < \infty$. Therefore if we denote $\widetilde{\sigma}^2 = \mathbb{E}[s_1^2 + 2s_1s_2]$, then

$$\sigma^2 = \frac{n_0}{\varepsilon}\mathbb{E}[s_1^2 + 2s_1s_2] = \frac{n_0}{\varepsilon}\widetilde{\sigma}^2. \tag{3.5}$$

As a result, we conclude that

$$\begin{aligned}\lim_{n\to\infty}\frac{\sum_{j=1}^{r(n)}\sum_{i=T_j}^{T_{j+1}-1}[h(X_i)-\pi(h)]}{n^{1/2}} &= \lim_{n\to\infty}\frac{\sum_{j=1}^{r(n)}\sum_{i=T_j}^{T_{j+1}-1}[h(X_i)-\pi(h)]}{r_n^{1/2}}\cdot\frac{r_n^{1/2}}{n^{1/2}} \\ &\longrightarrow_d \left(\frac{n_0}{\varepsilon}\right)^{1/2} N(0,\widetilde{\sigma}^2) \\ &= N(0,\sigma^2)\end{aligned}$$

as $n\to\infty$. □

Chapter 4

Adaptive MCMC Algorithm

4.1 Introduction

Markov chain Monte Carlo (MCMC) algorithms are widely used to generate samples from any probability distribution π on the state space $\mathcal{X}$. However it is generally acknowledged that the choice of an effective transition kernel is essential to obtain reasonable results by simulation in a limited amount of time. In practice, we can choose the transition probabiliy P from the family where $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is a collection of Markov chain kernels with stationary distribution $\pi(\cdot)$ on χ. Then the question is how to optimize the choice of the Markov chain's kernel. The initial idea is to choose a "best" P_γ, but it has been proved by Gilks et al[20](1998) that the optimal choice depends on the property of the target distribution π. So such "good" kernels are often very difficult to be well chosen (see also Gelman et al. [19]1996; Gilks et al [55] 1996 ; Haario et al [24] 1991; Roberts et al [40]1997). A possible solution so-called adaptive MCMC (AMCMC) has been proposed recently. The adaptive MCMC algorithm will tune the transition kernel at each step using the past simulations and try to "learn" the best parameter values while the chain runs. Adaptive MCMC methods using regeneration times and other complicate constructions have been propose by Gilks et al [20](1998), Brockwell and Kadane [10](2002). After

a significant step in this direction made by Haario et al. [26](1999), lots of adaptive algorithms were proposed, see [25](2001), [27] 2005, [23] 2006, Andrieu and Moulines [3](2005) , Andrieu and Robert [5](2001), Roberts and Rosenthal [48], [47](2005), Atchade and Rosenthal [6](2005), and Andieu and Achade [2](2007) for example.

4.2 Haario, Saksman and Tamminen's Adaptive MCMC Algorithm

A substantial amount of work has been done to validate adaptive Markov chain Monte Carlo algorithms in the seminal paper of Haario, Saksman and Tamminen [26]. We now explain how the algorithm works. Suppose, that at $n-$step we have sampled the states $X_0,\ X_1, \cdots, X_{n-1}$, where X_0 is the initial state. Then a candidate point Y is sampled from the (asymptotically symmetric) proposal distribution $q_n(\cdot|X_0,\ X_1, \cdots, X_{n-1})$, which now may depend on the whole history $(X_0,\ X_1, \cdots, X_{n-1})$. The candidate point Y is accepted with probability

$$\alpha(X_{n-1}, Y) = \min\{1, \frac{\pi(Y)}{\pi(X_{n-1})}\},$$

in which case we set $X_n = Y$, and otherwise $X_n = X_{n-1}$. Observe that the chosen probability for the acceptance resembles the familiar acceptance probability of the Metropolis algorithm. However, here the choice for the acceptance probability is not based on symmetry (reversibility) conditions since these cannot be satisfied in our case-the corresponding stochastic chain is no longer Markovian.

The proposal distribution $q_n(\cdot|X_0,\ X_1, \cdots, X_{n-1})$ here is the Gaussian distribution q_n with mean at the current point X_{n-1} and covariance $C_n = C_n(X_0, X_1, \cdots, X_{n-1})$.

The crucial thing regarding the adaption is how the covariance of the proposal distribution depends on the history of the chain. In the algorithm this is solved by setting $C_n = s_d cov(X_0, \cdots, X_{n-1}) + s_d \epsilon I_d$ after an initial period, where s_d is a parameter that de-

pends only on dimension d , $\epsilon > 0$ is a constant that we may choose very small compared to the size of S, I_d denotes the d-dimensional identity matrix and the initial covariance C_0 is an arbitrary strictly positive definite matrix according to our best prior knowledge. We select an index $n_0 > 0$ for the length of an initial period and define:

$$C_n = \begin{cases} C_0, & n \leq n_0; \\ s_d cov(X_0, \cdots, X_{n-1}) + s_d \epsilon I_d, & n > n_0. \end{cases}$$

The definition of the empirical covariance matrix determined by points $x_0, \cdots, x_x \in \mathbb{R}^d$:

$$cov(x_0, \cdots, x_k) = \frac{1}{k}(\sum_{i=0}^{k} x_i x_i^T - (k+1)\bar{x}_k \bar{x}_k^T).$$

where $\bar{x}_k = \frac{1}{k+1}\sum_{i=0}^{k} x_i$ and the elements $x_i \in \mathcal{R}^d$ are considered as column vectors. So one obtains that for $n \geq n_0 + 1$ the covariance C_n satisfies the recursion formula:

$$C_{n+1} = \frac{n-1}{n} C_n + \frac{s_d}{n}(n\bar{X}_{n-1}\bar{X}_{n-1}^T - (n+1)\bar{X}_n \bar{X}_n^T + X_n X_n^T).$$

This allows one to calculate C_n without too much computational cost since the mean $\bar{X}_n$ also satisfies an obvious recursion formula.

The choice for the length of the initial segment $n_0 > 0$ is free, but the bigger it is chosen the more slowly the effect of the adaption is felt. In a sense the size of n_0 reflects our trust in the initial covariance C_0. The role of the parameter ϵ is just to ensure that C_n will not become singular. As a basic choice for the scaling parameter we have adopted the value $s_d = \frac{(2.4)^2}{d}$ from Gelman et al.(1996), where it was shown that in a certain sense this choice optimizes the mixing properties of the Metropolis search in the case of Gaussian targets and Gaussian proposals, and further optimal results proved by [42] and [44]. We can observe that the algorithm continually adapt Σ using the empirical distribution of the available samples which makes the adaption tend to zero in some sense. Actually they provide a theoretical justification for adapting the covariance matrix Σ of the Gaussian proposal density used in a random walk Metropolis and proved the ergodicity of the above adaptive MCMC algorithm.

Theorem 4.1. *Let π be the density of a target distribution supported on a bounded measurable subset $\mathcal{X} \subset \mathbb{R}^d$, and assume that π is bounded. Let $\epsilon > 0$ and let ν_0 be any initial distribution on $\mathcal{X}$. Then the above adaptive MCMC simulates properly the target distribution π: for any bounded and measurable function $f : \mathcal{X} \to \mathbb{R}$, the equality*

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} f(X_i) = \int_{\mathcal{X}} f(x)\pi(dx)$$

holds almost surely.

These convergence results of adaptive algorithms have been made more general in [4], [3], [6], and [48]. An adaptive algorithm for the independent Metropolis sampler was proposed by [18] and [27] extended their previous work to Metropolis-within-Gibbs sampling. A class of quasi-perfect adaptive MCMC algorithms is introduced by [2]. Alternative approaches to adaptation within MCMC can be found in [10], [38], [21].

4.3 Ergodicity of General Adaptive MCMC (AMCMC) Algorithms

An important paper about the ergodicity of AMCMC was written by Roberts and Rosenthal [48] (2007). They present some simpler conditions, which still ensure the ergodicity of the specified target distribution. Before describing the procedure under study, it is necessary to introduce some notation and definitions.

4.3.1 General AMCMC

Here we will formalize the AMCMC as what Roberts and Roenthal [48](2007) did.
We let $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain kernels on $\mathcal{X}$, each of which is $\phi-$irreducible and aperiodic(which it usually will be) and has $\pi(\cdot)$ as a stationary distribution: $(\pi P_\gamma)(x,\cdot) = \pi(\cdot)$, and we call the set $\mathcal{Y}$ parameter space. Let Γ_n be $\mathcal{Y}-$valued random variables which

are updated according to specific rules. Consider a discrete time series $\{X_n\}$ on χ as below:

$$P[X_{n+1} \in A | X_n = x, \Gamma_n = \gamma, \mathcal{G}_n] = P_\gamma(x, A), \tag{4.1}$$

where $\mathcal{G}_n = \sigma(X_0, \cdots, X_n, \Gamma_0, \cdots, \Gamma_n)$. Then we call $\{X_n\}$ an adaptive MCMC with adaptive scheme Γ_n. Let

$$A^{(n)}((x, \gamma), B) = P[X_n \in B | X_0 = x, \Gamma_0 = \gamma], \quad B \in \mathcal{F};$$

and

$$T(x, \gamma, n) = \|A^{(n)}((x, \gamma), \cdot) - \pi(\cdot)\|.$$

We call an AMCMC algorithm an *independent adaptation* if for all n, Γ_n is independent of X_n. Obviously we have the following proposition:

Proposition 4.1. *Consider an independent adaptation algorithm $A^{(n)}((x, \gamma), \cdot)$, where $\pi(\cdot)$ is stationary for each $P_\gamma(x, \cdot)$. Then $\pi(\cdot)$ is also stationary for $A^{(n)}((x, \gamma), \cdot)$.*

When the AMCMC is to introduce some stopping time τ, such that no adaptations are done after time τ, i.e. such that $\Gamma_n = \Gamma_\tau$ whenever $n \geq \tau$. This scheme, which we refer to as finite adaptation, has been proposed by e.g. Pasarica and Gelman [39](2003). The finite sampling schemes always have the ergordic property:

Proposition 4.2. *Consider a finite AMCMC algorithm, in which each individual P_γ is ergodic for $\pi(\cdot)$. Then the finite AMCMC algorithm is also ergodic for $\pi(\cdot)$.*

4.3.2 The Ergodicity of AMCMC

In Roberts and Rosenthal [48](2007), they proved the following ergodic theorem in the uniformly convergence case:

Theorem 4.2. *Consider an adaptive MCMC algorithm on a state space χ, with adaptation index $\mathcal{Y}$ and the adaptive scheme is Γ_n. $\pi(\cdot)$ is stationary for each kernel P_γ for $\gamma \in \mathcal{Y}$. Suppose also that:*

Condition (a)[Simultaneous Uniform Ergodicity] *For all ϵ, there is $N = N(\epsilon) \in N$ such that $\|P_\gamma^N(x,\cdot) - \pi(\cdot)\| \leq \epsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$; and*

Condition (b)[Diminishing Adaption] *$lim_{n\to\infty} D_n = 0$ in probability, where $D_n = \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}} - P_{\Gamma_n}\|$ is a $\mathcal{G}_{n+1}$-measurable random variable.*

Then $lim_{n\to\infty} T(x,\gamma,n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.

They showed the Weak Law of Large Numbers (WLLN) under the same conditions.

Theorem 4.3. *Consider an adaptive MCMC algorithm. Suppose that conditions (a) and (b) hold. Let $g : \mathcal{X} \to R$ be a bounded measurable function. Then for any starting values $x \in \mathcal{X}$ and $\gamma \in \Gamma$, conditional on $X_0 = x$ and $\Gamma_0 = \gamma$ we have:*

$$\frac{\Sigma_{i=1}^n g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$.

Regarding the non-uniformly case, they also proved the ergodicity using the similar proof. Before we introduce the results, let us recall some definitions. According to the definition in Roberts, Rosenthal, and Schwartz [49] (1998), we say a family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ of Markov chain kernels is simultaneously strongly aperiodically geometrically ergodic (we denote it by **condition (c)**) if there is $C \in \mathcal{F}$, $V : \mathcal{X} \to [1,\infty)$, $\delta > 0$, $\lambda < 1$, and $b < \infty$, such that $\sup_C V = v < \infty$, and

(i) for each $\gamma \in \mathcal{Y}$, there exists a probability measure $\nu_\gamma(\cdot)$ on C with $P_\gamma(x,\cdot) \geq \delta\nu_\gamma(\cdot)$ for all $x \in C$; and

(ii) $(P_\gamma V)(x) \leq \lambda V(x) + b\mathbb{I}_C(x)$.

In Roberts and Rosenthal [48] (2007), they proved the following ergodic theorems:

Theorem 4.4. *Consider an adaptive MCMC algorithm on a state space χ, with adaptation index $\mathcal{Y}$ and the adaptive scheme is Γ_n. $\pi(\cdot)$ is stationary for each kernel P_γ for $\gamma \in \mathcal{Y}$. Suppose also that $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is simultaneously strongly aperiodically geometrically ergodic and the Adaptive scheme satisfies the following condition:*
[Diminishing Adaption] *$lim_{n\to\infty} D_n = 0$ in probability, where $D_n = \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}} - P_{\Gamma_n}\|$ is a $\mathcal{G}_{n+1}$-measurable random variable.*
Then $lim_{n\to\infty} T(x, \gamma, n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.

Furthermore, they also tried to relax the uniform convergence condition (a) of Theorem 4.2. Actually the proof of the Theorem 4.2 shows that condition (a) was used only to ensure $P^N_{\Gamma_{K-N}}(X_{K-N}, \cdot)$ was close to $\pi(\cdot)$. Therefore for any $\epsilon > 0$, define "ϵ convergence time function" $M_\epsilon : \mathcal{X} \times \mathcal{Y} \to N$ such that

$$M_\epsilon(x, \gamma) = \inf\{n \geq 1 : \|P^n_\gamma(x, \cdot) - \pi(\cdot)\| \leq \epsilon\}.$$

Obviously if each individual P_γ is ergodic, then $M_\epsilon(x, \gamma) < \infty$. We denote that for all $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^\infty$ is bounded in probability given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$ by **condition (d)**. That is:
Condition (d): for all $\delta > 0$, there is $N \in \mathbb{N}$ such that $P[M_\epsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \leq 1 - \delta$ for all $n \in \mathbb{N}$.

Theorem 4.5. *Consider an adaptive MCMC algorithm with Diminishing Adaption (i.e., $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| = 0$ in probability). Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{Y}$. Then $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$ provided condition (d) holds.*

Chapter 5

Recurrent And Ergodic Properties of AMCMC

5.1 Introduction

In Roberts and Rosenthal [48] 2007, they not only present some ergodicity results under more general conditions but also mentioned some research directions. We will continue to study the ergodicity of AMCMC along these directions, try to find some weaker conditions to ensure the ergodicity and discuss the relationship between the recurrence on the product space (of the state space and the parameter space) and the ergodicity.

The chapter is organized as follows. Section 5.2 we will present our main results: the ergodic theorem of AMCMC under the weakest drift conditions such that each kernel is positive recurrence. Further we will discuss the uniformly recurrent conditions in the same section after constructing some simple examples to show that usually AMCMC does not have good recurrence property. In section 5.3 we will give the proof of the ergodic theorem. In section 5.4, we consider the recurrent property on the product space of the state space and the parameter one. We will give the negative answer to the Open Problem 21 in Roberts and Rosenthal [48](2005) using a counter example, and present

some positive results under stronger conditions. Finally we will construct two examples to discuss the convergence rate of AMCMC.

5.2 The Ergodicity Under Minimal Uniformly Recurrent Conditions

Consider Theorem 4.2, Roberts and Rosenthal proved the ergodicity with simultaneously geometrically ergodic condition. However we note that Theorem 2.5 part (iv) indicates that to merely prove convergence (as opposed to geometric convergence), it suffices to have an even weaker drift condition of the form

$$PV(x) \leq V(x) - 1 + b\mathbb{I}_C.$$

So perhaps it suffices for the validity of adaptive MCMC algorithms that such drift conditions hold uniformly for all P_γ. Unfortunately, the available results appears not to provide any explicit quantitative bounds on convergence. However if the parameter space is compact in some sense, we can prove the ergodicity with the minimal uniformly recurrent conditions.

First let us think about how to measure the difference between two elements γ_1 and γ_2 in the parameter space $\mathcal{Y}$. Actually what we need to describe is the difference between the respective kernels P_{γ_1} and P_{γ_2}, i.e. $\sup_{x\in\mathcal{X}} \|P_{\gamma_1}(x,\cdot) - P_{\gamma_2}(x,\cdot)\|$. Therefore we will define the metric $d(\gamma_1,\gamma_2)$ on $\mathcal{Y} \subset R^q$ as:

$$d(\gamma_1,\gamma_2) = \sup_{x\in\mathcal{X}} \|P_{\gamma_1}(x,\cdot) - P_{\gamma_2}(x,\cdot)\|.$$

We suppose there exists a transition kernel P_γ corresponding to each $\gamma \in R^q$, and consider the following set:

$$\Delta = \{\gamma \in R^q |\ P_\gamma V \leq V - 1 + b1_C\}.$$

Now we can state our main result.

Theorem 5.1. ***(Ergodicity Theorem)*** *Consider an adaptive MCMC algorithm with Diminishing Adaption, such that there is* $C \in \mathcal{F}$, $V : \mathcal{X} \to [1, \infty)$ *such that* $\pi(V) < \infty$, $\delta > 0$, *and* $b < \infty$, *with* $\sup_C V = \nu < \infty$, *and:*

Condition (e):

(i) for each $\gamma \in \mathcal{Y}$, *there exists a probability measure* $\nu_\gamma(\cdot)$ *on* C *with* $P_\gamma(x, \cdot) \geq \delta\nu_\gamma(\cdot)$ *for all* $x \in C$; *and*

(ii)$P_\gamma V \leq V - 1 + b\mathbb{I}_C$ *for each* γ;

Condition (f):

(iii) the set Δ *is compact w.r.t the metric* d.

Suppose further that the sequence $\{V(X_n)\}_{n=0}^{\infty}$ *is bounded in probability, given* $X_0 = x_*$ *and* $\Gamma_0 = \gamma_*$. *Then* $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$.

5.2.1 The Uniform Minimal Drift Condition

Intuitively, we hope the AMCMC is recurrent whenever each kernel is positive recurrent with respect to the target distribution π. However following the example below, we get the negative conclusion. Consider the following adaptive MCMC: suppose the state space $\mathcal{X} = \{1, 2\}$, the parameter space $\mathcal{Y} = \mathbb{N} \times \{1, 2\}$ with each kernel $P_{n,1} = \begin{pmatrix} 1 - \frac{1}{2^n} & \frac{1}{2^n} \\ \frac{1}{2^n} & 1 - \frac{1}{2^n} \end{pmatrix}$ and $P_{n,2} = \begin{pmatrix} \frac{1}{2^n} & 1 - \frac{1}{2^n} \\ 1 - \frac{1}{2^n} & \frac{1}{2^n} \end{pmatrix}$, and the stationary distribution $\pi(1) = \pi(2) = \frac{1}{2}$. We design an adaptive algorithm as:

$$\Gamma_n = \begin{cases} (n, 1) & \text{, if } X_n = 1; \\ (n, 2) & \text{, if } X_n = 2. \end{cases}$$

Lemma 5.1. *The above adaptive MCMC is NOT recurrent, although each kernel is positive recurrent with respect to the distribution* $\pi(\cdot)$. *Actually we have* $\mathbb{E}_2[\eta_2] < \infty$, *which means that the chain will NOT come back to* $\{2\}$ *after a long run when it starts from* $\{2\}$. *Therefore* $\lim_{n\to\infty} P(X_n = 2 | X_0 = i) = 0$ *for* $i = 1,\ 2$, *which is not equal to* $\pi(2)$.

Proof. Suppose $\eta_2 = \sum_{n=1}^{\infty} \mathbb{I}\{X_n = 2\}$. Then according to the adaptive algorithm, we have:

$$\begin{aligned} P_2(\eta_2 = n) &= \sum_{1\le i_1<i_2\cdots<i_n<\infty} \frac{\Pi_{i=1}^{\infty}(1-\frac{1}{2^i})}{\Pi_{j=1}^{n}(1-\frac{1}{2^{i_j}})}\Pi_{j=1}^{n}\frac{1}{2^{i_j}} \\ &\le \sum_{1\le i_1<i_2\cdots<i_n<\infty} \Pi_{j=1}^{n}\frac{1}{2^{i_j}} \\ &= \sum_{1\le i_1<i_2\cdots<i_n<\infty} \frac{1}{2^{\sum_{j=1}^{n} i_j}} \\ &\le \sum_{m=\frac{n(n+1)}{2}}^{\infty} C_m^n \frac{1}{2^m} \\ &= \frac{1}{n!}\sum_{m=\frac{n(n+1)}{2}}^{\infty} m(m-1)\cdots(m-n+1)\frac{1}{2^m}. \end{aligned}$$

Consider the functional series $S_n(x) = \sum_{m=\frac{n(n+1)}{2}}^{\infty} m(m-1)\cdots(m-n+1)x^m$ for $0 < x < 1$, then we have:

$$\begin{aligned} S_n(x) &= x^n\big[\sum_{m=\frac{n(n+1)}{2}}^{\infty} x^m\big]^{(n)} \\ &= x^n\big[\frac{x^{\frac{n(n+1)}{2}}}{1-x}\big]^{(n)} \\ &= x^n\sum_{i=0}^{n} C_n^i \frac{(\frac{n(n+1)}{2})!}{(\frac{n(n+1)}{2}-i)!}x^{\frac{n(n+1)}{2}-i} i!(1-x)^{-i} \\ &\le x^{\frac{n(n+1)}{2}} \times \sum_{i=0}^{n} C_n^i x^{n-i}(x-1)^{-i}\frac{(\frac{n(n+1)}{2})!}{(\frac{n(n+1)}{2}-n)!}n! \\ &\le x^{\frac{n(n+1)}{2}} \times (x+\frac{1}{1-x})^n(\frac{n(n+1)}{2})^n(n!)\ . \end{aligned}$$

Therefore we have:

$$\begin{aligned} P_2(\eta_2 = n) &\le (\frac{1}{2})^{\frac{n(n+1)}{2}} \times (\frac{5}{2})^n \times (\frac{n(n+1)}{2})^n \\ &= \Big[(\frac{1}{2})^{\frac{(n+1)}{2}} \times (\frac{5}{2}) \times (\frac{n(n+1)}{2})\Big]^n. \end{aligned}$$

We know that $\lim_{n\to\infty}(\frac{1}{2})^{\frac{(n+1)}{2}} \times (\frac{5}{2}) \times (\frac{n(n+1)}{2}) = 0$, i.e. there exists $N > 0$ such that for

any $n > N$ we have $(\frac{1}{2})^{\frac{(n+1)}{2}} \times (\frac{5}{2}) \times (\frac{n(n+1)}{2}) < \frac{1}{2}$. So

$$\begin{aligned} E_2[\eta_2] &= \sum_{n=1}^{\infty} P_2(\eta_2 = n)n \\ &< \sum_{i=1}^{N} i + \sum_{i=N+1}^{\infty} i \times [\frac{1}{2}]^i \\ &< \infty. \end{aligned}$$

Therefore the set $\{2\}$ is a transient set. Furthermore following that $\sum_{n=1}^{\infty} P_2(\eta_2 = n)n < \infty$, we know that $\lim_{n\to\infty} P(\eta_2 = n) = 0$, which is NOT equal to $\pi(2)$. □

In the above example, we can ascribe the transience of the AMCMC to increasing of probability to $\{2\}$ as $n \to \infty$. Therefore we need the "uniform" recurrence property with respect to the parameter γ. Following the theorem 11.0.1 in Meyn and Tweedie [37], we know that an irreducible Markov chain is positive recurrent if and only if there exists some petite set C and some extend valued, non-negative test function V, which is finite for at least one state in the state space $\mathcal{X}$, satisfying:

$$PV(x) \leq V(x) - 1 + b\mathbb{I}_C(x), \quad x \in \mathcal{X}.$$

Therefore we will suppose all the $\gamma \in \mathcal{Y}$ satisfy:

$$P_\gamma V(x) \leq V(x) - 1 + b\mathbb{I}_C(x), \quad x \in \mathcal{X}.$$

5.3 The Proof of Ergodicity Theorem

Before we prove the theorem 5.1, let us think about the following lemma:

Lemma 5.2. *Consider an adaptive MCMC algorithm with Diminishing Adaptation, with a regular stationary measure π and an accessible atom $\alpha \in \mathcal{F}$ such that $P_\gamma(x, B) = \nu_\gamma(B)$ for any $x \in \alpha$ and $B \in \mathcal{B}(\mathcal{X})$, where $\nu_\gamma(\cdot)$ is a regular probability measure, let measurable function $W : \mathcal{X} \to [0, \infty)$, $0 < K < \infty$,*
(i) $E_{\alpha,\gamma}[\tau_\alpha] \leq K$ and $E_{x,\gamma}[\tau_\alpha] \leq W(x)$ for any $x \in \alpha^c$ and $\gamma \in \mathcal{Y}$.

(ii) The parameter space $\mathcal{Y}$ is a closed complete subset w.r.t the metric d of the set Δ. Suppose further that the sequence $\{W(X_n)\}_{n=0}^{\infty}$ is bounded in probability, given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$. Then we have:

$$\lim_{n\to\infty} T(x_*, y_*, n) = 0.$$

5.3.1 The Proof Of Theorem 5.1

Suppose we have the lemma 5.2 hold. Let us recall what the splitting chain is. Actually outside C the chain $\{\check{X}_n^\gamma\}$ behaves just like $\{X_n^\gamma\}$, moving on the "top" half $\mathcal{X}_0$ of the split space. Each time it arrives in C, it is "split"; with probability $1-\delta$ it remain in C_0, with probability δ it drops to C_1, and C_1 is the atom of the splitting chain, set $C_1 = \alpha$. We can prove Theorem 5.1 as below.

Proof. Consider the splitting chain $\{\check{X}_n^\gamma\}$, we know that the subset $\alpha = C_1 \in \check{\mathcal{X}}$ is an accessible atom of any chain $\{X_n^\gamma\}$.

Step 1: Prove that there exists $K > 0$ such that

$$E_{\alpha,\gamma}(\tau_\alpha) \le K;$$

Step 2: Prove that there exists a measurable function $W : \check{\mathcal{X}} \to [0, \infty)$ such that:

$$E_{x,\gamma}(\tau_\alpha) \le W(x);$$

Step 3: Check the regularity of ν_γ and π.

Suppose $\check{\tau}_{A,\gamma}^{(m)}(B)$ is the $m-$th hitting time of B from A and with the kernel $\check{P}_\gamma$. Consider the random variable $\check{\tau}_{\alpha,\gamma}(\alpha)$, then $\check{\tau}_{\alpha,\gamma}(\alpha) = \check{\tau}_{\alpha,\gamma}(\check{C}) + \check{\tau}_{\check{C},\gamma}^{(k-1)}(\check{C})$with probability $(1-\delta)^{k-1}\delta$. If we denote the random variable T = the number of $\{n \le \check{\tau}_{\alpha,\gamma}(\alpha) | \check{X}_n \in \check{C}\}$, where $\check{C} = C_0 \cup C_1$, we have:

$$\begin{aligned} E_{\alpha,\gamma}(\tau_\alpha) &= E[E(\check{\tau}_{\alpha,\gamma}(\alpha)|T)] \\ &= \sum_{k=1}^{\infty} \left(E(\check{\tau}_{\alpha,\gamma}(\check{C})) + (k-1)E_{\check{C},\gamma}(\tau_{\check{C}}) \right)(1-\delta)^{k-1}\delta \\ &= E(\check{\tau}_{\alpha,\gamma}(\check{C})) + \frac{1-\delta}{\delta} E_{\check{C},\gamma}(\tau_{\check{C}}) \end{aligned}$$

and we also know that for any $x \in \check{C}$,$\gamma \in \mathcal{Y}$ $E[\check{\tau}_{x,\gamma}(\check{C})] = E_{x,\gamma}(\tau_C) \leq V(x)+b \leq v+b = K$. Therefore $E_{\alpha,\gamma}(\tau_\alpha) \leq K + \frac{1-\delta}{\delta}K = \frac{K}{\delta}$.

Similarly for any $x \notin \alpha$, we know that $\check{\tau}_{x,\gamma}(\alpha) = \check{\tau}_x(\check{C}) + \check{\tau}^{(k-1)}_{\check{C},\gamma}(\check{C})$with probability $(1-\delta)^{k-1}\delta$. Therefore we have:

$$\begin{aligned} E_{x,\gamma}(\tau_\alpha) &= E[E(\check{\tau}_{x,\gamma}(\alpha)|T)] \\ &= \sum_{k=1}^{\infty}\Big(E(\check{\tau}_{x,\gamma}(\check{C})) + (k-1)E_{\check{C},\gamma}(\tau_{\check{C}})\Big)(1-\delta)^{k-1}\delta \\ &= E(\check{\tau}_{x,\gamma}(\check{C})) + \frac{1-\delta}{\delta}E_{\check{C},\gamma}(\tau_{\check{C}}) \end{aligned}$$

and we also have for any x,$\gamma \in \mathcal{Y}$, $E[\check{\tau}_{x,\gamma}(\check{C})] = E_{x,\gamma}(\tau_C) \leq V(x) + b = W(x)$. Since $V(X_n)$ is bounded in probability, $W(X_n)$ is also bounded in probability.

Finally since $\int_{\mathcal{X}} V(y)\nu_\gamma(dy) < v$ and $\pi(V) < \infty$,the probability measures ν_γ and π are both regular. Then we can prove the theorem 5.1 following the lemma 5.2. □

5.3.2 The Proof Of Lemma 5.2

Following the last section, it suffices to prove the lemma 5.2. For any initial value $x \in \mathcal{X}$ and measurable function $|f| \leq 1$, denote: $a_{x,\gamma}(n) = P_{x,\gamma}(\tau_\alpha = n)$, that is the first hitting time of α is n when the kernel is P_γ and the start value is x; similarly denote $u_\gamma(n) = (P_\gamma)_\alpha(\Phi_n \in \alpha)$ and define:

$$t_{f,\gamma}(n) = \int_\alpha P^n_\gamma(\alpha, dy)f(y) = (E_\gamma)_\alpha[f(\Phi_n)1\{\tau_{\alpha\geq n}\}].$$

Then following the first-entrance last-exit decomposition we have:

$$P^n_\gamma(x,B) =_\alpha P^n_\gamma(x,B) + \sum_{j=1}^{n-1}[\sum_{k=1}^{j} {}_\alpha P^k_\gamma(x,\alpha)P^{j-k}(\alpha,\alpha)]_\alpha P^{n-j}_\gamma(\alpha,B),$$

where ${}_\alpha P^{n-j}_\gamma(\alpha,B)$ is the taboo probability given by

$${}_\alpha P^{n-j}_\gamma(\alpha,B) = P_\gamma(X_{n-i} \in B, \tau_\alpha \geq n-j | X_0 \in \alpha)$$

Therefore for any $x \in \mathcal{X}$ and f, we have:

$$\int P_\gamma^n(x, d\omega) f(\omega) = \int {}_\alpha P_\gamma^n(x, d\omega) f(\omega) + a_{x,\gamma} * u_\gamma * t_{f,\gamma}(n),$$

then we will get:

$$\begin{aligned}
|E_{x,\gamma}[f(\Phi_n)] - E_\pi[f(\Phi_n)]| &\leq E_{x,\gamma}[f(\Phi_n)\mathbb{I}\{\tau_{\alpha\geq n}\}] \\
&+ |a_{x,\gamma} * u_\gamma - \pi(\alpha)| * t_{f,\gamma}(n) \\
&+ \pi(\alpha) \sum_{j=n+1}^{\infty} t_{f,\gamma}(j) \\
&\leq E_{x,\gamma}[f(\Phi_n)\mathbb{I}\{\tau_{\alpha\geq n}\}] + \sum_{j=1}^{n} |\sum_{i=1}^{j} a_x(j) u(j-i) - \pi(\alpha)| t_1(n-j) \\
&+ \pi(\alpha) \sum_{j=n+1}^{\infty} t_{f,\gamma}(j) \\
&\leq E_{x,\gamma}[f(\Phi_n)\mathbb{I}\{\tau_{\alpha\geq n}\}] + \sum_{j=1}^{n} \sum_{i=1}^{j} a_x|(j) u(j-i) - \pi(\alpha)| t_1(n-j) \\
&+ \pi(\alpha) \sum_{j=n+1}^{\infty} t_{f,\gamma}(j) \\
&\leq E_{x,\gamma}[f(\Phi_n)\mathbb{I}\{\tau_{\alpha\geq n}\}] + \sum_{j=1}^{n} \sum_{i=1}^{j} a_x(i) |u(j-i) - \pi(\alpha)| t_1(n-j) \\
&+ \pi(\alpha) \sum_{j=1}^{n} \sum_{i=j+1}^{\infty} a_x(i) t_1(n-j) + \pi(\alpha) \sum_{j=n+1}^{\infty} t_{1,\gamma}(j).
\end{aligned}$$

Now we can denote the first term as I, the second as II, the third as III and the fourth term as IV. And we have the following estimations.

The Estimation Of I and III

Lemma 5.3. $I \leq \frac{W(x)}{n}$.

Proof.

$$\begin{aligned} I &\leq E_{x,\gamma}[1_{\tau_\alpha \geq n}] \\ &= P_{x,\gamma}(\tau_\alpha \geq n) \\ &\leq \frac{E_{x,\gamma}(\tau_\alpha)}{n} \\ &\leq \frac{W(x)}{n}. \end{aligned}$$

□

Lemma 5.4. *Let* $a_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{i}$, *then* $III \leq 2a_nKW(x)$ *for any* $x \in \mathcal{X}$.

Proof.

$$\begin{aligned} III &\leq \sum_{j=1}^{n} P_x(\tau_\alpha \geq j)P_\alpha(\tau_\alpha \geq n-j) \\ &\leq \sum_{j=1}^{n} \frac{W(x)}{j} \times \frac{K}{n-j} \\ &= KW(x)\frac{2}{n}\sum_{i=1}^{n}\frac{1}{i} \\ &= 2Ka_nW(x). \end{aligned}$$

And we know that $\lim_{n\to\infty} a_n = 0$. □

The Estimation Of Term IV

Following the structure of stationary distribution π, we know that

$$\sum_{j=1}^{\infty} P_{\alpha,\gamma}(\tau_\alpha > j) = \frac{1}{\pi(\alpha)} = M,$$

so for any $\epsilon > o$, there exists N_γ, such that for any $n_\gamma > N_\gamma$:

$$\sum_{j=1}^{n_\gamma} P_{\alpha,\gamma}(\tau_\alpha) > M - \epsilon$$

We define $n_\epsilon(\gamma) = \inf\{n : \sum_{j=1}^{n} P_{\alpha,\gamma}(\tau_\alpha > j) > M - \epsilon\}$, and prove that:

Lemma 5.5. *For any fixed γ_0, there exists $\delta > 0$ such that for any $d(\gamma, \gamma_0) < \delta$, we have $n_\epsilon(\gamma) = n_\epsilon(\gamma_0)$.*

Proof. Denote $\eta_1 = \sum_{j=1}^{n_{\gamma_0}} P_{\alpha,\gamma_0}(\tau_\alpha > j) - (M - \epsilon)$ and $\eta_2 = M - \epsilon - \sum_{j=1}^{n_{\gamma_0}-1} P_{\alpha,\gamma_0}(\tau_\alpha > j)$. Set $\delta = \frac{2\min\{\eta_1,\eta_2\}}{n_\epsilon(\gamma_0)(n_\epsilon(\gamma_0)+1)}$, then consider two Markov chain $\{X_i\}$ with kernel P_{γ_0} and $\{X_i^{'}\}$ with kernel P_{γ_1} such that $d(\gamma_0, \gamma_1) < \delta$. Then

$$\begin{aligned}
\mathbb{P}_x(X_i \neq X_i^{'} | X_{i-1} = X_{i-1}^{'}) &= \mathbb{E}(P_x(X_i \neq X_i^{'} | X_{i-1} = X_{i-1}^{'}, X_{i-1} = y)) \\
&\leq \mathbb{E}(P(X_i \neq X_i^{'} | X_{i-1} = X_{i-1}^{'} = y)) \\
&= \mathbb{E}(\|P_{\gamma_0}(y, \cdot) - P_{\gamma_1}(y, \cdot)\|) \\
&\leq \mathbb{E}(d(\gamma_0, \gamma_1)) \\
&< \delta.
\end{aligned}$$

The third equation $P(X_i \neq X_i^{'} | X_{i-1} = X_{i-1}^{'} = y) = \|P_{\gamma_0}(y, \cdot) - P_{\gamma_1}(y, \cdot)\|$ is following the Proposition 3(g) in [46]. Then we have

$$P_x(X_i \neq X_i^{'}, X_{i-1} = X_{i-1}^{'}) = P_x(X_i \neq X_i^{'} | X_{i-1} = X_{i-1}^{'}) P_x(X_{i-1} = X_{i-1}^{'}) \leq \delta$$

With the same start value $x \in \alpha$, then we have:

$$\begin{aligned}
\mathbb{P}(X_i \neq X_i^{'} | X_0 = X_0^{'} = x) &= P_x(X_i \neq X_i^{'}, X_{i-1} \neq X_{i-1}^{'}) + P_x(X_i \neq X_i^{'}, X_{i-1} = X_{i-1}^{'}) \\
&\leq P_x(X_{i-1} \neq X_{i-1}^{'}) + \delta \\
&\leq P_x(X_{i-1} \neq X_{i-1}^{'}, X_{i-2} \neq X_{i-2}^{'}) + P_x(X_{i-1} \neq X_{i-1}^{'}, X_{i-2} = X_{i-2}^{'}) + \delta \\
&\leq P_x(X_{i-2} \neq X_{i-2}^{'}) + 2\delta \\
&\leq \cdots \\
&\leq i\delta.
\end{aligned}$$

Therefore:

$$\sum_{i=1}^{n_{\gamma_0}(\epsilon)} P_x(X_i \neq X_i^{'}) \leq \sum_{i=1}^{n_{\gamma_0}(\epsilon)} i\delta \leq \min\{\eta_1, \eta_2\}.$$

So we still have $n_\epsilon(\gamma) = n_\epsilon(\gamma_0)$. □

Lemma 5.6. *For any $\epsilon > 0$, there exits $N > 0$ which is independent with γ, such that for any $n > N$, we have: $\sum_{j=n+1}^{\infty} P_{\alpha,\gamma}(\tau_\alpha > j) < \epsilon$.*

Proof. Suppose there exists $\epsilon > 0$ and a sequence $\{\gamma_k\}$ such that $n_\epsilon(\gamma_k) \to \infty$. Following the compactness of the parameter space $\mathcal{Y}$, there exists $\{\gamma_{k_i}\} \to \gamma_0$, i.e. $|\gamma_{k_i} - \gamma_0| \to 0$, and $\gamma_0 \in \Delta$. Now let $k_i \to \infty$, we will get $\sum_{i=1}^{\infty} P_{\alpha,\gamma_0}(\tau_\alpha > j) \leq M - \epsilon$ which is conflicting with that: for any $\gamma \in \Delta$, we have $\sum_{i=1}^{\infty} P_{\alpha,\gamma_0}(\tau_\alpha > j) = M$. So for any $\epsilon > 0$, there exits $N > 0$ which is independent with γ, such that for any $n > N$, we have: $\sum_{j=n+1}^{\infty} P_{\alpha,\gamma}(\tau_\alpha > j) < \epsilon$ □

Lemma 5.7. *For any $\epsilon > 0$, there exits $N > 0$ which is independent with γ, such that for any $n > N$, we have: $IV < \epsilon$.*

Proof. Since

$$\begin{aligned} IV &\leq \pi(\alpha) \sum_{j=n+1}^{\infty} t_{1,\gamma}(j) \\ &= \pi(\alpha) \sum_{j=n+1}^{\infty} E_{\alpha,\gamma}[1_{\tau_\alpha \geq j}] \\ &= \pi(\alpha) \sum_{j=n+1}^{\infty} P_{\alpha,\gamma}(\tau_\alpha > j), \end{aligned}$$

following lemma 5.6, we know that for any $\epsilon > 0$, there exits $N > 0$ which is independent with γ, such that for any $n > N$, we have: $\sum_{j=n+1}^{\infty} P_{\alpha,\gamma}(\tau_\alpha > j) < \frac{\epsilon}{\pi(\alpha)}$. That is $IV \leq \epsilon$ for any $n > N$. □

The Estimation On Term II

Lemma 5.8. *For any $\epsilon > 0$, there exists $N > 0$ which is independent with γ such that $II \leq \epsilon W(x)$.*

$$\begin{aligned} II &\leq \sum_{j=1}^{n} t_{1,\gamma}(n-j) \sum_{i=1}^{j} a_{x,\gamma}(i) i \frac{|u_\gamma(j-i)-\pi(\alpha)|}{i} \\ &\leq \sum_{j=1}^{n} t_{1,\gamma}(n-j) [\sum_{i=1}^{\infty} a_{x,\gamma}(i) i] \sum_{i=1}^{j} \frac{|u(j-i)-\pi(\alpha)|}{i} \\ &\leq \sum_{j=1}^{n} t_{1,\gamma}(n-j) E_{x,\gamma}(\tau_\alpha) \sum_{i=1}^{j} \frac{|u(j-i)-\pi(\alpha)|}{i} \\ &\leq W(x) \sum_{j=1}^{n} t_{1,\gamma}(n-j) \sum_{i=1}^{j} \frac{|u_\gamma(j-i)-\pi(\alpha)|}{i}. \end{aligned}$$

Lemma 5.9. $\sum_{i=1}^{\infty} |u_\gamma(i) - \pi(\alpha)| < \infty$ *for each* γ.

Proof. Since $\sup_{\check{C}} V(x) = v$ and ν_γ is probability measure on α, $\int_{\mathcal{X}} V(x)\nu_\gamma(dx) < \infty$ and $\pi(V) < \infty$, following Theorem 11.3.12 of Meyn and Tweedie [37], we know that ν_γ and $\pi(\cdot)$ are both regular measure. Then following Theorem 13.4.5 in Meyn and Tweedie's book, we know that:

$$\sum_{n=1}^{\infty} \|\nu_\gamma P_\gamma^n - \pi\| < \infty.$$

Therefore we have $\sum_{n=1}^{\infty} \|P_\gamma^n(\alpha,\alpha) - \pi(\alpha)\| < \infty$. □

Lemma 5.10. $\lim_{n\to\infty} \sum_{j=1}^{n} t_{1,\gamma}(n-j) \sum_{i=1}^{j} \frac{|u_\gamma(j-i)-\pi(\alpha)|}{i} = 0$ *for any* $\gamma \in \mathcal{Y}$.

Proof. Let $s_j(\gamma) = \sum_{i=1}^{j} \frac{|u_\gamma(j-i)-\pi(\alpha)|}{i}$, following bounded convergence theorem and lemma 5.9, we have $s_j(\gamma) \to_{j\to\infty} 0$. Similarly following $\sum_{j=1}^{\infty} t_{1,\gamma}(j) = E_{\gamma,\alpha}(\tau_\alpha) \leq v < \infty$, we have $\lim_{n\to\infty} \sum_{j=1}^{n} t_{1,\gamma}(n-j) \sum_{i=1}^{j} \frac{|u_\gamma(j-i)-\pi(\alpha)|}{i} = 0$. □

Lemma 5.11. *For any* $\epsilon > 0$ *there exists* N *which is independent with* γ*, such that for any* $n > N$*, we have* $\sum_{j=1}^{n} t_{1,\gamma}(n-j) \sum_{i=1}^{j} \frac{|u_\gamma(j-i)-\pi(\alpha)|}{i} < \epsilon$.

Proof. Suppose there exist $\epsilon > 0$, and strictly increasing $\{n_i\}_{i=1}^{\infty}$ and $\gamma_{n_i} \in \mathcal{Y}$ such that $\sum_{j=1}^{n_i} t_{1,\gamma_{n_i}}(n-j) \sum_{i=1}^{j} \frac{|u_{\gamma_{n_i}}(j-i)-\pi(\alpha)|}{i} > \epsilon$. Then there exists γ_0 such that $\gamma_{n_i} \to \gamma_0$. Therefore we have:

$$\sum_{j=1}^{\infty} t_{1,\gamma_0}(n-j) \sum_{i=1}^{j} \frac{|u_{\gamma_0}(j-i)-\pi(\alpha)|}{i} > \epsilon.$$

Contradiction. So $\lim_{n\to\infty}\sum_{j=1}^{n} t_{1,\gamma}(n-j)\sum_{i=1}^{j}\frac{|u_\gamma(j-i)-\pi(\alpha)|}{i}=0$. □

From all above estimations of I, II, III and IV, we have the following lemma:

Lemma 5.12. *For any $\epsilon>0$, there exists $N>0$ which is independent with the choice of γ, such that for any $n>N$, we have:*

$$\|P_\gamma^n(x,\cdot)-\pi(\cdot)\| \leq \frac{W(x)}{n}+\epsilon W(x)+\epsilon.$$

The Proof Of Lemma 5.2

Proof. Let $M_\epsilon(x,\gamma)=\inf\{n\geq 1:\|P_\gamma^n(x,\cdot)-\pi(\cdot)\|\leq\epsilon\}$. Then following the theorem 13 in Roberts and Rosenthal [48] (2007), it suffices to prove that $\{M_\epsilon(X_n,\Gamma_n)\}_{n=0}^{\infty}$ is bounded in probability given $X_0=x_*$ and $\Gamma_0=\gamma_*$, i.e. for all $\delta>0$, there is $N\in\mathbb{N}$ such that:

$$P[M_\epsilon(X_n,\Gamma_n)\leq N|X_0=x_*,\Gamma_0=\gamma_*]\geq 1-\delta.$$

Since for any $\epsilon>0$, there exists $N>0$ which is independent with the choice of γ, such that for any $n>N$, we have:

$$\|P_\gamma^n(x,\cdot)-\pi(\cdot)\| \leq \epsilon W(x)+\epsilon,$$

and $W(X_n)$ is bounded in probability, we have the conclusion hold. □

5.4 Recurrence On The Product Space $\mathcal{X}\times\mathcal{Y}$

The adaptive MCMC induces sample paths on the product space $\mathcal{X}\times\mathcal{Y}$. We will study the recurrent property on the product space in this section. When each kernel P_γ has good ergodic property and the random variable sequence (X_n,Γ_n) is also recurrent on the $\mathcal{X}\times\mathcal{Y}$, we hope to get the ergodicity of AMCMC. But following the computation in section 5.4.1, we get the negative answer. Fortunately Roberts and Rosenthal's paper [14]

(2007) offered us a proper condition–"Diminishing Adaptation conditions" and showed some positive results, however they mentioned an open problem as well. We will state the open problem in section 5.4.3 and give a counter-example to the open problem 21 in Roberts and Rosenthal's paper [14] (2007) in section 5.4.3. Finally we present some positive results about the relationship between ergoidicity and recurrence on the space $\mathcal{X} \times \mathcal{Y}$.

5.4.1 Recurrence On The Product Space Is NOT Sufficient For Ergodicity

Even we take finite kernels with good ergodic property(uniformly ergodic) so that we can make the adaptive MCMC recurrent, we still can not guarantee the AMCMC is ergodic with respect to the target distribution π. A good counter example is one-two version running example which was presented in Roberts and Rosenthal(2005) [14] and simulated in the related Java applet. The example was also discussed in Atchade and Rosenthal (2005) [17]. Here we will consider the AMCMC algorithm as a general Markov chain on the product space $\mathcal{X} \times \mathcal{Y}$. We will give the explicit form of the transition matrix on the product space, and analysis the recurrent and ergodic property of such a Markov chain on the product space $\mathcal{X} \times \mathcal{Y}$.

Let $\mathcal{X} = \{1, 2, 3, 4\}$, $\pi(2) = b > 0$ be very small, and $\pi(1) = a$ and $\pi(2) = \pi(3) = \frac{1-a-b}{2} > 0$. Let $\mathcal{Y} = \{1, 2\}$. For $\gamma \in \mathcal{Y}$, let P_γ be the kernel corresponding to a random-walk Metropolis algorithm for $\pi(\cdot)$, with proposal distribution:

$$Q_\gamma(x, \cdot) = Uniform\{x - \gamma, x - \gamma + 1, \cdots, x - 1, x + 1, x + 2, \cdots, x + \gamma\}$$

i.e.uniform on all the integers within γ of x, aside from x itself. The kernel P_γ then proceeds, given X_n and Γ_n, by first choosing a proposal state $Y_{n+1} \sim Q_{\Gamma_n}(X_n, \cdot)$. With probability $\min[1, \frac{\pi(Y_{n+1})}{\pi(X_n)}]$ it then accepts this proposal by setting $X_{n+1} = Y_{n+1}$. Otherwise, with probability $1 - \min[1, \frac{\pi(Y_{n+1})}{\pi(X_n)}]$, it rejects this proposal by setting $X_{n+1} = X_n$.

(If $Y_{n+1} \notin \mathcal{X}$, then the proposal is always rejected; this corresponds to setting $\pi(y) = 0$ for $y \notin \mathcal{X}$.). We define the adaptive scheme such that $\Gamma_n = 2$ if the previous proposal was accepted, otherwise $\Gamma_n = 1$ if the previous proposal was rejected.

We can compute the kernels induced by the proposals Q_i, $i = 1, 2$:

$$P_1 = \begin{pmatrix} \frac{2a-b}{2a} & \frac{b}{2a} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{b}{1-a-b} & \frac{1}{2} - \frac{b}{1-a-b} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$P_2 = \begin{pmatrix} \frac{3}{4} - \frac{b}{4a} & \frac{b}{4a} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{a}{2(1-a-b)} & \frac{b}{2(1-a-b)} & \frac{3}{4} - \frac{a+b}{2(1-a-b)} & \frac{1}{4} \\ 0 & \frac{b}{2(1-a-b)} & \frac{1}{4} & \frac{3}{4} - \frac{b}{2(1-a-b)} \end{pmatrix}.$$

In the above AMCMC, we can observe that the distribution of Γ_n given X_0 and Γ_0 does NOT depend on the value of $\{X_i | 0 \leq i \leq n-1\}$, therefore we call this kind of Markovian AMCMC. The $n-$th transition kernel $Q_{(n)}$ induced by Markovian adaptive algorithm is as below:

$$Q^{(n)}((x, \gamma), A \times B) = \int_A \int_B \Gamma_n(d\gamma_1 | x, y, \gamma) P_\gamma(x, dy).$$

Then in the one-two running example, if given the value of $X_{n-1} = x, X_n = y$ and $\Gamma_{n-1} = \gamma$, then Γ_n is a measurable function of x, y and γ. We have:

$$\Gamma_n(x, y, \gamma) = \delta(x = y) + 2\delta(x \neq y).$$

So we can compute the $n - th$ transition kernel on $(\mathcal{X} \times \mathcal{Y})$:

$$\begin{aligned} Q((x, \gamma), y \times \gamma_1) &= \int_A \int_B \Gamma_n(d\gamma_1 | x, y, \gamma) P_\gamma(x, dy) \\ &= P_\gamma(x, y)\delta(x = y)\delta(\gamma_1 = 1) + P_\gamma(x, y)\delta(x \neq y)\delta(\gamma_1 = 2). \end{aligned}$$

Since the transition kernel is independent of n, the one-two version running example presents a general Markov Chain with transition kernle Q as:

$$Q = \begin{pmatrix} \frac{2a-b}{2a} & 0 & 0 & \frac{b}{2a} & 0 & 0 & 0 & 0 \\ \frac{3}{4}-\frac{b}{4a} & 0 & 0 & \frac{b}{4a} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{b}{1-a-b} & \frac{1}{2}-\frac{b}{1-a-b} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{a}{2(1-a-b)} & 0 & \frac{b}{2(1-a-b)} & \frac{3}{4}-\frac{a+b}{2(1-a-b)} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{b}{2(1-a-b)} & 0 & \frac{1}{4} & \frac{3}{4}-\frac{b}{2(1-a-b)} & 0 \end{pmatrix}.$$

Now we take the value $a = 0.1$ and $b = 0.01$, then $\pi(1) = 0.1$; $\pi(2) = 0.01$; $\pi(3) = \pi(4) = 0.445$.

And we have the following lemma:

Lemma 5.13. *The above one-two version running example is recurrent, but for any starting value (x_*, γ_*), and $A \in \mathcal{B}\{\mathcal{X}\}$, we have:*

$$\lim_{n\to\infty} P_{(x_*,\gamma_*)}(X_n \in A) \neq \pi(A).$$

Proof. Let us calculate the eigenvalues of the above transition matrix, we have: $\lambda_1 = 1$; $\lambda_2 = 0.95445494$; $\lambda_3 = 0.12887658 + 0.4670861i$; $\lambda_4 = 0.12887658 - 0.4670861i$; $\lambda_5 = -0.25615654$; $\lambda_6 = 0.03778642 + 0.1057364i$; $\lambda_7 = 0.03778642 - 0.1057364i$; $\lambda_8 = -0.09286036$. Then compute the eigenvector of Q^T with respect to the eigenvalue $\lambda_0 = 1$, it is

$$(-0.48637045, -0.03354279, -0.00867102, -0.03468408,$$
$$-0.49208038, -0.36554543, -0.51525761, -0.34609757)$$

i.e the stationary distribution $\tilde{\pi}$ is: $\tilde{\pi}(1,1) = 0.213110130$, $\tilde{\pi}(1,2) = 0.014697250$, $\tilde{\pi}(2,1)0.003799331$, $\tilde{\pi}(2,2) = 0.015197323$, $\tilde{\pi}(3,1) = 0.215612017$, $\tilde{\pi}(3,2) = 0.160168927$,

$\tilde{\pi}(4,1) = 0.225767451$, $\tilde{\pi}(4,2) = 0.151647571$. Therefore for any start value (x_*, γ_*), we have:

$$\begin{aligned}\lim_{n\to\infty} P_{(x_*,\gamma_*)}(X_n = 1) &= \lim_{n\to\infty} P_{(x_*,\gamma_*)}(X_n = 1, \Gamma_n = 1) + P_{(x_*,\gamma_*)}(X_n = 1, \Gamma_n = 1) \\ &= 0.21311 + 0.014697 = 0.227807.\end{aligned}$$

similarly

$$\lim_{n\to\infty} P_{(x_*,\gamma_*)}(X_n = 2) = 0.003799 + 0.015197 = 0.018996,$$

$$\lim_{n\to\infty} P_{(x_*,\gamma_*)}(X_n = 3) = 0.215612 + 0.160168 = 0.37578,$$

$$\lim_{n\to\infty} P_{(x_*,\gamma_*)}(X_n = 4) = 0.225767 + 0.151647 = 0.377414.$$

Therefore for any $1 \leq i, j \leq 4$, we have:

$$E_i[\eta_j] = \infty$$

because $P_i(\eta_j = \infty) = 1$. But we can observe that $P_{(x_*,\gamma_*)}(X_n \in A) \to_{n\to\infty} \pi'(A)$ which is the marginal distribution of $\tilde{\pi}$, however $\pi'(\cdot) \neq \pi(\cdot)$. □

5.4.2 $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}$ is bounded in probability is NOT Necessary For Ergodicity

Consider theorem 4.5, we are wondering whether $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}$ is bounded in probability is necessary and sufficient under diminishing adaption condition or not. Unfortunately it is not a necessary condition of the ergodicity. That is:

Theorem 5.2. *Under Diminishing Adaption condition, $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}$ is bounded in probability is NOT a necessary condition of the ergodicity, although it is sufficient.*

Proof. Consider the state space $\mathcal{X} = [0,1]$, $\pi(\cdot) = Unif[0,1]$,the parameter space $\mathcal{Y} = \{k \in \mathbb{Z} | k \geq 2\}$ and proposal distribution $Q_k(x,\cdot) \sim Uniform[x - \frac{k}{2}, x + \frac{k}{2}]$. Denote P_k is the transition kernel induced by Metropolis-Hasting algorithm with proposal distribution Q_k. Obviously P_k is uniformly ergodic. Note that if the proposal is not in $[0,1]$, then the proposal is always rejected. Since for any fixed $\epsilon > 0$, we can prove that:

$$\begin{aligned}
& \inf\{n \geq 1 : \|P_k^n(x,\cdot) - \pi(\cdot)\| \leq \epsilon\} \\
\geq\ & \inf\{n \geq 1 : \|P_k^n(x,\{x\}) - \pi(\{x\})\| \leq \epsilon\} \\
\geq\ & \inf\{n \geq 1 : P_k^n(x,\{x\}) \leq \epsilon\} \\
=\ & \inf\{n \geq 1 : [\int_{R-[0,1]} (1 - \min\{1, \frac{\pi(y)q_k(y,x)}{\pi(x)q_k(x,y)}\})q_k(x,y)dy]^n \leq \epsilon\} \\
=\ & \inf\{n \geq 1 : [1 - \frac{1}{k}]^n \leq \epsilon\}
\end{aligned}$$

Suppose $a_k = \inf\{n \geq 1 : [1 - \frac{1}{k}]^n \leq \epsilon\}$, then we have:

$$\begin{aligned}
\lim_{k\to\infty} M_\epsilon(x, \frac{1}{k}) &= \lim_{k\to\infty} \inf\{n \geq 1 : \|P_k^n(x,\cdot) - \pi(\cdot)\| \leq \epsilon\} \\
&\geq \lim_{k\to\infty} a_k \\
&= \infty.
\end{aligned}$$

Next we can construct the adaptive scheme, let

$$r_k = \inf\{r : \|P_k^r(x,\cdot) - \pi(\cdot)\| \leq \frac{1}{k}\},$$

and $s_k = \sum_{i=1}^k r_i$. Then consider the independent adaptive scheme, we will use the kernel P_k from the $s_{k-1}-$step to $(s_k - 1)-$step. Obviously, as $k \to \infty$, such an adaptive MCMC satisfies the Diminishing Adaption property. Following that

$$\lim_{k\to\infty} \|P_k^{r_k}(x,\cdot) - \pi(\cdot)\| \leq \lim_{k\to\infty} \frac{1}{k} = 0,$$

we can prove the ergodicity. However for any $x \in \mathcal{X}$, we have

$$\lim_{k\to\infty} M_\epsilon(x,k) = \infty,$$

which is NOT bounded in probability. □

5.4.3 The Open Problem 21 In Roberts And Rosenthal [48]

Following theorem 6.2, it is possible to find out weaker conditions to ensure the ergodicity. We can observe that in the theorem 4.5 the adaptive chain pair (X_n, Γ_n) has good "fast convergence" property in probability. If we denote:

Condition (d_1)**:** for all $\epsilon > 0$, there is $m \in \mathbb{N}$ such that $P[M_\epsilon(X_n, \Gamma_n) < m \text{ i.o.} | X_0 = x_*, \Gamma_0 = \gamma_*] = 1$.

Then we can state the following open problem.

Open Problem 21. Consider an adaptive MCMC algorithm with Diminishing Adaptation. Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{Y}$. Does condition (d_1) imply that $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$? The problem seems reasonable, however the following example gives us the negative answer.

Consider $\mathcal{X} = R \bmod Z$ i.e. the state space is the real number mod the integers. Define $\mathcal{Y} = \mathbf{N} \cup \mathcal{X}$, and suppose $Z_{k,x}$ are random variable with distribution $Uniform[x - \frac{1}{2^{k+1}}, x + \frac{1}{2^{k+1}}]$ for any $(x, \gamma) \in \mathcal{X} \times \mathcal{Y}$. When $k \in N$, we define:

$$P_k(x, A) = \frac{1}{2^k} P(Z_{k,x} \in A) + (1 - \frac{1}{2^k}) \delta_x(A).$$

When $y \in \mathcal{X}$, suppose $\pi(\cdot)$ is the Lebesgue measure on $\mathcal{X}$.

we define:

$$P_y(x, A) = \begin{cases} \frac{2}{3}\pi(A) + \frac{1}{3}\delta_x(A) & \text{, if } x \neq y; \\ \frac{2}{3} Uniform[0, \frac{3}{4}] + \frac{1}{3}\delta_0(A) & \text{, if } x = y. \end{cases}$$

Lemma 5.14. *For each $k \in \mathbb{N}$, P_k is stationary with respect to π.*

Proof. It is suffice to prove that for any interval $A = [a, b] \subset [0, 1]$ we have:

$$\int_{\mathcal{X}} P_k(x, A)\pi(dx) = \pi(A).$$

Case 1:$|b - a| \geq \frac{1}{2^k}$

$$\begin{aligned}
\int_{\mathcal{X}} P_k(x,A)\pi(dx) &= \frac{1}{2^k} \times \int_0^1 P(Z_{x,k} \in A)dx + (1-\frac{1}{2^k})\pi(A) \\
&= \frac{1}{2^k} \times [2^k \int_{a-\frac{1}{2^{k+1}}}^{a+\frac{1}{2^{k+1}}} [x + \frac{1}{2^{k+1}} - a]dx + 2^k \int_{b-\frac{1}{2^{k+1}}}^{b+\frac{1}{2^{k+1}}} [-x + \frac{1}{2^{k+1}} + b]dx \\
&+ (b-a-\frac{1}{2^k})] + (1-\frac{1}{2^k})\pi(A) \\
&= \frac{1}{2^k} \times [2^{k+1} \int_0^{\frac{1}{2^k}} tdt + (b-a-\frac{1}{2^k})] + (1-\frac{1}{2^k})\pi(A) \\
&= b-a.
\end{aligned}$$

Similarly we can prove **Case 2:**$|b-a| < \frac{1}{2^k}$. □

Lemma 5.15. *For each $y \in \mathcal{X}$, P_y is stationary with respect to π.*

Proof.

$$\begin{aligned}
\int_{\mathcal{X}} P_y(x,A)\pi(dx) &= \int_{x \neq y} [\frac{2}{3}\pi(A) + \frac{1}{3}\delta_x(A)]\pi(dx) \\
&= \frac{2}{3}\pi(A) + \frac{1}{3}\pi(A) \\
&= \pi(A).
\end{aligned}$$

□

Define the independent random variable I_n as below:

$$I_n = \begin{cases} 1 & \text{w.p.} \frac{\sqrt{n}-1}{\sqrt{n}}; \\ 0 & \text{w.p.} \frac{1}{\sqrt{n}}. \end{cases}$$

And independent random variable Y_n as below: $Y_0 = Y_1 = 1$ and

$$Y_n = \begin{cases} n+1 & \text{with probability } \frac{1}{n}, \\ n+2 & \text{with probability } \frac{1}{n}, \\ . & \\ . & \\ . & \\ 2n & \text{with probability } \frac{1}{n}. \end{cases}$$

Define the adaptive scheme as:

$$\Gamma_n = \begin{cases} Y_n & \text{, if } I_n = 1; \\ X_n & \text{, if } I_n = 0. \end{cases}$$

Lemma 5.16. *Such an adaptive scheme satisfies the diminishing condition.*

Proof. Actually $P_{Y_n}(x, A) = \frac{1}{n}\sum_{i=n+1}^{2n} P_i(x, A)$, so

$$\begin{aligned} & |P_{\Gamma_{n+1}}(x, A) - P_{\Gamma_n}(x, A)| \\ \leq\ & |P_{Y_{n+1}}(x, A) - P_{Y_n}(x, A)| + P(I_n = 0 \text{ or } I_{n+1} = 0) \\ \leq\ & |\frac{1}{n+1}\sum_{i=n+2}^{2n+2} P_i(x, A) - \frac{1}{n}\sum_{i=n+1}^{2n} P_i(x, A)| + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} \\ \leq\ & \frac{1}{n(n+1)}\sum_{i=n+2}^{2n} P_i(x, A) + |\frac{1}{n+1}P_{2n+1}(x, A) + \frac{1}{n+1}P_{2n+2}(x, A) - \frac{1}{n}P_{n+1}(x, A)| \\ +\ & \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} \\ \leq\ & \frac{1}{n} + \frac{3}{n} + \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} \\ \to\ & 0 \text{ as } n \to \infty. \end{aligned}$$

□

Lemma 5.17. *Given $x_* = 0$ and $\gamma_* = 0$. Then for any $\epsilon > 0$, there is $m \in N$ such that:*

$$P[(X_n, \Gamma_n) \in \mathcal{Z}_{m,\epsilon} \ \ i.o. \ \ |X_* = 0, \Gamma_0 = 0] = 1.$$

Proof. We know P_0 is uniformly ergordic with respect to $\pi(\cdot)$, so for any $\epsilon > 0$ there exists m such that:

$$\|P_0^m(0,\cdot) - \pi(\cdot)\| < \epsilon. \tag{5.1}$$

If we suppose

$$J = \begin{cases} 1 & \text{w.p.}\frac{2}{3}; \\ 0 & \text{w.p.}\frac{1}{3}. \end{cases}$$

Then we can consider $P_x(x, A)$ as the following: if $J = 0$, the chain will move to 0, otherwise select one point on the interval $[0, \frac{3}{4}]$ with uniform distribution.
And we have:

$$P[X_{n+1} = 0, \Gamma_{n+1} = 0 \text{ i.o.}] \geq P[I_n = 0, I_{n+1} = 0 \text{ and} J = 0 \text{ i.o}].$$

Since $\sum_{i=1}^{\infty} P(I_{2i} = 0, I_{2i+1} = 0, J = 0) = \sum_{i=1}^{\infty} \frac{1}{3}\frac{1}{\sqrt{2i(2i+1)}} = \infty$. Following The Borel-Cantelli Lemma in [50] we have:

$$P[I_{2n} = 0, I_{2n+1} = 0 \text{ and} J = 0 \text{ i.o}] = 1.$$

Therefore $P[(X_n, \Gamma_n) = (0,0) \text{ i.o.}] = 1$. Following (5.1) we know that

$$\begin{aligned} 1 &\geq P[(X_n, \Gamma_n) \in \mathcal{Z}_{m,\epsilon} \ \ i.o. \ \ |X_* = 0, \Gamma_* = 0] && (5.2) \\ &\geq P[(X_n, \Gamma_n) = (0,0) \ \ i.o. \ \ |X_* = 0, \Gamma_* = 0] = 1. && (5.3) \end{aligned}$$

□

Lemma 5.18. *Suppose $\{a_i\}_{i=1}^{\infty}$ is a decreasing positive sequence such that $0 < a_i < 1$, and if $\sum_{i=1}^{\infty} a_i < \infty$, then*

$$\lim_{N\to\infty} \prod_{i=N}^{\infty} (1 - a_i) = 1. \tag{5.4}$$

Proof. When $0 < a_i < 1$, we have:

$$ln(1 - a_i) \leq -a_i.$$

Therefore

$$\begin{aligned} 1 &\geq \lim_{N\to\infty} \prod_{i=N}^{\infty} (1 - a_i) \\ &\geq \lim_{N\to\infty} e^{\sum_{i=N}^{\infty}(-a_i)} \\ &= 1. \end{aligned}$$

□

Lemma 5.19. *Given $X_* = 0$ and $\Gamma_* = 0$, we do NOT have $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$.*

Proof. Suppose $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$, that is for any $\epsilon > 0$, there exists N_1 such that for any $n > N$ and $A \in \mathcal{B}(\mathcal{X})$,

$$|P[X_n \in A | X_* = 0, \Gamma_* = 0] - \pi(A)| < \epsilon. \tag{5.5}$$

According to the above adaptive scheme, if $\Gamma_n \in [0, 1]$, then Γ_n must be equal to X_n, in other words the case of kernel $P_y(x, \cdot)$ but $y \neq x$ will NOT happen in this adaptive Markov Chain. So if $X_n \in [0, \frac{3}{4}]$, there are four cases maybe happen at X_{n+1}

Case 1: $X_{n+1} = X_n$;

Case 2: $X_{n+1} = 0$;

Case 3: $X_{n+1} = Z_{x_n,n}$;

Case 4: $X_{n+1} \sim Uniform[0, \frac{3}{4}]$.

Only in the case 3, X_{n+1} maybe jump out of $[0, \frac{3}{4}]$, so $P(X_{n+1} \in [0, \frac{3}{4}] | X_n \in [0, \frac{3}{4}]) > 1 - \frac{1}{2^n}$. Since this is a Markovian adaptive MCMC,

$$\begin{aligned} & P(X_{n+2} \in [0, \frac{3}{4}] | X_n \in [0, \frac{3}{4}]) \\ \geq\ & P(X_{n+2} \in [0, \frac{3}{4}] | X_{n+1} \in [0, \frac{3}{4}]) P(X_{n+1} \in [0, \frac{3}{4}] | X_n \in [0, \frac{3}{4}]) \\ \geq\ & (1 - \frac{1}{2^n})(1 - \frac{1}{2^{n+1}}). \end{aligned}$$

Similarly for any $m > 0$, we have:

$$P(X_{n+m} \in [0, \frac{3}{4}] | X_n \in [0, \frac{3}{4}]) \geq \prod_{i=n}^{n+m-1} (1 - \frac{1}{2^i}). \tag{5.6}$$

Following lemma 5.18 we select $N_2 > 0$ such that $\prod_{i=N_2}^{\infty}(1 - \frac{1}{2^i}) > 1 - \frac{\epsilon}{2}$ Let $N = max\{N_1, N_2\}$, then following (5.3) there exist K large enough such that:

$$P[\exists N \leq n < N^K \text{ such that } (X_n, \Gamma_n) = (0, 0)] > \frac{\frac{3}{4} + 2\epsilon}{1 - \frac{\epsilon}{2}}, \tag{5.7}$$

whenever $(X_n, \Gamma_n) = (0, 0)$, then X_{n+1} must be in $[0, \frac{3}{4}]$, so following (5.6) we have:

$$\begin{aligned}
& P(X_{N^K+1} \in [0, \frac{3}{4}]) \\
= \; & P(X_{N^K+1} \in [0, \frac{3}{4}] | \exists N < n \leq N^K \text{ s.t. } X_n \in [0, \frac{3}{4}]) \cdot P(\exists N < n \leq N^K \text{ s.t. } X_n \in [0, \frac{3}{4}]) \\
\geq \; & \prod_{i=N}^{\infty}(1 - \frac{1}{2^i}) \cdot P[\exists N < n \leq N^K \text{ s.t. } (X_{n-1}, \Gamma_{n-1}) = (0, 0)] \\
\geq \; & (1 - \frac{\epsilon}{2}) \times \frac{\frac{3}{4} + 2\epsilon}{1 - \frac{\epsilon}{2}} \\
= \; & \frac{3}{4} + 2\epsilon.
\end{aligned}$$

Which is conflicting with (5.5). □

5.4.4 Strengthen The Diminishing Adaption Condition

Following the counterexample in the section 5.4.3, we know that the Diminishing Adaption condition and the recurrence property to the "good convergence" set are not sufficient to get the ergodicity of the AMCMC. Therefore we can strengthen the Diminishing Adaption condition such that it can match with the recurrence condition, so that we can use the coupling methods to prove the ergodicity.

For any $m \in \mathbb{N}$ and $\epsilon > 0$, we can define the $i-$th hitting time $\tau^{(i)}_{x,\gamma}(m, \epsilon)$ as below:

$$\tau^{(i)}_{x,\gamma}(m, \epsilon) = \min\{n > \tau^{(i-1)}_{x,\gamma}(m, \epsilon) | M_\epsilon(X_n, \Gamma_n) \leq m \text{ given } X_0 = x, \Gamma_0 = \gamma\},$$

and the hitting number within n step

$$c_{x,\gamma}^{m,\epsilon}(n) = \text{ the number of } \{0 \le j \le n | M_\epsilon(X_j, \Gamma_j) \le m \text{ given } X_0 = x, \Gamma_0 = \gamma\}.$$

Furthermore we can define:

$$s_{x,\gamma}^{(i)}(m,\epsilon) = \sum_{j=\tau_{x,\gamma}^{(i)}(m,\epsilon)+1}^{\tau_{x,\gamma}^{(i+1)}(m,\epsilon)} D_j,$$

and denote

Condition (d_2): Suppose that for all $\epsilon > 0$, there is $m \in \mathbb{N}$ such that $P[M_\epsilon(X_n, \Gamma_n) < m \text{ i.o.} | X_0 = x_*, \Gamma_0 = \gamma_*] = 1$ and $s_{x,\gamma}^{(i)}(m,\epsilon) \to_{i\to\infty} 0$ in probability.

Then we have the following theorem:

Theorem 5.3. *Consider an adaptive MCMC algorithm , let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{Y}$. Then conditon (d_2) implies $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$.*

Proof. For any $\epsilon > 0$, there is $m \in \mathbb{N}$ such that

$$P[M_\epsilon(X_n, \Gamma_n) < m \text{ i.o.} | X_0 = x_*, \Gamma_0 = \gamma_*] \;=\; 1,$$

and there exists $N_1 > 0$ such that for any $n > N_1$ we have:

$$P\left[\sum_{j=n}^{n+m} s_{x,\gamma}^{(i)}(m,\epsilon) > \epsilon\right] \;\le\; \epsilon.$$

Following $P[M_\epsilon(X_n, \Gamma_n) < m \text{ i.o.} | X_0 = x_*, \Gamma_0 = \gamma_*] = 1$, we know that there is $N > 0$ such that

$$P[c_{x,\gamma}^{m,\epsilon}(N) > N_1 + m] > 1 - \epsilon. \tag{5.8}$$

Consider any $n > N$, the above formula indicates that:

$$P[\exists k > N_1 + m \text{ such that } \tau_{x,\gamma}^{(k)}(m,\epsilon) \le n < \tau_{x,\gamma}^{(k+1)}(m,\epsilon)] > 1 - \epsilon.$$

We set $l = \tau_{x,\gamma}^{(k-m)}(m,\epsilon)$. we can construct a second chain $\{X_i'\}_{i=l}^n$ such that $X_l' = X_l$ and $X_i' \sim P_{\Gamma_l}(X_{i-1}, \cdot)$ for $l \le i \le n$. If we denote the event $E = \{\sum_{i=l}^n P(X_i' \ne X_i) < \epsilon\}$, then from (5.8) we have:

$$P[E] > 1 - \epsilon.$$

On the other hand we have:

$$\|P_{\Gamma_l}^{n-l}(X_l,\cdot)-\pi(\cdot)\| \;\leq\; \|P_{\Gamma_l}^{l+m}(X_l,\cdot)-\pi(\cdot)\leq\epsilon.$$

We can construct $Z\sim\pi(\cdot)$, then

$$\begin{aligned}
&\|P(X_n\in\cdot|X_0=x,\Gamma_0=\gamma)-\pi(\cdot)\|\\
\leq\;& P(X_n\neq Z|X_0=x,\Gamma_0=\gamma)\\
\leq\;& P(X_n\neq X_n^{'},E|X_0=x,\Gamma_0=\gamma)+P(X_n^{'}\neq Z,E|X_0=x,\Gamma_0=\gamma)+P(E^c|X_0=x,\Gamma_0=\gamma)\\
\leq\;& 3\epsilon
\end{aligned}$$

i.e. $T(x,\gamma,n)<3\epsilon$. □

Following the Theorem 5.3, we can get the following corollary easily.

Corollary 5.1. *Consider an adaptive MCMC algorithm such that* $\sum_{i=1}^{\infty}D_i<\infty$ *in probability. Let* $x_*\in\mathcal{X}$ *and* $\gamma_*\in\mathcal{Y}$*. Suppose that for all* $\epsilon>0$*, there is* $m\in\mathbb{N}$ *such that* $P[M_\epsilon(X_n,\Gamma_n)<m\ i.o.|X_0=x_*,\Gamma_0=\gamma_*]=1$*. Then* $\lim_{n\to\infty}T(x_*,\gamma_*,n)=0$.

Proof. Since $\sum_{i=1}^{\infty}D_i<\infty$ in probability, we know that $s_{x,\gamma}^{(i)}(m,\epsilon)\to_{i\to\infty}0$ in probability. Therefore following the Theorem 5.3, we have the conclusion. □

5.5 The Convergence Rate Of AMCMC

5.5.1 Discussion On The Convergence Rate Of Finite AMCMC

Let us start our discussion with some special adaptive scheme- finite AMCMC algorithm. Following proposition 4.2, we know that the finite AMCMC algorithm is ergodic for the target distribution $\pi(\cdot)$. Intuitively if each kernel P_γ is geometrically ergodic, we hope the finite AMCMC is also geometrically ergodic, i.e. there is $\rho<1$ and $K(x,\gamma)<\infty$ such that $T(x,\gamma,n)\leq K(x,\gamma)\rho^n$ for all $n\in\mathbb{N}$. However we have the following negative theorem:

Theorem 5.4. *There exists a finite adaptive scheme where each P_γ is geometrically ergodic with respect to $\pi(\cdot)$, but where the finite adaptive scheme fails to be geometrically ergodic.*

Proof. Let $\chi = \mathbf{R}$ with $\pi(\cdot) = N(0,1)$. Let $\mathcal{Y} = (0,\infty)$, and for $\gamma \in \mathcal{Y}$ let P_γ be a Metropolis algorithm with proposal distribution $Q(x,\cdot) = N(x,\gamma^2)$. Then each such P_γ is geometrically ergodic (See e.g. Roberts and Tweedie [41],1996). On the other hand, consider an adaptive scheme such that $\Gamma_0 = 1$ and τ is the first time a proposal is accepted, and $\Gamma_{n+1} = 2\Gamma_n$ for $n < \tau$, with $\Gamma_{n+1} = \Gamma_\tau$ for $n \geqslant \tau$. Now we suppose that there exist $M(x,\gamma) < \infty$ and $\gamma \in (0,1)$ such that:

$$|P(X_n \in A | X_0 = x, \Gamma_0 = \gamma) - \pi(A)| \leqslant M(x,\gamma)\rho^n, \tag{5.9}$$

for each $A \in \mathcal{B}(\chi)$ and $n \in \mathbf{N}$. Consider $X_0 = 0$, $\Gamma_0 = 1$ and $A = \mathbf{R} \setminus \{0\}$. Then we have:

$$|1 - P(X_n \in A | X_0 = 0, \Gamma_0 = 1)| \leqslant M(0,0)\rho^n. \tag{5.10}$$

Now we denote $P_{0,1}(X_n \in A) = P(X_n \in A | X_0 = 0, \Gamma_0 = 1)$, and we can write it in following form:

$$\begin{aligned} P_{0,1}(X_n \in A) &= \sum_{j=1}^{\infty} P_{0,1}(X_n \in A, \tau = j) && (5.11)\\ &= \sum_{j \leq n} P_{0,1}(X_n \in A, \tau = j) && (5.12)\\ &= \sum_{j \leq n} P_{0,1}(\tau = j). && (5.13) \end{aligned}$$

Equation (5.12) follows that if $\tau > n$, which means X_n is still zero, X_n is not in A; and equation (5.13) follows that if $\tau \leqslant n$, $X_n \in A$ with probability 1. So (5.10) can be written as:

$$P_{0,1}(\tau \geqslant n+1) \leqslant M(0,1)\rho^n. \tag{5.14}$$

Suppose Y_i denote the random variable generated by the n-th proposal distribution, and we find that when $\tau \geqslant n+1$, the first n Y_i are independent, so we have:

$$\begin{aligned}
P_{0,1}(\tau \geqslant n+1) &= E(P_{0,1}(\tau \geqslant n+1|Y_1, \cdots, Y_n)) \\
&= \Pi_{i=0}^{n} \int_{\mathcal{X}} (1 - \exp\{-\frac{y_i^2}{2}\}) d\mathcal{L}\{Y_i\} \\
&= \Pi_{i=0}^{n} (1 - \frac{1}{\sqrt{2\pi}2^i} \int_{\mathcal{X}} \exp\{-\frac{y^2}{2}\} \exp\{-\frac{y^2}{2 \cdot 2^{2i}}\} dy) \\
&= \Pi_{i=0}^{n} (1 - \frac{1}{\sqrt{2\pi}2^i} \cdot \sqrt{2\pi} \frac{2^i}{\sqrt{2^{2i}+1}} \cdot \frac{1}{\sqrt{2\pi}\frac{2^i}{\sqrt{2^{2i}+1}}} \int_{\mathcal{X}} \exp\{-\frac{y^2}{2 \cdot \frac{2^{2i}}{2^{2i}+1}}\} dy) \\
&= \Pi_{i=0}^{n} (1 - \frac{1}{\sqrt{2^{2i}+1}}) \\
&\geq \Pi_{i=1}^{n} (1 - \frac{1}{2^i}) \\
&\geq \Pi_{i=1}^{n} (1 - \frac{1}{i}) \\
&= \frac{1}{n}.
\end{aligned}$$

So following (5.10), we have:

$$\begin{aligned}
M(0,0)\rho^n &\geq |1 - P_{0,1}(X_n \in A)| \\
&= |1 - \sum_{j \leq n} P_{0,1}(\tau = j)| \\
&= P_{0,1}(\tau \geq n+1) \\
&\geq \frac{1}{n}.
\end{aligned}$$

That is we have $\rho \geq [\frac{1}{n \cdot M(0,0)}]^{\frac{1}{n}}$, and we know that $lim_{n\to\infty}[\frac{1}{n \cdot M(0,0)}]^{\frac{1}{n}} = 1$, so we have $\rho \geq 1$ which is contradicting with assumption! □

5.5.2 Discussion On The Convergence Rate Of Uniformly Converging AMCMC

Following the Theorem 4.2, we know that if the kernel family $\{P_\gamma\}$ is simultaneous uniform ergordicity, we can prove that the AMCMC is ergodic under the Diminishing

Adaptation. We also hope that the AMCMC keep the uniformly ergordicity with respect to the target distribution as each kernel does. However the following example shows that it may not be true.

Consider $\mathcal{X} = (0, 1]$, $\mathcal{Y} = (0, 1] \times \mathbf{N}$, $\pi(\cdot)$ is the Lebesgue measure on $\mathcal{X}$, and

$$g(x) = x^{-\frac{1}{2}},$$

therefore $\pi(g) = 2$. Furthermore, for $(\gamma, k) \in \mathcal{Y}$ define the kernel $P_{(\gamma,k)}$ by:

$$P_{(\gamma,k)}(x, A) = \begin{cases} \frac{2}{3}\pi(A) + \frac{1}{3}\delta_x(A) & \text{if } x \neq \gamma \\ \frac{2}{3}\pi(A) + \frac{1}{3}\delta_{\frac{1}{4^k}}(A) & \text{if } x = \gamma. \end{cases}$$

We construct the adaptive scheme as below:

first we define $\{I_n\}_{n=1}^{\infty}$ to be an independent random variable sequence such that:

$$I_n = \begin{cases} 1 & \text{with probability} \frac{1}{n} \\ 0 & \text{with probability} \frac{n-1}{n}; \end{cases}$$

secondly we let $\Gamma_{n+1} = \Gamma_n \times (1 - I_n) + (X_{n+1}, n + 1) \times I_n$.

Theorem 5.5. *For the above adaptive MCMC which satisfies conditions (a) and (b), for any $x \in \mathcal{X}$, there exits a measurable set B such that $\sum_{n=1}^{\infty} A^{(n)}(x, B) = \infty$.*

Proof. Consider the set $B = \{\frac{1}{4^k} | k = 1, 2, \cdots, \}$, suppose for any start valuve $X_0 = x$ and $\Gamma_0 = \gamma$, we have:

$$\sum_{i=1}^{\infty} A^i((x, \gamma), B) < \infty,$$

then for any $0 < \epsilon < 1$, there exists $N_{x,\gamma} > 0$ such that

$$\sum_{i=N+1}^{\infty} A^i((x, \gamma), B) \leq \epsilon.$$

Because

$$P(X_{n+1} = \frac{1}{4^n} | \Gamma_n = (X_n, n)) \geq \frac{1}{3}$$

and

$$P(\Gamma_n = (X_n, n)) \geq P(I_n = 1),$$

we can get

$$\begin{aligned} P(X_{n+1} = \frac{1}{4^n}) &\geq P(X_{n+1} = \frac{1}{4^n}, \Gamma_n = (X_n, n)) \\ &\geq \frac{1}{3} P(\Gamma_n = (X_n, n)) \\ &\geq \frac{1}{3} P(I_n = 1) \\ &= \frac{1}{3n}. \end{aligned}$$

Then following the Borel-Cantelli lemma see Jeffrey S. Rosenthal [50] (2000), we have:

$$\begin{aligned} 1 &= P(\exists\ m > N_{x,\gamma} \text{ s.t. } I_m = 1) \\ &\leq \sum_{i=N+1}^{\infty} P^i((X_i = \frac{1}{4^i} | X_0 = x, \Gamma_0 = \gamma) \\ &\leq \sum_{i=N+1}^{\infty} A^i((x,\gamma), B) \\ &\leq \epsilon, \end{aligned}$$

Contradiction!! So we have $\sum_{i=1}^{\infty} A^i((x,\gamma), B) = \infty$. Since $\pi(B) = 0$, we can get:

$$\sum_{i=1}^{\infty} T^i((x,\gamma), B) < \infty.$$

Therefore $A^i((x,\gamma), \cdot)$ is neither uniformly nor geometrically ergodic. □

Chapter 6

Weak Law of Large Numbers for AMCMC

6.1 Introduction

Usually we also want to estimate the integral $\pi(g) = \int_{\mathcal{X}} g(x)\pi(dx)$ of various functions $g : \mathcal{X} \to R$ using the laws of large numbers for ergodic averages of the form:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to_{n\to\infty} \pi(g) \text{ in probability or almost surely}$$

There are many references e.g Tierney [53](1994), Meyn and Tweedie [37](1993) which give the proof and applications of the LLN of general Markov Chains. Regarding the LLN of AMCMC, there are also many papers e.g. Andieu and Achade [2] (2007), Andrieu and Moulines [3](2005), Andrieu and Robert [5](2001), Atchade and Rosenthal [6](2005) giving the proof under various conditions. Based on theorem 4.3, we know that the simultaneous uniform ergodicity and diminishing adaptation are sufficient to ensure the WLLLN for bounded function. This also leads another questions: Does the WLLN hold for all unbounded $g \in L(\pi)$ under the same conditions? We will present counter-examples to demonstrate that when g is unbounded the conditions in the Theorem 4.3 are not enough to guarantee that the the weak law of large numbers (WLLN) holds. Then

we show various theoretical results of the WLLN for the adaptive Metropolis-Hasting algorithm and unbounded measurable function g, then we will apply our results to the Adaptive Metropolis algorithm proposed by Haario et al.[26] (2001). Finally we will prove the WLLN under the conditions of Theorem 5.1.

6.2 The Counter Example

Consider the example constructed in section 5.5.2 and we have the following theorem:

Theorem 6.1. *There exists adaptive MCMC algorithm satisfies conditions (a) and (b) and $\pi(|g|) < \infty$, but the WLLN does NOT hold.*

According to the construction of the example, we can show that

Lemma 6.1. *The adaptive MCMC algorithm in section 5.5.2 satisfies conditions (a) and (b).*

Proof. Obviously each $P_{(\gamma,k)}$ is stationary with respect to π, and $\|P_{(\gamma,k)}(x,\cdot)-\pi(\cdot)\|_{var} \leq \frac{1}{3}$ for any (γ, k), so such a family of kernels satisfy the condition (a) following the Proposition 7 in Roberts and Rosenthal (2004);
And following the definition of Γ_n, we have:

$$\begin{aligned} D_n &= \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \\ &\leq P(\Gamma_{n+1} \neq \Gamma_n) \\ &= P(I_n = 1) \\ &= \frac{1}{n}. \end{aligned}$$

Therefore we have the conditions (a) and (b) holds. □

To prove the Theorem 6.1, we show the following lemmas first:

Lemma 6.2. *For any $\epsilon > 0$ and any sequence $\{x_i\}_{i=0}^{\infty}$, if n and k are two positive integers such that $n < k < \frac{2^n-(1+\epsilon)n-1}{1+\epsilon}$ and we also have $g(x_n) = 2^n$ then:*

$$\frac{\sum_{i=1}^{k} g(x_i)}{k} - 2 > \epsilon. \tag{6.1}$$

Proof. Since $\frac{k+2^n-1}{k}$ strictly decreases with respect to k and $g(x) \geq 1$, we have:

$$\begin{aligned}
\frac{\sum_{i=1}^{k} g(x_i)}{k} - 2 - \epsilon &\geq \frac{k-1+2^n}{k} - 2 - \epsilon \\
&\geq \frac{\frac{2^n-(1+\epsilon)n-1}{1+\epsilon} - 1 + 2^n}{\frac{2^n-(1+\epsilon)n-1}{1+\epsilon}} - 2 - \epsilon \\
&\geq 1 + \frac{2^n - 1}{\frac{2^n-(1+\epsilon)n-1}{1+\epsilon}} - 2 - \epsilon \\
&\geq 1 + \frac{2^n - 1}{2^n - (1+\epsilon)n - 1} \times [1+\epsilon] - 2 - \epsilon \\
&> 1 + 1 + \epsilon - 2 - \epsilon \\
&= 0.
\end{aligned}$$

□

Lemma 6.3. *For any $\epsilon > 0$, there exists M_ϵ such that for any $m > M_\epsilon$ we have:*

$$\frac{2^{m+1} - (m+1)(1+\epsilon) - 1}{1+\epsilon} > m^2.$$

Proof. Denote $h_m = \frac{2^{m+1}-(m+1)(1+\epsilon)-1}{1+\epsilon} - m^2$, then we have $\lim_{m\to\infty} h_m = \infty$. Therefore there exists M_ϵ such that for any $m > M_\epsilon$ we have $h_m > 0$, i.e. $\frac{2^{m+1}-(m+1)(1+\epsilon)-1}{1+\epsilon} > m^2$. □

For any $0 < \epsilon < \frac{1}{6}$, we define $N_\epsilon = \max\{M_\epsilon, \frac{1}{1-6\epsilon}\}$, then we can prove that:

Lemma 6.4. *For any $X_0 = x$, $\Gamma_0 = \gamma$ and $0 < \epsilon < \frac{1}{6}$, then we have:*

$$P(|\frac{\sum_{i=1}^{n} g(X_i)}{n} - \pi(g)| > \epsilon | X_0 = x, \Gamma_0 = \gamma) > 2\epsilon \quad \textit{for any} \quad n > N_\epsilon^2.$$

Proof. For any $n > N_\epsilon^2$, we have

$$\begin{aligned} & P(I_m = 0, \text{for any } m \text{ satisfies} \lfloor\sqrt{n}\rfloor + 1 \le m \le n) \\ = & \prod_{i=\lfloor\sqrt{n}\rfloor+1}^{n} \frac{i}{i+1} \\ = & \frac{\lfloor\sqrt{n}\rfloor}{n} \\ \le & \frac{1}{\sqrt{n}} \\ \le & \frac{1}{N_\epsilon}, \end{aligned}$$

then

$$P(\exists m, \lfloor\sqrt{n}\rfloor + 1 < m < n, I_m = 1) \ge \frac{N_\epsilon - 1}{N_\epsilon}.$$

Whenever $\Gamma_{n+1} = (X_{n+1}, n+1)$, we have $g(X_{n+1}) = 2^{n+1}$ w.p. $\frac{1}{3}$. Since $N_\epsilon > \frac{1}{1-6\epsilon}$, we have $\frac{N_\epsilon - 1}{3N_\epsilon} > 2\epsilon$. Therefore:

$$P(\exists m,\ \lfloor\sqrt{n}\rfloor + 1 < m < n, g(X_m) = 2^m)) > \frac{N_\epsilon - 1}{3N_\epsilon} > 2\epsilon. \tag{6.2}$$

Also since $m > N_\epsilon$, lemma 6.3 indicates that for any $\lfloor\sqrt{n}\rfloor + 1 < m < n$ we have:

$$\frac{2^m - (1+\epsilon)m - 1}{1+\epsilon} > m^2 + 1 > (\lfloor\sqrt{n}\rfloor + 1)^2 + 1 > n + 1.$$

Following lemma 6.4 and $m < n < \frac{2^m - (1+\epsilon)m - 1}{1+\epsilon}$, we know that

$$\frac{\sum_{i=1}^n g(x_i)}{n} - 2 \ \ge \ \epsilon.$$

Therefore:

$$\begin{aligned} & P(|\frac{\sum_{i=1}^n g(X_i)}{n} - 2| > \epsilon) \\ \ge & P(\frac{\sum_{i=1}^n g(X_i)}{n} - 2 \ge \epsilon) \\ \ge & P(\exists m, \lfloor\sqrt{n}\rfloor < m < n, g(X_m) = 2^m)) \\ > & 2\epsilon, \end{aligned}$$

the last inequality is from (6.2). □

Based on all above technical results, we start to prove the Theorem 6.1.

Proof. Consider the above example, following all the lemmas above we know that for any $\epsilon > 0$, we have:

$$\limsup_{n\to\infty} P(|\frac{\sum_{i=1}^{n} g(X_i)}{n} - \pi(g)| > \epsilon) > 2\epsilon.$$

In other words, we do NOT have:

$$\lim_{n\to\infty} P(|\frac{\sum_{i=1}^{n} g(X_i)}{n} - \pi(g)| > \epsilon) = 0.$$

So the WLLN does NOT hold in this example. □

6.3 Summable Adaptive Conditions

From the above counter-example, we know that conditions (a) and (b) are not sufficient conditions to the WLLN of unbounded functions, so we need to strengthen them. Intuitively if n is large enough, for any k, $l > n$, Γ_k and Γ_l are "almost" the same, then the WLLN may hold for any $g \in L(\pi)$. Let us consider the following condition:
$(b^{'})$[Summable Adaption]$\sum_{i=1}^{\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{i+1}}(x,\cdot) - P_{\Gamma_i}(x,\cdot)\| < \infty$. Actually we can prove the following theorem:

Theorem 6.2. *Consider an adaptive MCMC algorithm. Suppose that conditions (a) and* $(b^{'})$ *hold. Let* $g : \mathcal{X} \to R$ *be a measurable function such that* $\pi(|g|) < \infty$. *Then for any starting values* $x \in \mathcal{X}$ *and* $\gamma \in \Gamma$, *conditional on* $X_0 = x$ *and* $\Gamma_0 = \gamma$ *we have:*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$.

Proof. Denote

$$S_n = \sum_{i=n}^{\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{i+1}}(x,\cdot) - P_{\Gamma_i}(x,\cdot)\|.$$

For any $\epsilon > 0$, following condition $(b^{'})$, there exists N_1 such that

$$P(S_{N_1} > \epsilon) < \frac{\epsilon}{4}.$$

We can denote $E = \{S_{N_1} < \epsilon\}$. Since $|g| < \infty$, there exists N_2, such that for any $n > N_2$

$$P(|\frac{\sum_{i=1}^{N_1} g(X_i)}{n}| > \frac{\epsilon}{2}) < \frac{\epsilon}{4}.$$

Define $N = \max\{N_1, N_2\}$, and we can construct a second chain $\{X_n^{'}\}_{n=N}^{\infty}$ on E such that $X_N^{'} = X_N$ and $X_n^{'} \sim P_{\Gamma_N}(X_{n-1}^{'}, \cdot)$ for $n > N$, and such that:

$$\sum_{n=N}^{\infty} P(X_n \neq X_n^{'}, E) < \frac{\epsilon}{4}.$$

Define the events: $B^n(\epsilon) = \{|\frac{\sum_{i=N+1}^{n} g(X_i^{'})}{n}| > \frac{\epsilon}{2}\}$, then following the Law of Large Numbers of Markov chain (See Theorem 17.3.2 in [37]), we can get

$$\lim_{n\to\infty} P(B^n(\epsilon)|X_N, \Gamma_N) = 0.$$

Then

$$\begin{aligned}
\lim_{n\to\infty} P(B^n(\epsilon)) &= \lim_{n\to\infty} E(P(B^n(\epsilon)|X_N, \Gamma_N)) \\
&= E(\lim_{n\to\infty} P(B^n(\epsilon)|X_N, \Gamma_N)) \\
&= 0
\end{aligned}$$

That is when n is large enough we have $P(B^n(\epsilon)) < \frac{\epsilon}{4}$. Therefore we have

$$\begin{aligned}
& P(|\frac{\sum_{i=1}^{n} g(X_i)}{n}| > \epsilon) \\
\leq\ & P(|\frac{\sum_{i=1}^{N} g(X_i)}{n}| > \frac{\epsilon}{2}) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}) \\
\leq\ & P(|\frac{\sum_{i=1}^{N} g(X_i)}{n}| > \frac{\epsilon}{2}) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}, E) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}, E^c) \\
\leq\ & P(|\frac{\sum_{i=1}^{N} g(X_i)}{n}| > \frac{\epsilon}{2}) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}, E^c) \\
+\ & P(|\frac{\sum_{i=N+1}^{n} g(X_i^{'})}{n}| > \frac{\epsilon}{2}, E) + \sum_{i=N+1}^{n} P(X_i \neq X_i^{'}, E) \\
\leq\ & \epsilon.
\end{aligned}$$

□

Remark: According to the conditions in the above proposition, we know that when N is large enough, the sequence $\{X_n\}_{n=N}^{\infty}$ is almost equal to $\{X_n^{'}\}_{n=N}^{\infty}$ which is a Markov chain with transition kernel P_{Γ_n}. At the first sight, adaptive algorithms that satisfy the conditions (a) and $(b^{'})$ cannot show the adaptive MCMC's advantages sufficiently. But following Roberts and Rosenthal [47] (2005), we know that in lots of cases, the adaptive MCMC will tune the parameter to an "optimal" one after "learning" the information from the historical samples. So we can adjust the convergence speed of S_n such that the adaptive chain can learn enough to find the optimal parameter, that is we can make N very large, such that Γ_N is almost a "good" parameter.

6.4 The WLLN For Adaptive Metropolis-Hastings Algorithm

Usually we construct the transition kernel using Metropolis-Hastings algorithms. If we tune the proposal distribution at each step as Harrio eg did in [26], we hope to prove the WLLN for unbounded function with respect to adaptive Metropolis-Hasting algorithm. Furthermore, when the proposal kernels have uniformly bounded densities, Roberts and Rosenthal [48] (2005) have proved the following ergodicity corollary with respect to adaptive Metropolis-Hastings algorithm.

Corollary 6.1. *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergordic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, P_γ represents a Metropolis-Hastings algorithm with proposal kernel $Q_\gamma(x, dy) = f_\gamma(x, y)\lambda(dy)$ having a density $f_\gamma(x, y)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$, with corresponding density w for $\pi(\cdot)$ so that $\pi(dy) = w(y)\lambda(dy)$. Finally, suppose $f_\gamma(x, y)$ are uniformly bounded, and that for each fixed $y \in \mathcal{X}$, the mapping $(x, \gamma) \mapsto f_\gamma(x, y)$ is continuous with respect to some product metric space topology, with*

respect to which $\mathcal{X} \times \mathcal{Y}$ *is compact. Then* $\lim_{n\to\infty} T(x, \gamma, n) = 0$ *for all* $x \in \mathcal{X}$ *and* $\gamma \in \mathcal{Y}$.

If we denote the conditions in corollary 6.1 by condition (j), then we have the following theorem:

Theorem 6.3. *Consider an adaptive MCMC that satisfies the condition (j). Then for any measurable function* g *such that* $\lambda(|g|) < \infty$ *and* $\pi(|g|) < \infty$ *we have:*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$, *conditional on* $X_0 = x_*$ *and* $\Gamma_0 = \gamma_*$.

Remark: If there exist $M > m > 0$ such that $m < w(x) < M$, where $\pi(dy) = w(y)\lambda(dy)$, then we know that $\lambda(|g|) < \infty$ if and only if $\pi(|g|) < \infty$. A typical case is that the state space $\mathcal{X}$ is compact set in R^d, $w(y)$ is continuous function on $\mathcal{X}$ and λ is Lebesgue measure. Then we have $M > w(x) > m > 0$, and the WLLN of the adaptive MCMC satisfying the conditions in corollary 6.1 will hold for any measurable function g such that $\pi(|g|) < \infty$.

We will prove the theorem following the steps below:

Step 1: For all $M > 0$, denote $E_M = \{x \in \mathcal{X} || g(x)| \leq M\}$ and for all ε define:

$$\begin{aligned} M_\varepsilon &= inf\{M > 0 | \lambda(E_M) \geq 1 - \varepsilon, \int_{E_M} |g(x)|\lambda(dx) \geq s - \varepsilon\} \\ &= inf\{M > 0 | \lambda(E_M^c) \leq \varepsilon, \int_{E_M^c} |g(x)|\lambda(dx) \leq \varepsilon\}. \end{aligned}$$

If $\lambda(|g|) < \infty$, we will prove that $\varepsilon \cdot M_\varepsilon \to 0$ as $\varepsilon \to 0$;

Step 2: Suppose $P_\gamma(x, A) = \int_A \tilde{f}_\gamma(x, y)\lambda(dx) + r_\gamma(x)\delta_x(A)$ then Under the conditions of the Theorem 6.3 we have $0 < r_\gamma(x) < \eta$;

Step 3: Suppose $A_\gamma^n(x, A) = P(X_n \in A | X_0 = 0, \Gamma_0 = \gamma)$, then there exist $L > 0$ and $0 < \eta < 1$, then under the conditions of the theorem, we have

$$A_\gamma^n(x, B) = \int_B h_\gamma^{(n)}(x, y)\lambda(dy) + w_\gamma^{(n)}(x)\delta_x(B),$$

such that $h_\gamma^{(n)}(x, y) < L$ and $w_\gamma^{(n)}(x) < \eta^n$;

Step 4: Prove the WLLN using coupling methods.

6.4.1 Some Technical Results

Suppose the probability of accepting a proposal y generated from x according to Q_γ is given by $\alpha_\gamma(x,y) = \min\{1, \frac{g(y)f_\gamma(y,x)}{g(x)f_\gamma(x,y)}\}$, so we have:

$$P_\gamma(x,B) = \int_B f_\gamma(x,y)\alpha_\gamma(x,y)\lambda(dy) + (1 - \int_{\mathcal{X}} \alpha_\gamma(x,y)\lambda(dy))\delta_x(B).$$

We can denote $\tilde{f}_\gamma(x,y) = f_\gamma(x,y)\alpha_\gamma(x,y)$, $r_\gamma(x) = (1 - \int_{\mathcal{X}} \alpha_\gamma(x,y)\lambda(dy))$ and suppose $f_\gamma(x,y) < F$. Obviously we have $\tilde{f}_\gamma(x,y) < F$ since $\alpha_\gamma(x,y) \leq 1$. We also need to prove the following lemmas before we prove the theorem.

Lemma 6.5. *Suppose $(\chi, \mathfrak{F}, \lambda)$ is a probability space, and $g : \chi \to R$ is a measurable function such that $\lambda(|g|) = s < \infty$. Then for $\forall \varepsilon > 0$, there exists $M > 0$, such that: $\lambda(E_M) \geq 1 - \varepsilon$ and $\int_{E_M} |g(x)|\lambda(dx) \geq s - \varepsilon$*

Proof. Suppose there exits $\varepsilon_0 > 0$, for each M, we have

$$\lambda(E_M^c) \geq \varepsilon_0 \tag{6.3}$$

or

$$\int_{E_M} |g(x)|\lambda(dx) \leq s - \varepsilon_0 \tag{6.4}$$

If (6.3) holds, we have $\int_{E_M^c} |g(x)|\pi(dx) \geq M\varepsilon_0$ for all M, contradiction!
If (6.4) holds, we have $\int_\chi |g(x)|1_{E_n}(x)\pi(dx) \leq s - \varepsilon_0$ for all $n \in \mathbb{N}$. Suppose

$$Y_n = |g(X)|1_{E_n}(x).$$

Obviously $Y_n \uparrow |g(X)|$, then by the monotone convergence theorem

$$E_\lambda(|g(x)|) = lim_{n\to\infty} E(Y_n) \leq s - \varepsilon_0,$$

which is contradicting with $E_\lambda(|g(x)|) = s$. □

Lemma 6.6. *Suppose $g : \chi \to R$ is a measurable function such that $\lambda(|g|) = s < \infty$. Then for each sequence $\{\varepsilon_n\} \to 0$, there exists a subsequence $\varepsilon_{n_k} \searrow 0$ such that $\varepsilon_{n_k} M_{\varepsilon_{n_k}} \to 0$ as $n \to 0$.*

Proof. Following lemma 6.5 we know that $0 \leq \frac{\lambda(E^c_{M_{\varepsilon_n}})}{\varepsilon_n} \leq 1$, there is a subsequence $\varepsilon_{n_k} \searrow 0$ such that $\{\frac{\lambda(E^c_{M_{\varepsilon_{n_k}}})}{\varepsilon_{n_k}}\}$ is convergent to some a. Then we can think about the problem in the following two cases:

(1). $0 < a \leq 1$; then there exists $N > 0$ such that for each $k > N$, $|\frac{\lambda(E^c_{M_{\varepsilon_{n_k}}})}{\varepsilon_{n_k}} - a| < \frac{a}{2}$, i.e. $\lambda(E^c_{M_{\varepsilon_{n_k}}}) > \frac{a}{2}\varepsilon_{n_k}$, so

$$\begin{aligned} 0 &= lim_{k\to\infty} \int_{E_{M^c_{\varepsilon_{n_k}}}} |g(x)|\pi(dx) \\ &\geq lim_{k\to\infty} \lambda(E^c_{M_{\varepsilon_n}}) M_{\varepsilon_{n_k}} \\ &\geq lim_{k\to\infty} \frac{a}{2}\varepsilon_{n_k} M_{\varepsilon_{n_k}} \\ &\geq 0. \end{aligned}$$

So $lim_{k\to\infty}\varepsilon_{n_k} M_{\varepsilon_{n_k}} = 0$.

(2). $a = 0$; then there exists N, $k > N$, such that

$$\lambda(E^c_{M_{\varepsilon_{n_k}}}) < \frac{1}{2}\varepsilon_{n_k}. \tag{6.5}$$

And following (6.3) for each $\delta > 0$,

$$\lambda(|g(x)| \geq M_{\varepsilon_{n_k}} - \delta) > \varepsilon_{n_k}. \tag{6.6}$$

Following (6.5) and (6.6), let $\delta \to 0$, we can get:

$$\begin{aligned} \lambda(|g(x)| = M_{\varepsilon_{n_k}}) &\geq \varepsilon_{n_k} - \frac{1}{2}\varepsilon_{n_k} \\ &= \frac{1}{2}\varepsilon_{n_k}. \end{aligned}$$

Since $\varepsilon_{n_{k+1}} < \varepsilon_{n_k}$, $M_{\varepsilon_{n_k}} \leq M_{\varepsilon_{n_{k+1}}}$,

$$\begin{aligned} 0 &= lim_{k\to\infty} \int_{E^c_{M_{\varepsilon_{n_k}}}} |g(x)|\lambda(dx) \\ &\geq lim_{k\to\infty} \int_{\{|g(x)|=M_{\varepsilon_{n+1}}\}} |g(x)|\lambda(dx) \\ &= lim_{k\to\infty} M_{\varepsilon_{n_{k+1}}} \cdot \lambda(|g(x)| = M_{\varepsilon_{n_{k+1}}}) \\ &\geq lim_{k\to\infty} \frac{1}{2} M_{\varepsilon_{k_{n+1}}} \cdot \varepsilon_{n_{k+1}} \\ &\geq 0. \end{aligned}$$

So $lim_{k\to\infty} M_{\varepsilon_{n_{k+1}}} \cdot \varepsilon_{n_{k+1}} = 0$. □

Lemma 6.7. *Suppose $g : \chi \to R$ is a measurable function such that $\lambda(|g|) = s < \infty$. Then $\varepsilon \cdot M_\varepsilon \to 0$ as $\varepsilon \to 0$.*

Proof. Suppose there exists $c > 0$ such that for each $n \in N$, there exists $\varepsilon_n < \frac{1}{n}$ and $\varepsilon_n \cdot M_{\varepsilon_n} \geq c$ for all n, then every subsequence $\{\varepsilon_{n_k}\}$ of $\{\varepsilon_n\}$ satisfies that $\varepsilon_{n_k} \cdot M_{\varepsilon_{n_k}} \geq c$, which is contradicting with the lemma 6.6. So $\varepsilon \cdot M_\varepsilon \to 0$ as $\varepsilon \to 0$. □

Lemma 6.8. *Under the conditions of corollary 6.1, we have that condition (a) holds.*

Proof. Following the proof of Corollary 12 in Roberts and Rosenthal [48](2005), we can get the lemma directly. □

Lemma 6.9. *Condition (a) is equivalent to: There exist $M > 0$ and $0 < \rho < 1$ such that for any x, γ we have:*

$$\|P^n_\gamma(x, \cdot) - \pi(\cdot)\| \leq M\rho^n.$$

Proof. Suppose $t_\gamma(n) = 2\sup_{x\in\mathcal{X}} \|P^n_\gamma(x, \cdot) - \pi(\cdot)\|$, following Roberts and Rosenthal [46] (2004) Proposition 3(c), we know that $t_\gamma(m + n) \leq t_\gamma(m)t_\gamma(n)$. Under condition (a), there exists n which is independent of γ such that $t_\gamma(n) \equiv \beta < 1$, so for all $j \in \mathbf{N}$,

$t_\gamma(jn) \le (t_\gamma(n))^j = \beta^j$. Therefore, we have:

$$\|P_\gamma^m(x,\cdot) - \pi(\cdot)\| \le \|P_\gamma^{\lfloor m/n \rfloor n}(x,\cdot) - \pi(\cdot)\| \le \frac{1}{2} t_\gamma(\lfloor m/n \rfloor n) \le \beta^{\lfloor m/n \rfloor} \le \beta^{-1}(\beta^{1/n})^m.$$

So all the kernels are uniformly ergodic with $M = \beta^{-1}$ and $\rho = \beta^{1/n}$. □

Lemma 6.10. *Suppose $P_\gamma(x,A) = \int_A \tilde{f}_\gamma(x,y)\lambda(dx) + r_\gamma(x)\delta_x(A)$, then there exist measurable functions $\tilde{f}_\gamma^{(n)}(x,y)$ on χ^2 such that $P_\gamma^n(x,A) = \int_A \tilde{f}_\gamma^{(n)}(x,y)\lambda(dx) + r_\gamma^n(x)\delta_x(A)$.*

Proof. We will prove it by induction, and obviously the conclusion holds when $n = 1$. We suppose it also holds when $n = k$, then let's consider the case when $n = k+1$:

$$\begin{aligned}
P_\gamma^{k+1}(x,A) &= \int_\chi P_\gamma^k(y,A) P_\gamma(x,dy) \\
&= \int_\chi [\int_A \tilde{f}_\gamma^{(k)}(y,z)\lambda(dz) + r_\gamma^k(y)\delta_y(A)][\tilde{f}_\gamma(x,y)\pi(dy) + r_\gamma(x)\delta_x(dy)] \\
&= \int_\chi \int_A \tilde{f}_\gamma^{(k)}(y,z)\pi(dz) f_\gamma(x,y)\lambda(dy) + \tilde{f}_\gamma^{(k)}(y,z)\pi(dz) r_\gamma(x)\delta_x(dy) \\
&+ r_\gamma^k(y)\delta_y(A)\tilde{f}_\gamma(x,y)\lambda(dy) + r_\gamma^k(y)\delta_y(A) r_\gamma(x)\delta_x(dy) \\
&= \int_A [\int_\chi \tilde{f}_\gamma^{(k)}(y,z) f_\gamma(x,y)\lambda(dy)]\pi(dz) + \int_A r_\gamma(x)\tilde{f}_\gamma^k(x,z)\pi(dz) \\
&+ \int_A r_\gamma^k(y)\tilde{f}_\gamma(x,y)\lambda(dy) + r_\gamma^{k+1}(x)\delta_x(A) \\
&= \int_A \tilde{f}_\gamma^{(k+1)}(x,z)\lambda(dz) + r_\gamma^{k+1}(x)\delta_x(A),
\end{aligned}$$

where

$$\tilde{f}_\gamma^{(k+1)}(x,z) = \int_\chi \tilde{f}_\gamma^{(k)}(y,z)\tilde{f}_\gamma(x,y)\pi(dy) + r_\gamma(x)\tilde{f}_\gamma^k(x,z) + r_\gamma^k(x)\tilde{f}_\gamma(x,z). \quad (6.7)$$

□

Lemma 6.11. *Suppose $P_\gamma(x,A) = \int_A \tilde{f}_\gamma(x,y)\lambda(dx) + r_\gamma(x)\delta_x(A)$ where $\lambda(\cdot)$ is a finite reference measure on $\mathcal{X}$ such that $\lambda(\{x\}) = 0$ for any x, with corresponding density w for $\pi(\cdot)$ so that $\pi(dy) = w(y)\lambda(dy)$. Then under condition (a),we have $0 < r_\gamma(x) < \eta$, where the η is the same as in Lemma 6.9.*

Proof. Because $P_\gamma(x, \{x\}^c) = \int_{\chi - x} \tilde{f}_\gamma(x, y)\pi(dx)$, $P_\gamma(x, x) = r_\gamma(x)$ and $\pi(x) = 0$, and following that $P_\gamma^{k+1}(x, A) = \int_A \tilde{f}_\gamma^{(k+1)}(x, z)\lambda(dz) + r_\gamma^{k+1}(x)\delta_x(A)$, we know that $|P_\gamma^n(x, \{x\}) - \pi(\{x\})| = r_\gamma^n(x)$ for each $x \in \chi$. Then following condition (a), we know for $\forall \epsilon > 0$, there exists N such that $r_\gamma^N(x) < \epsilon$, that is $r_\gamma(x) < \epsilon^{\frac{1}{N}}$ for each γ and x. Then we take $\epsilon < 1$, and we can get $\eta = \epsilon^{\frac{1}{N}} < 1$. □

Lemma 6.12. *Suppose $A_\gamma^n(x, A) = P(X_n \in A | X_0 = 0, \Gamma_0 = \gamma)$, then under the conditions of corollary 6.1, there exist $L > 0$ and $0 < \eta < 1$, such that*

$$A_\gamma^n(x, B) = \int_B h_\gamma^{(n)}(x, y)\lambda(dy) + w_\gamma^{(n)}(x)\delta_x(B),$$

where $h_\gamma^{(n)}(x, y) < L$ and $w_\gamma^{(n)}(x) < \eta^n$.

Proof. Suppose the joint distribution of $(X_1, X_2, \cdots, X_n, \Gamma_1, \Gamma_2, \cdots, \Gamma_{n-1})$ given $X_0 = x$ and $\Gamma_0 = \gamma$ is $\mu_{(x,\gamma)}^{(n)}$, obviously the marginal distribution of X_n is $A^{(n)}((x, \gamma), \cdot)$. Since γ_n is a measurable function of $(x_1, x_2, \cdots, x_n, \gamma_1, \gamma_2, \cdots, \gamma_{n-1})$, we have:

$$\begin{aligned}
A^{(n+1)}((x, \gamma), B) &= \int_{\mathcal{X}^n \times \mathcal{Y}^{n-1}} P_{\Gamma_n}(x_n, B)\mu_{(x,\gamma)}^{(n)}(dx_1 \cdots dx_n d\gamma_1 \cdots d\gamma_{n-1}) \\
&= \int_{\mathcal{X}^n \times \mathcal{Y}^{n-1}} [\int_B \tilde{f}_{\gamma_n}(x_n, y)\lambda(dy) + r_{\gamma_n}(x_n)(\delta_{x_n}(B))]\mu_{(x,\gamma)}^{(n)}(dx_1 \cdot \cdot dx_n d\gamma_1 \cdot \cdot d\gamma_{n-1}) \\
&= \int_B \int_{\mathcal{X}^n \times \mathcal{Y}^{n-1}} \tilde{f}_{\gamma_n}(x_n, y)\mu_{(x,\gamma)}^{(n)}(dx_1 \cdots dx_n d\gamma_1 \cdots d\gamma_{n-1})\lambda(dy) \\
&+ \int_{\mathcal{X}^n \times \mathcal{Y}^{n-1}} r_{\gamma_n}(x_n)\delta_{x_n}(B)\mu_{(x,\gamma)}^{(n)}(dx_1 \cdots dx_n d\gamma_1 \cdots d\gamma_{n-1}).
\end{aligned}$$

We can observe that the second term:

$$\begin{aligned}
& \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} r_{\gamma_n}(x_n)\delta_{x_n}(B)\mu^{(n)}_{(x,\gamma)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1}) \\
= & \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{\mathcal{X}} r_{\gamma_n}(x_n)\delta_{x_n}(B)P_{\gamma_{n-1}}(x_{n-1},dx_n)\mu^{(n-1)}_{(x,\gamma)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-1}) \\
= & \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)P_{\gamma_{n-1}}(x_{n-1},dx_n)\mu^{(n-1)}_{(x,\gamma)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
= & \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)\tilde{f}_{\gamma_{n-1}}(x_{n-1},x_n)\lambda(dx_n)\mu^{(n-1)}_{(x,\gamma)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
+ & \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)r_{\gamma_{n-1}}(x_{n-1})\delta_{x_{n-1}}(dx_n)\mu^{(n-1)}_{(x,\gamma)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}).
\end{aligned}$$

If $\gamma_n = \gamma_n(x, x_1, \cdots, x_n, \gamma, \gamma_1, \cdots, \gamma_{n-1})$, then we can define:

$$\gamma_n^i = \gamma_n(x, x_1, \cdots, x_{n-i-1}, x_{n-i}, x_{n-i}, \cdots, x_{n-i}, \gamma, \gamma_1, \cdots, \gamma^1_{n-i+1}, \cdots, \gamma^{i-1}_{n-1}).$$

Similarly we can compute the second term of the above inequality:

$$\begin{aligned}
& \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)r_{\gamma_{n-1}}(x_{n-1})\delta_{x_{n-1}}(dx_n)\mu^{(n-1)}_{(x,\gamma)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
= & \int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma^1_n}(x_{n-1})r_{\gamma_{n-1}}(x_{n-1})P_{\gamma_{n-2}}(x_{n-2},dx_{n-1})\mu^{(n-2)}_{(x,\gamma)}(dx_1\cdots dx_{n-2}d\gamma_1\cdots d\gamma_{n-3}) \\
= & \int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma^1_n}(x_{n-1})r_{\gamma_{n-1}}(x_{n-1})\tilde{f}_{\gamma_{n-2}}(x_{n-2},x_{n-1})\lambda(dx_{n-1})\mu^{(n-2)}_{(x,\gamma)}(dx_1\cdot\cdot dx_{n-2}d\gamma_1\cdot\cdot d\gamma_{n-3}) \\
+ & \int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma^1_n}(x_{n-1})r_{\gamma_{n-1}}(x_{n-1})r_{\gamma_{n-2}}(x_{n-2})\delta_{x_{n-2}}(dx_{n-1})\mu^{(n-2)}_{(x,\gamma)}(dx_1\cdots dx_{n-2}d\gamma_1\cdots d\gamma_{n-3}).
\end{aligned}$$

Inductively we have:

$$\begin{aligned}
h_{\gamma}^{(n+1)}(x,y) &= \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} \tilde{f}_{\gamma_n}(x_n,y)\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1}) \\
&+ \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-2}} r_{\gamma_n}(x_n)\tilde{f}_{\gamma_{n-1}}(x_{n-1},x_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
&+ \int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-3}}\int_B \prod_{i=0}^{1} r_{\gamma_{n-i}^i}(x_{n-i})\tilde{f}_{\gamma_{n-2}}(x_{n-2},x_{n-1})\mu_{(x,\gamma)}^{(n-2)}(dx\cdot\cdot dx_{n-2}d\gamma_1\cdot\cdot d\gamma_{n-3}) \\
&+ \cdots \\
&+ \int_B \prod_{i=0}^{n-1} r_{\gamma_{n-i}^i}(x_1)\tilde{f}_{\gamma}(x,x_1)\mu_{(x,\gamma)}^{1}(dx_1) \\
&\leq F\sum_{i=0}^{n-1}\eta^i \\
&\leq \frac{F}{1-\eta}
\end{aligned}$$

and

$$\begin{aligned}
w_{\gamma}^{(n+1)}(x) &= \prod_{i=0}^{n-1} r_{\gamma_{n-i}^{n-i}}(x) \\
&\leq \eta^n.
\end{aligned}$$

□

6.4.2 The proof of Theorem 6.2

Now we state the proof using the above lemmas as below:

Proof. Suppose $\pi(g) = 0$,$\lambda(|g|) = s$,$D_n = \sup_{x\in\chi}\|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}\|$ and $f_\gamma(x,y) < F$. Lemma 4.2 implies that given $\varepsilon > 0$, there exits $\eta_1 > 0$ such that $M_{\eta_1}\eta_1 < \epsilon$; denote $\eta_2 = \frac{\epsilon}{F}$, then we have:

$$\int_{E_{M_{\eta_2}}^c} |g(x)|\lambda(dx) \leq \varepsilon.$$

Following lemma 4.4, we can find $\eta < \min\{\eta_1,\eta_2\}$ such that $M_\eta\eta < \epsilon$ and

$$\int_{E_{M_{\eta}}^c} |g(x)|\lambda(dx) \leq \varepsilon.$$

Then we define $g_k(x) = g(x)\delta_{E_k}(x)$, Since $g_{M_\eta}(x)$ is a bounded measurable function, then we can find an integer N such that:

$$E_{\gamma,x}[|\frac{\sum_{i=1}^{N} g_{M_\eta}(X_i)}{N}|] < \epsilon, \quad x \in \mathcal{X} \quad \gamma \in \mathcal{Y}.$$

Denote $H_n = \{D_n \geq \frac{\eta}{N^2}\}$, then Diminishing Adaptive condition implies that we can find $N_1 \in N$ such that for each $n > N_1$, $P(H_n) \leq \frac{\eta}{N}$ and $\frac{|g(x_*)|\eta^{N_1}}{N(1-\eta)} < \epsilon$. Define the event $E = \bigcap_{i=n+1}^{n+N} H_i^c$. Then when $n > N_1$, we have $P(E^c) < \eta$.For all $n \geq N_1$, following the triangle inequality and induction, on event E we have:

$$\sup_{x\in\chi} \|P_{\Gamma_{n+k}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \leq \eta/N, k \leq N.$$

In particular, for all $x \in \chi$ and $k - N \leq m \leq k$

$$\|P_{\Gamma_{k-N}}(x,\cdot) - P_{\Gamma_m}(x,\cdot)\| \leq \eta, on \quad E.$$

So $\|P^N_{\Gamma_{k-N}}(x,\cdot) - P(X_k \in \cdot | X_{k-N} = x, G_{k-N})\| \leq \eta \quad on \quad E$ for all $x \in \chi$. Then we can construct a second chain $\{X_n^{'}\}_{n=k-N}^{k}$ such that $X_{k-N}^{'} = X_{k-N}$ and $X_n^{'} \sim P_{\Gamma_{k-N}}(X_{n-1}^{'}, \cdot)$ for $k - N + 1 \leq n \leq k$ such that $P(X_k^{'} \neq X_k) \leq \eta$. So for any $n > N_1$, we have the following inequality (*):

$$\begin{aligned}
& E(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|X_0 = x_*, \Gamma_0 = \gamma_*) \\
\leq\ & E(E(|\frac{\sum_{i=n+1}^{n+N} g_{M_\eta}(X_i)}{N}||\mathcal{G}_n)|X_0 = x, \Gamma_0 = \gamma) + E(\frac{|\sum_{i=n+1}^{n+N}(g - g_{M_\eta})(X_i)}{N}||X_0, \Gamma_0) \\
\leq\ & E(E_{\Gamma_n, X_n}(|\frac{\sum_{i=1}^{N} g_{M_\eta}(X_i)}{N}|)|X_0, \Gamma_0) + M_\eta\eta + M_\eta P(E^c) + \frac{\sum_{i=n+1}^{n+N} E(|(g - g_\eta)(X_i)||X_0, \Gamma_0)}{N} \\
\leq\ & \epsilon + \epsilon + M_\eta\eta + \frac{\sum_{i=n+1}^{n+N} \int_{E^c_{M_\eta}} |g|(y)|A^{(i)}((x_*, \gamma_*), dy)}{N} \\
\leq\ & \epsilon + \epsilon + \epsilon + \frac{\sum_{i=n+1}^{n+N} \int_{E^c_{M_\eta}} |g|(y)|h^{(i)}_{\gamma_*}(x_*, y)\lambda(dy) + w^{(i)}_{\gamma_*}(x_*)|g(x_*)|}{N} \\
\leq\ & 3\epsilon + \frac{\sum_{i=n+1}^{n+N} L\int_{E^c_{M_\eta}} |g|(y)|\lambda(dy) + \eta^i|g(x_*)|}{N} \\
\leq\ & (3 + L)\epsilon + \frac{|g(x_*)|\eta^{n+1}}{N(1-\eta)}. \\
\leq\ & (4 + L)\epsilon \qquad (*)
\end{aligned}$$

Now consider any integer T sufficiently large such that:

$$max[\frac{N_1 Fs + \frac{|g(x_*)|}{1-\eta}}{T}, \frac{NFs + \frac{|g(x_*)|}{1-\eta}}{T}] \leq \varepsilon. \tag{6.8}$$

Then we have

$$\begin{aligned}
& E(|\frac{\sum_{i=1}^{T} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \\
\leq\ & E(|\frac{\sum_{i=1}^{N_1} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \\
+\ & E(\frac{1}{\lfloor \frac{T-N_1}{N} \rfloor} \sum_{j=1}^{\lfloor \frac{T-N_1}{N} \rfloor} \frac{1}{N} \sum_{k=1}^{N} g(X_{N_1+(j-1)N+k}|X_0 = x_*, \Gamma_0 = \gamma_*)) \\
+\ & E(|\frac{\sum_{N_1+\lfloor \frac{T-N_1}{N} \rfloor N+1}^{T} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*).
\end{aligned}$$

For the first term we have:

$$\begin{aligned}
& E(|\frac{\sum_{i=1}^{N_1} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \\
\leq\ & \frac{\sum_{i=1}^{N_1} E(|g(X_i)||X_0 = x_*, \Gamma_0 = \gamma_*)}{T} \\
\leq\ & \frac{\sum_{i=1}^{N_1} \int_{\mathcal{X}} |g(y)| A^{(n)}((x_*, \gamma_*), dy)}{T} \\
\leq\ & \frac{\sum_{i=1}^{N_1} \int_{\mathcal{X}} |g(y)| h_{\gamma}^{(n)}(x_*, y)\lambda(dy) + |g(x_*)|\eta^i}{T} \\
\leq\ & \frac{N_1 Fs + \frac{|g(x_*)|}{1-\eta}}{T} \\
\leq\ & \epsilon,
\end{aligned}$$

and for the third one we know that:

$$\begin{aligned}
& E(|\frac{\sum_{N_1+\lfloor \frac{T-N^*}{N} \rfloor N+1}^{T} g(X_i)}{T}|) \\
\leq\ & \frac{\sum_{N^*+\lfloor \frac{T-N^*}{N} \rfloor N+1}^{T} E(|g(X_i)|}{T} \\
\leq\ & \frac{NFs + \frac{|g(x_*)|}{1-\eta}}{T} \\
\leq\ & \epsilon.
\end{aligned}$$

Finally following the inequality (*), the second term $\leq (4+L)\epsilon$, so we have

$$E(|\frac{\sum_{i=1}^{T} g(X_i)}{T}|) \quad \leq \quad (6+L)\epsilon.$$

Markov's inequality then gives that

$$P(|T^{-1}\sum_{i=1}^{T} g(X_i)| \geq \varepsilon^{\frac{1}{2}}) \leq (6+L)\epsilon^{\frac{1}{2}}.$$

Since this holds for all sufficiently large T, and since $\varepsilon > 0$ is arbitrary, the result follows. □

Remark: Here we actually get the conclusion: for any $\epsilon > 0$, $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, there exists N such that for any $n > N$ we have:

$$P(|\frac{\sum_{i=1}^{n} g(X_i)}{n}| > \epsilon) < \epsilon.$$

But here the "N" is **dependent** on the choice of the starting value x, but **independent** of the starting value γ. In fact, this kind of dependence of the starting value is reasonable when g is unbounded. Let us consider the following example which is a general Markov chain with the kernel being uniformly ergodic:

Consider $\mathcal{X} = (0, 1]$, and

$$P(x, A) = \frac{2}{3}\mu(A) + \frac{1}{3}\delta_x(A),$$

where μ is Lebesgue measure on $(0, 1]$. Since

$$\begin{aligned} \int_{\mathcal{X}} P(x, A)\mu(dx) &= \int_{\mathcal{X}} [\frac{2}{3}\mu(A) + \frac{1}{3}\delta_x(A)]\mu(dx) \\ &= \frac{2}{3}\mu(A) + \frac{1}{3}\mu(A) \\ &= \mu(A), \end{aligned}$$

π is stationary with respect to $P(x, \cdot)$. And following that:

$$\|P(x, \cdot) - \pi(\cdot)\|_{var} \quad = \quad \| - \frac{1}{3}\mu(A) + \frac{1}{3}\delta_x(A)\|_{var} \leq \frac{1}{3}.$$

Therefore, P is uniformly ergodic with respect to μ. Now suppose $g(x) = x^{-\frac{1}{2}}$, then $\mu(g) = 2$, and then $P(X_1 \in (0, \frac{1}{m^2}]|X_0 = \frac{1}{m^2}) = \frac{2}{3m^2} + \frac{1}{3}$ for each $m \in \mathbf{N}$. Suppose for

some $0 < \epsilon < \frac{1}{3}$, there exists N such that $P(|\frac{\sum_{i=1}^{N} g(X_i)}{N}| > \epsilon | X_0 = x_0) < \epsilon$ for all $x_0 \in \mathcal{X}$. If we take $x_0 = (3N)^{-2}$, since $g(X_i) > 0$, we have:

$$\begin{aligned} P(|\frac{\sum_{i=1}^{N} g(X_i)}{N} - \pi(g)| > \epsilon | X_0 = \frac{1}{(3N)^2}) &\geq P(\frac{g(X_1)}{N} - 2 > \epsilon | X_0 = \frac{1}{(3N)^2}) \\ &\geq P(g(X_1) \geq 3N | X_0 = \frac{1}{(3N)^2}) \\ &\geq P(X_1 \leq \frac{1}{(3N)^2} | X_0 = \frac{1}{(3N)^2}) \\ &> \frac{1}{3}. \end{aligned}$$

Contradiction!

6.4.3 A Corollary

In Roberts and Rosenthal [48] (2007), they also studied the adaptive MCMC with bounded densities and proved the following corollary:

Corollary 6.2. *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergordic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, $P_\gamma(x, dy) = f_\gamma(x, y)\lambda(dy)$ has a density $f_\gamma(x, y)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$. Finally, suppose $f_\gamma(x, y)$ are uniformly bounded, and that for each fixed $y \in \mathcal{X}$, the mapping $(x, \gamma) \mapsto f_\gamma(x, y)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then $\lim_{n\to\infty} T(x, \gamma, n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.*

We also have the WLLN for the unbounded measurable function g under the same conditions in the corollary 6.2. Actually $P_\gamma(x, A) = \int_A f_\gamma(x, y)\lambda(dy)$ is a special case of $P_\gamma(x, A) = \int_A f_\gamma(x, y)\lambda(dy) + r_\gamma(x)\delta_x(A)$ when $r_\gamma(x) \equiv 0$. We just plug in $\eta = 0$ to the proof of the Theorem 6.3, then we can prove the following corollary:

Corollary 6.3. *Consider an adaptive MCMC that satisfies the conditions in Corollary 6.2, then for any measurable function g such that $\lambda(|g|) < \infty$ and $\pi(g) < \infty$ we have:*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$*, conditional on* $X_0 = x$ *and* $\Gamma_0 = \gamma$*.*

Remark: The Corollary 6.3 indicates that: for any $\epsilon > 0$, $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, there exists N such that for any $n > N$ we have:

$$P(|\frac{\sum_{i=1}^{n} g(X_i)}{n}| > \epsilon) < \epsilon.$$

However it is not hard to find that such an "N" is **independent** of the choice of the initial values x and γ.

6.4.4 Applications

As an application of the Theorem 6.3, we will think about the Adaptive Metropolis algorithm of Haario et al. [25](2001) , in which the target distribution π is supported on the subset $S \subseteq \mathbb{R}^d$ and it has the density π with a slight abuse of notation with respect to the Lebesgue measure on S.

Haario et al. [25] (2001) have prove the following Strong Laws of Large Number(SLLN):

Theorem 6.4. *Let* π *be the density of a target distribution supported on a bounded measurable subset* $S \subseteq \mathbb{R}^d$*, and assume that* π *is bounded from above. Let* $\epsilon > 0$ *and let* μ_0 *be any initial distribution on* S*. Define the adaptive MCMC as above. Then the AMCMC simulates properly the target distribution* π*: for any bounded and measurable function* $f : S \to R$*, the equality:*

$$\lim_{n\to\infty} \frac{1}{n+1}(f(X_0) + f(X_1) + \cdots + f(X_n)) = \int_S f(x)\pi(dx)$$

holds almost surely.

However following the Theorem 6.3, we actually can prove that the WLLN holds for any unbounded measurable function g with $\lambda(|g|) < \infty$ where λ is Lesbesgue measure.

Corollary 6.4. *The WLLN holds for the above adaptive MCMC and any measurable function* g *satisfying* $\lambda(|g|) < \infty$ *and* $\pi(|g|) < \infty$*.*

Proof. In this adaptive algorithm, according to the formula (14) in Haario et al. [25] (2001), the parameter space $\mathcal{Y}$ consists of all the $d \times d$ matrix γ satisfying that $c_1 I_d \leq \gamma \leq c_2 I_d$ for some $c_1 > 0$ and $c_2 > 0$. If we consider $\mathcal{Y}$ as a d^2 vector space and define the metric on it as $d(\gamma_1, \gamma_2) = \sqrt{\sum_{1 \leq i \leq j \leq d} \left((\gamma_1)_{ij} - (\gamma_2)_{ij} \right)^2}$. Obviously $\mathcal{Y}$ is compact with respect this metric topology, hence $\mathcal{X} \times \mathcal{Y}$ is also compact. Furthermore since the proposal distribution $Q_\gamma(x, \cdot) = MVN(x, \gamma)$, P_γ is ergodic for $\pi(\cdot)$ and the density mapping $(x, \gamma) \to f_\gamma(x, y)$ are continuous and bounded. Therefore following the Theorem 6.3 we have the conclusion. □

6.5 WLLN Under Conditions of Theorem 6.5

Here we will prove the WLLN of AMCMC for bounded function under the conditions of the Theorem 5.1.

Theorem 6.5. ***(WLLN)*** *Consider an adaptive MCMC algorithm. Suppose that the conditions of the Theorem 5.1 hold. Let $g : \mathcal{X} \to R$ be a bounded measurable function. Then for any starting values $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, conditional on $X_0 = x$ and $\Gamma_0 = \gamma$ we have*

$$\frac{\sum_{i=1}^n g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$.

Similar to the proof of theorem 5.1, it suffices to prove the following lemma before we prove the Theorem 6.5:

Lemma 6.13. *Under the conditions of lemma 5.2. Let $g : \mathcal{X} \to R$ be a bounded measurable function. Then for any starting values $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, conditional on $X_0 = x$ and $\Gamma_0 = \gamma$ we have*

$$\frac{\sum_{i=1}^n g(X_i)}{n} \to \pi(g)$$

in probability as $n \to \infty$.

6.5.1 Some Technical Results

Following the usual laws of large numbers for Markov chain (see e.g. Meyn and Tweedie [37]) imply that for each fixed $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(X_i^\gamma) \to \pi(g)$ in probability, where $\{X_n^\gamma\}$ is the usual Markow chain with kernel P_γ. Actually we will prove that under the conditions in lemma 5.2 the above convergence is uniformly with respect to the parameter γ. Before we start the proof, let us define some symbols, let

$$s_i^\gamma(g) = \sum_{j=\tau_{\check{\alpha}}(i)+1}^{\tau_{\check{\alpha}}(i+1)} g(X_j^\gamma),$$

and

$$l_n^\gamma = \max\{i \geq 0 : \tau_{\check{\alpha}}(i) \leq n\}.$$

Lemma 6.14. *Under the conditions of lemma 5.2, for any $\epsilon > 0$ and fixed start value x, there exists N which is independent with the choice of γ such that for any $n > N$ we have:*

$$P_x(|\frac{\sum_{i=1}^{n} g(X_i^\gamma)}{n} - \pi(g)| > \epsilon) < \epsilon W(x) + \epsilon.$$

Proof. Without losing generalities, we suppose $\pi(g) = 0$ and $|g(x)| \leq M$ then

$$\begin{aligned}
& P_x(|\frac{\sum_{i=1}^{n} g(X_i^\gamma)}{n}| > 3\epsilon) \\
= \; & P_x(|\frac{\sum_{i=1}^{\tau_{\check{\alpha}}} g(X_i^\gamma)}{n} + \frac{\sum_{i=0}^{l_n} s_i(g)}{n} + \frac{\sum_{i=\tau_{\check{\alpha}}(l_n)+1}^{n} g(X_i^\gamma)}{n}| > 3\epsilon) \\
\leq \; & P_x(|\frac{\sum_{i=1}^{\tau_{\check{\alpha}}} g(X_i^\gamma)}{n}| > \epsilon) + P_x(|\frac{\sum_{i=0}^{l_n} s_i^\gamma(g)}{n}| > \epsilon) + P_x(|\frac{\sum_{i=\tau_{\check{\alpha}}(l_n)+1}^{n} g(X_i^\gamma)}{n}| > \epsilon).
\end{aligned}$$

Regarding the first term we have:

$$\begin{aligned}
P_x(|\frac{\sum_{i=1}^{\tau_{\check{\alpha}}} g(X_i^\gamma)}{n}| > \epsilon) \; & \leq \; \frac{E_x[|\sum_{i=1}^{\tau_{\check{\alpha}}} g(X_i^\gamma)|]}{n\epsilon} \\
& \leq \; \frac{E_x[\tau_{\check{\alpha}}]M}{n\epsilon} \\
& \leq \; \frac{W(x)M}{n\epsilon}.
\end{aligned}$$

Regarding the third term we have:

$$\begin{aligned} P_x(|\frac{\sum_{i=\tau_{\check{\alpha}}(l_n)+1}^{n} g(X_i^{\gamma})}{n}| > \epsilon) &\leq \frac{E_{\check{\alpha}}[|\sum_{i=\tau_{\check{\alpha}}(l_n)+1}^{n} g(X_i^{\gamma})|]}{n\epsilon} \\ &\leq \frac{E_{\check{\alpha}}[\tau_{\check{\alpha}}]M}{n\epsilon} \\ &\leq \frac{KM}{n\epsilon}. \end{aligned}$$

Actually the second term is independent with the choice of start value x,i.e.

$$P_x(|\frac{\sum_{i=0}^{l_n} s_i^{\gamma}(g)}{n}| > \epsilon) = P_{\check{\alpha}}(|\frac{\sum_{i=0}^{l_n} s_i^{\gamma}(g)}{n}| > \epsilon).$$

Suppose for any $n \in \mathbb{N}$, there exists γ_n such that $P_{\check{\alpha}}(|\frac{\sum_{i=0}^{l_n} s_i^{\gamma_n}(g)}{n}| > \epsilon) > \frac{\epsilon}{2}$, same as the proof of lemma 5.6, we can find certain $\gamma_0 \in \Delta$ such that:

$$\lim_{n\to\infty} P_{\check{\alpha}}(|\frac{\sum_{i=0}^{l_n} s_i^{\gamma_n}(g)}{n}| > \epsilon) > \frac{\epsilon}{2},$$

Which is conflicting with the fact that for any $\gamma \in \Delta$ and $\epsilon > 0$, we have:

$$\lim_{n\to\infty} P_{\check{\alpha}}(|\frac{\sum_{i=0}^{l_n} s_i^{\gamma_n}(g)}{n}| > \epsilon) = \pi(g) = 0.$$

Therefore there exists N_1, such that for any $n > N_1$ and γ, we have:

$$P_x(|\frac{\sum_{i=0}^{l_n} s_i^{\gamma}(g)}{n}| > \epsilon) < \frac{\epsilon}{2}.$$

We also can find N_2 such that for any $n > N_2$ we have $\frac{M}{n} < \epsilon^2$ and $\frac{KM}{n} < \frac{\epsilon^2}{2}$. Then let $N = \max\{N_1, N_2\}$ we can get the conclusion. □

Lemma 6.15. *Given $\epsilon > 0$, we can find $N > 0$ such that when $n > N$ we have:*

$$E_{\gamma,x}[|\frac{\sum_{i=1}^{N} g(X_i)}{N}|] \leq \epsilon W(x) + \epsilon.$$

Proof. Following Lemma 6.14, we know that for any $\epsilon > 0$, there exists N such that:

$$P_x(|\frac{\sum_{i=1}^{n} g(X_i^{\gamma})}{n}| > \epsilon) < \frac{\epsilon}{M} W(x) + \frac{\epsilon}{2M}$$

We also have $|\frac{\sum_{i=1}^{n} g(X_i^{\gamma})}{n}| \leq M$. If we denote $\Lambda = \{\omega \in \Omega || \frac{\sum_{i=1}^{n} g(X_i^{\gamma})}{n}| > \frac{\epsilon}{2} \text{ given } X_0 = x\}$. Then we have:

$$\begin{aligned} E_{\gamma,x}[|\frac{\sum_{i=1}^{N} g(X_i)}{N}|] &= E_{\gamma,x}[|\frac{\sum_{i=1}^{N} g(X_i)}{N}| \times \mathbb{I}_{\omega}(\Lambda)] + E_{\gamma,x}[|\frac{\sum_{i=1}^{N} g(X_i)}{N}| \times \mathbb{I}_{\omega}(\Lambda^c)] \\ &\leq M[W(x)\frac{\epsilon}{M} + \frac{\epsilon}{2M}] + \frac{\epsilon}{2} \\ &\leq \epsilon W(x) + \epsilon. \end{aligned}$$

□

6.5.2 The Proof Of Theorem 6.5

First we can prove the Lemma 6.13:

Proof. Given starting value $X_0 = x$, $\Gamma_0 = \gamma$ and $\epsilon > 0$, $W(X_n)$ is bounded in probability, i.e. for any $\epsilon > 0$, there exists $a > 0$ such that:

$$P(W(X_n) > a) < \frac{\epsilon}{4M} \text{ for all } n \in \mathbb{N}.$$

Following the Lemma 6.15, we know that there exists $N = N(\epsilon)$, such that for any x and γ we have:

$$E_{\gamma,x}[|\frac{\sum_{i=1}^{N} g(X_i)}{N}|] \leq \frac{\epsilon W(x)}{4a} + \frac{\epsilon}{4}.$$

Then let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and $H_n = D_n \geq \frac{\epsilon}{4MN^2}$. Using the Diminishing Adaptation condition to choose $n^* = n^*(\epsilon) \in \mathbb{N}$ large enough so that

$$P(H_n) \leq \frac{\epsilon}{4NM}, \quad n \leq n^*.$$

To continue, fix a "target time" $K \geq n^* + N$. We shall construct a coupling which depends on the target time K (cf. Roberts and Rosenthal [45], 2002), to prove that $\mathcal{L}(X_k) \approx \pi(\cdot)$.

Define the event $E = \cap_{i=n+1}^{n+N} H_i^c$, we have $P(E) \geq 1 - \frac{\epsilon}{4M}$. Now, it follows from the triangle inequality and induction that on the event E, we have:

$$\sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+k}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| < \frac{\epsilon}{4MN}, \quad k \leq N.$$

In particular, on E we have $\|P_{\Gamma_{L-N}}(x,\cdot) - P_{\Gamma_m}(x,\cdot)\| < \frac{\epsilon}{4MN}$ for all $x \in \mathcal{X}$ and $L-N \leq m \leq L$, so by induction again,

$$\|P^N_{\Gamma_{L-N}}(x,\cdot) - P_{\Gamma_n}(X_k \in \cdot | X_{L-N} = x, \mathcal{G}_{L-N})\| < \frac{\epsilon}{4M} \text{ on E, for } x \in \mathcal{X}.$$

To construct the coupling, first construct the original adaptive chain $\{X_n\}$ together with its adaption sequence $\{\Gamma_n\}$, starting with $X_0 = x$ and $\Gamma_0 = \gamma$.

We now claim that on E, we can construct a second chain $\{X'_n\}_{n=L-N}^{L}$ such that $X'_{L-N} = X_{L-N}$ and $X'_n \sim P_{\Gamma_{L-N}}(X'_{n-1}, \cdot)$ for $L-N+1 \leq n \leq L$, and such that $P(X'_L \neq X_L) < \epsilon$. Indeed, conditional on $\mathcal{G}_{L-N}$, we have $X'_L \tilde{P}^N_{\Gamma_{L-N}}(X_{L-N}, \cdot)$. Then we have:

$$\|\mathcal{L}(X'_k) - \mathcal{L}(X_k)\| < \frac{\epsilon}{4M}.$$

The claim then follows from e.g. Roberts and Rosenthal [46](2004, Proposition 3(g)).

Since $|g| \leq M$, we have:

$$\begin{aligned} E\left(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|\mathcal{G}_n\right) &\leq E_{\Gamma_n, X_n}\left(\frac{1}{N}|\sum_{i=1}^{N} g(X_i)|\right) + M\frac{\epsilon}{4M} + MP(E^c) \\ &\leq \frac{\epsilon W(X_n)}{4a} + \frac{\epsilon}{2}, \end{aligned}$$

and we also have:

$$E\left(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|\mathcal{G}_n\right) \leq M.$$

Therefore,

$$\begin{aligned} & E((\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)) \\ = \; & E\left(E(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|\mathcal{G}_n)\right) \\ = \; & E\left(E(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|\mathcal{G}_n, W(X_n) \leq a)\right) + E\left(E(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|\mathcal{G}_n, W(X_n) > a)\right) \\ \leq \; & \frac{\epsilon}{2} + \frac{\epsilon}{4} + M\frac{\epsilon}{4M} \\ = \; & \epsilon. \end{aligned}$$

Now consider any integer T sufficiently large that:

$$\max[\frac{Mn^*}{T}, \frac{MN}{T}] \leq \epsilon.$$

Then we have:

$$\begin{aligned}
& E(|\frac{\sum_{i=1}^{T} g(X_i)}{T}||X_0 = x, \Gamma_0 = \gamma) \\
\leq\ & E(|\frac{\sum_{i=1}^{n^*} g(X_i)}{T}||X_0 = x, \Gamma_0 = \gamma) \\
+\ & E(\frac{1}{\lfloor \frac{T-n^*}{N} \rfloor} \sum_{j=1}^{\lfloor \frac{T-n^*}{N} \rfloor} \frac{1}{N} \sum_{k=1}^{N} g(X_{N_1+(j-1)N+k}|X_0 = x, \Gamma_0 = \gamma)) \\
+\ & E(|\frac{\sum_{n^*+\lfloor \frac{T-n^*}{N} \rfloor N+1}^{T} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma) \\
\leq\ & \epsilon + \epsilon + \epsilon \\
=\ & 3\epsilon.
\end{aligned}$$

Markov's inequality then gives that:

$$P(|\frac{\sum_{i=1}^{T} g(X_i)}{T}| \geq \epsilon^{\frac{1}{2}}|X_0 = x, \Gamma_0 = \gamma)) \leq 3\epsilon^{\frac{1}{2}}.$$

Since this holds for all sufficiently large T and since $\epsilon > 0$ was arbitrary, the results follows. □

Secondly we can prove the Theorem 6.5 easily using the lemma 6.13.

Proof. Similar to proof of theorem 5.1, the splitting chain of $\{X_n^\gamma\}$ satisfies the conditions of lemma 5.11 for any $\gamma \in \mathcal{Y}$. Therefore we have the WLLN hold. □

Chapter 7

Regional Adaption Algorithm

7.1 Introduction

We notice that the HST algorithm and many modern MCMC algorithms with certain notions of local adaptation e.g. [20], [35] and [13], [22], [16] are not efficient when the target distribution is multi-model. One obvious reason is that different "optimal" kernels are needed in different regions of the state space in many practical problems, however many current adaptive MCMC algorithms try to find the uniformly efficient transition kernel on all regions of the state space through the adaptation. Another reason is that the switches between different models are not continual enough, even in lots of cases the algorithms cannot find the other models except the one that contains the initial value. The last reason is that we do not know how to make the exact partition of the state space.

Regarding the first reason above, we will propose the regional adaptive MCMC algorithm in which the parameters of the proposal distribution with respect to different regions are adapted carefully using the historical samples from the same region so that the performance of the algorithm is "optimal". Regarding the second reason, we will design the mixed regional adaptive MCMC algorithm in which we add another Gaussian

proposal to the regional adaptive MCMC and expect the new part in the proposal distribution will switch the models fluently. Regarding the third reason, we propose a parallel chain adaptation strategy that incorporates multiple Markov chains which are run in parallel and tempered inter-chain adaptation to detect different models. One can find more details about these two strategies in section 2 of R. Craiu, J. Rosenthal and C. Yang [14]. Further we construct the coefficients of different proposal distributions with respect to regions using jump distance under the assumption that the partitions are not optimal, so that we can select the optimal proposal distribution at each rough partitions with more possibilities.

We not only provide theoretical justification using the Theorem 5.4, but also show the performance of the methods using simulations. In addition, we conduct analysis on a mixture model for real data using an algorithm combining the two methods together.

Focusing on the practical aspects of AMCMC, we try to realize the above ideas in this chapter. Section 7.2 is about the regional adaptation. Section 7.3 shows the ergodicity of RAPT first, then using the same idea we prove that Dual RAPT algorithm and Mixed RAPT algorithm are both ergodic too. Section 7.4 presents the real data analysis.

7.2 Regional Adaptation

Consider the target distribution

$$\pi(x|\mu_1, \mu_2, \Sigma_1, \Sigma_2) = 0.5N_{10}(x; \mu_1, \Sigma_1) + 0.5N_{10}(x; \mu_2, \Sigma_2),$$

with $N_d(x; \mu, \Sigma)$ denoting the density of a d-dimensional Gaussian random variable with mean μ and covariance matrix Σ and where $\mu_1 = (3, 3, 3, \cdots, 3)^T$, $\mu_2 = (-3, -3, -3, \cdots, -3)^T$, $\Sigma_1 = I_{10}$ and $\Sigma_2 = 5I_{10}$. The target distribution consists of two different models with the same weight (see Figure 7.1). Obviously due to the different covariance matrix of each model, the "optimal" proposals of each model should be different. For instance, following Roberts and Rosenthal [44] the "optimal" covariance matrix of the Gaussian

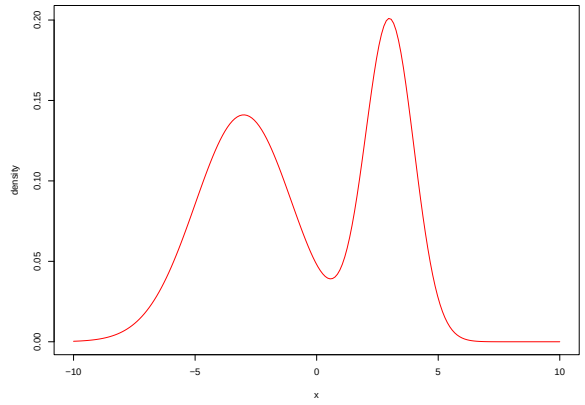

Figure 7.1: *The marginal distribution for each coordinate.*

proposal distribution of Metropolis-Hasting algorithm should be $\frac{2.38^2}{10}I_{10}$ for the model centered at μ_1. Similarly the "optimal" one for the model centered at μ_2 should be $\frac{5\times 2.38^2}{10}I_{10}$. In one word, there does not exist a common "optimal" proposal distribution for both regions, therefore we need to tune the empirical covariance matrices by learning the "history" of different regions of the state space. We assume that there is a partition consists of two regions $\mathcal{S}_{01}, \mathcal{S}_{02}$, that is $\mathcal{S}_{01} \cap \mathcal{S}_{02} =$ and $\mathcal{S}_{01} \cup \mathcal{S}_{02} = \mathcal{S}$. Then following the above analysis we hope to use different proposal distributions $Q_i,\ i = 1, 2$ with respect to different regions $\mathcal{S}_{01}, \mathcal{S}_{02}$. Formally we will use the proposal as:

$$q(x, y) = \sum_{i=1}^{2} \delta_{\mathcal{S}_{0i}}(x) q_i(x, y) \tag{7.1}$$

where $\delta_{\mathcal{S}_{0i}}(x)$ is the indicator function of region $\mathcal{S}_{0i}$, and $q_i(x, y),\ i = 1, 2$ are Gaussian distribution with covariance matrix collecting the information independently in region $\mathcal{S}_{0i}$. For an adaptive Metropolis algorithm with two regions, the acceptance ratio is:

$$\alpha(x, y) = \begin{cases} \frac{\pi(y)}{\pi(x)} & \text{, if } x, y \in \mathcal{S}_{0i} \\ \frac{\pi(y) q_1(y,x)}{\pi(x) q_2(x,y)} & \text{, if } x \in \mathcal{S}_{02}, y \in \mathcal{S}_{01} \\ \frac{\pi(y) q_2(y,x)}{\pi(x) q_1(x,y)} & \text{, if } x \in \mathcal{S}_{01}, y \in \mathcal{S}_{02} \end{cases},$$

where q_i is the density of Q_i.

However the critical problem is that we usually do not know exactly how to split the

state space into two parts $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$. Actually in most cases the true boundary should be certain surface which depends on the target distribution and is hard to compute, so our assumption in this chapter is that the partition is not good. To illustrate easily , let us see Figure 7.2 in which $\mathcal{S}_1$ and $\mathcal{S}_2$ form the partition in practice, and $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$ form the perfect partition. The solid black line indicates the true boundary between $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$ which we do not known. The dashed red line denotes the boundary of the regions $\mathcal{S}_1$ and $\mathcal{S}_2$ used for the regional adaptation. Now we can find that there are still two models in the region $\mathcal{S}_1$. If we still use the proposal distribution (7.1), the wrong proposal will be used in the region between the true boundary and the estimated one. Intuitively we can mix both Q_1 and Q_2 linearly with different weights for each region $\mathcal{S}_i$. So we suggest the proposal as

$$q^{(t)}(x,y) = \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)} q_1(x,y) + \lambda_2^{(i)} q_2(x,y)], \tag{7.2}$$

Obviously fixed coefficients $\lambda_1^{(i)},\ i = 1,2$ are not reasonable. Therefore we hope to modify the weights $\lambda_1^{(i)},\ i = 1,2$ of $q_i^{(t)},\ i = 1,2$ regionally so that we can get some optimal values finally. Then the problem arises: how to adapt the weights of $q_i^{(t)}(x,y),\ i = 1,2$ using the past simulations? We need to find out some statistics which can reflect how good the proposal fits the given region. One possible option using the average square jump distance up to time t is:

$$\lambda_j^{(i)}(t) = \frac{d_j^{(i)}(t)}{\sum_{h=1}^{K} d_h^{(i)}(t)},$$

where $d_j^{(i)}(t)$ is the average square jump distance up to time t computed when the accepted proposals are distributed with $Q_j^{(t)}$ and the current state of the chain lies in $\mathcal{S}_i$. So far using this Dual Regional Adaptive MCMC in which both $q_i^{(t)}(x,y),\ i = 1,2$ and their coefficients are adapted, we have already found more "optimal" proposals than the RAPT algorithm. Formally we will use the proposal distribution at the $t-th$ step as:

$$q^{(t)}(x,y) = \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)}(t) q_1(x,y) + \lambda_2^{(i)}(t) q_2(x,y)], \tag{7.3}$$

where $\lambda_1^{(i)}(t) + \lambda_2^{(i)}(t) = 1$. Then the adaptive Metropolis Hastings algorithm with the above proposal distribution is called Regional Adaptive MCMC(RAPT).

We know that the "optimal" proposal distribution depends on the properties of target distribution, even when we consider all the Gaussian proposals. Therefore we adapt the $q_i(x, y)$ using the past simulations. That is we will use the proposal distribution at the $t - th$step as:

$$q^{(t)}(x, y) = \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)}(t)q_1^{(t)}(x, y) + \lambda_2^{(i)}(t)q_2^{(t)}(x, y)], \tag{7.4}$$

where $\lambda_1^{(i)}(t) + \lambda_2^{(i)}(t) = 1$. We call this adaptive Metropolis MCMC algorithm as Dual RAPT.

When we start the Dual Regional Adaptive MCMC at one of the regions, its performance in this region will be better and better. However it is not very efficient to switch the models. To switch the models continually, we need the proposals to have bigger log than the locally "optimal" ones and have precise jump directions. Therefore we add a third component to the proposal distribution in the Dual RAPT algorithm and hope this part will make a good flow between different regions. From all analysis above, we set up the proposal distribution at the $t - th$ step as:

$$q^{(t)}(x, y) = (1 - \beta) \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)}(t)q_1^{(t)}(x, y) + \lambda_2^{(i)}(t)q_2^{(t)}(x, y)] + \beta q_{whole}^{(t)}(x, y), \tag{7.5}$$

where $q_{whole}^{(t)}$ is adapted using all the samples till t in $\mathcal{S}$, and β is always a constant.

7.3 Theoretical Results

In this section we will prove the ergodicity of the Mixed RAPT algorithm for random walk Metropolis using the Theorem 5 in Roberts and Rosenthal [48] when the state space is compact. We notice there are too many variables in the parameter space when we consider the kernel family generated by the Mixed RAPT. Therefore to make the

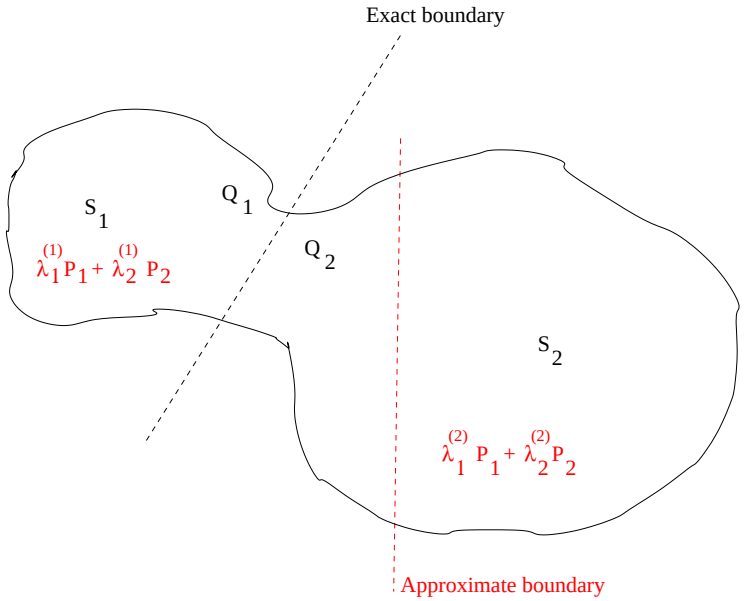

Figure 7.2: *Illustration of the regional adaptive MCMC sampler. The dashed black line indicates the true boundary between $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$ which we do not know. The dashed red line denotes the boundary of $\mathcal{S}_1$ and $\mathcal{S}_2$ used for the regional adaptation.*

main idea more clearly and avoid tedious calculations, at first we will introduce the proof from RAPT case, i.e. only the weights $\lambda_j^{(i)}$, $1 \leq i, j \leq 2$ are adapted. Secondly we will prove that the Dual RAPT algorithm is ergodic. Finally we show that the same idea can be applied to prove the ergodicity of the Mixed RAPT.

Before we start the proof, we introduce some notations first. Let $\{x_i\}_{i=0}^t$ be the samples obtained by time t and $N_i^{(t)}$ be the total number of sample points $\{x_{t_g^i}^i\}_{g=0}^{N_i(t)}$ generated up to time t that are in $\mathcal{S}_i$. We also define the set of time points where the proposal is generated from Q_j and the current state is in $\mathcal{S}_i$, $W_{jt}^{(i)} = \{0 \leq s \leq t : \ x_s \in \mathcal{S}_i$ and proposal at time s is generated from $Q_j\}$.

7.3.1 The Ergodicity of the RAPT Algorithm

Let $\mathcal{M}(\mathcal{S})$ denote the class of densities π with $\pi(x)$ being continuous, $\pi(x) > 0$ for any $x \in \mathcal{S}$ and $\pi(x) = 0$ for $x \notin \mathcal{S}$, where $\mathcal{S} \subset \mathbb{R}^k$ is a compact set. Now we will prove the

ergodicity of the RAPT algorithm with the linear coefficients $\lambda_j^{(i)} = \frac{d_j^{(i)}(t)}{\sum_{h=1}^{2} d_h^{(i)}(t)}$, where $d_j^{(i)}(t)$ is the average jump distance until time t computed for proposals generated from Q_j. And recall that $q_i(x,y),\ i =, 2$ are fixed at each step for the RAPT algorithm. Since $\lambda_2^{(i)} = 1 - \lambda_1^{(i)}$, the *adaption parameter space* consists of $(\lambda_1^{(1)}, \lambda_1^{(2)})|(\lambda_1^{(1)}, \lambda_1^{(2)}) \in [0,1] \times [0,1]$, that is: $\mathcal{Y} = \{(\lambda_1^{(1)}, \lambda_1^{(2)})|(\lambda_1^{(1)}, \lambda_1^{(2)}) \in [0,1] \times [0,1]\}$.

Theorem 7.1. *Let $\mathcal{S} \subset \mathbb{R}^k$ be compact, $\pi \in \mathcal{M}(\mathcal{S})$ and assume $q_i(x,y)$ is positive and continuous for all $x, y \in \mathcal{S}$. Then the RAPT algorithm is ergodic with respect to the target distribution π.*

Following the Theorem 5 in [48] it suffices to prove the following lemmas.

Lemma 7.1. *Under the conditions of the theorem 4.4. There exists $0 < \rho < 1$, for any $\gamma = (\gamma_1, \gamma_2) \in \mathcal{Y}$ such that:*

$$\|P_\gamma^n(x,\cdot) - \pi(\cdot)\| \leq \rho^n.$$

Proof. Since $\mathcal{S}$ is compact and non-empty; by positivity and continuity we have $d = \sup_{x\in\mathcal{S}} \pi(x) < \infty$ and $\epsilon = \min\{\inf_{x,y\in\mathcal{S}} q_1(x,y), \inf_{x,y\in\mathcal{S}} q_2(x,y)\} > 0$. Following (7.4), we have:

$$q_\gamma(x,y) = \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\gamma_i q_1(x,y) + (1-\gamma_i) q_2(x,y)] \geq \epsilon,$$

for any $x, y \in \mathcal{S}$. Choose $B \subseteq \mathcal{S}$. By construction, for fixed x, denote

$$R_x(B) = \left\{ y \in B : \frac{\pi(y) q_\gamma(y,x)}{\pi(x) q_\gamma(x,y)} < 1 \right\}$$

and $A_x(B) = B - R_x(B)$. We have

$$\begin{aligned}
& P_\gamma(x,B) \geq \\
\geq\ & \int_{R_x(B)} q_\gamma(x,y) \min\left\{ \frac{\pi(y) q_\gamma(y,x)}{\pi(x) q_\gamma(x,y)}, 1 \right\} \mu^{Leb}(dy) + \int_{A_x(B)} q_\gamma(x,y) \min\left\{ \frac{\pi(y) q_\gamma(y,x)}{\pi(x) q_\gamma(x,y)}, 1 \right\} \mu^{Leb}(dy) \\
=\ & \int_{R_x(B)} \frac{\pi(y) q_\gamma(y,x)}{\pi(x)} \mu^{Leb}(dy) + \int_{A_x(B)} q_\gamma(x,y) \mu^{Leb}(dy) \\
\geq\ & \frac{\epsilon}{d} \int_{R_x(B)} \pi(y) \mu^{Leb}(dy) + \frac{\epsilon}{d} \int_{A_x(B)} \pi(y) \mu^{Leb}(dy) = \frac{\epsilon}{d} \pi(B).
\end{aligned}$$

Thus $\mathcal{S}$ is small and we have

$$P_\gamma(x, B) \geq \nu(B),$$

where $\nu(B) = \frac{\epsilon}{d}\pi(B)$ is a non-trivial measure on $\mathcal{S}$. Therefore the chain is automatically aperiodic. Note that the measure $\nu(\cdot)$ is independent of γ.

Following the Theorem 16.0.2 in [37]

$$\|P_\gamma(x, \cdot) - \pi(\cdot)\| \leq \rho^n,$$

where $\rho = 1 - \nu(\mathcal{S}) = 1 - \frac{\epsilon}{d}$. □

Lemma 7.2. *Under the conditions of the theorem 7.1. The Diminishing Adaption condition holds when* $\lambda_j^{(i)}(k) = \frac{d_j^{(i)}(k)}{d_1^{(i)}(k)+d_2^{(i)}(k)},\ i = 1, 2;\ j = 1, 2.$

Proof. Denote $f_\lambda(x, y) = \lambda q_1(x, y) + (1 - \lambda)q_2(x, y)$. Since $\mathcal{S}$ is compact, we let $M = \max\{\sup_{x,y\in\mathcal{S}} q_1(x, y), \sup_{x,y\in\mathcal{S}} q_2(x, y)\} > 0$. For any $x \in S_1$ and $A \in \mathcal{B}(\mathcal{S})$, we have:

$$\begin{aligned}
P_{\gamma_k}(x, A) &= \int_{A\cap S_1} f_{\lambda_1^{(1)}(k)}(x, y) \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} dy \\
&+ \int_{A\cap S_2} f_{\lambda_1^{(1)}(k)}(x, y) \min\left\{1, \frac{\pi(y) f_{\lambda_1^{(2)}(k)}(x, y)}{\pi(x) f_{\lambda_1^{(1)}(k)}(x, y)}\right\} dy \\
&+ \delta_x(A) \int_{S_1} f_{\lambda_1^{(i)}(k)}(x, y) \cdot \left[1 - \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\right] dy \\
&+ \delta_x(A) \int_{S_2} f_{\lambda_1^{(1)}(k)}(x, y) \left[1 - \min\left\{1, \frac{\pi(y) f_{\lambda_1^{(2)}(k)}(x, y)}{\pi(x) f_{\lambda_1^{(1)}(k)}(x, y)}\right\}\right] dy.
\end{aligned}$$

Denote the first term $I_k(x, A)$, the second term $II_k(x, A)$, the third term $III_k(x, A)$ and the fourth term $IV_k(x, A)$. Then we have:

$$\begin{aligned}
|P_{\gamma_{k+1}}(x, A) - P_{\gamma_k}(x, A)| &\leq |I_{\gamma_{k+1}}(x, A) - I_{\gamma_k}(x, A)| + |II_{\gamma_{k+1}}(x, A) - II_{\gamma_k}(x, A)| \\
&+ |III_{\gamma_{k+1}}(x, A) - III_{\gamma_k}(x, A)| + |IV_{\gamma_{k+1}}(x, A) - IV_{\gamma_k}(x, A)|.
\end{aligned}$$

Let

$$\alpha_{k_1^{(i)}}(x, y) = \min\left\{1, \frac{\pi(y)[\lambda_1^{(i)}(k)q_1(y, x) + (1 - \lambda_1^{(i)}(k))q_2(y, x)]}{\pi(x)[\lambda_1^{(1)}(k)q_1(x, y) + (1 - \lambda_1^{(1)}(k))q_2(x, y)]}\right\}.$$

Then

$$\begin{aligned}
|II_{\gamma_{k+1}}(x,A) - II_{\gamma_k}(x,A)| &\leq \int_{A\cap S_2} |f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{(k+1)_1^{(2)}}(x,y) - f_{\lambda_1^{(1)}(k)}(x,y)\alpha_{k_1^{(2)}}(x,y)|dy \\
&\leq \int_{A\cap S_2} |f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{(k+1)_1^{(2)}}(x,y) - f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{k_1^{(2)}}(x,y) \\
&+ f_{\lambda_1^{(1)}(k+1)}(x,y)\alpha_{k_1^{(2)}}(x,y) - f_{\lambda_1^{(1)}(k)}(x,y)\alpha_{k_1^{(2)}}(x,y)|dy \\
&\leq \int_{A\cap S_2} f_{\lambda_1^{(1)}(k+1)}(x,y)|\alpha_{(k+1)_1^{(1)}}(x,y) - \alpha_{k_1^{(1)}}(x,y))|dy \\
&+ \int_{A\cap S_2} \alpha_{k_1^{(1)}}(x,y))|f_{\lambda_1^{(1)}(k+1)}(x,y) - f_{\lambda_1^{(1)}(k)}(x,y)|dy \\
&\leq M\int_{A\cap S_2} |\alpha_{(k+1)_1^{(1)}}(x,y) - \alpha_{k_1^{(1)}}(x,y))|dy \\
&+ \int_{A\cap S_2} |f_{\lambda_1^{(1)}(k+1)}(x,y) - f_{\lambda_1^{(1)}(k)}(x,y)|dy.
\end{aligned}$$

For the second term, following the fact that $|f_{\lambda_1^{(1)}(k+1)}(x,y) - f_{\lambda_1^{(1)}(k)}(x,y)| \leq 2M|\lambda_1^{(1)}(k+1) - \lambda_1^{(1)}(k)|$, it suffices to prove $\lim_{k\to\infty}|\lambda_1^{(1)}(k+1) - \lambda_1^{(1)}(k)| = 0$. For the first term, we have:

$$\begin{aligned}
& M\int_{A\cap S_2} |\alpha_{(k+1)_1^{(1)}}(x,y) - \alpha_{k_1^{(1)}}(x,y))|dy \\
= \; & M\int_{A\cap S_2} \frac{\pi(y)}{\pi(x)}\left|\frac{f_{\lambda_{k+1}^{(2)}}(x,y)}{f_{\lambda_{k+1}^{(1)}}(x,y)} - \frac{f_{\lambda_k^{(2)}}(x,y)}{f_{\lambda_k^{(1)}}(x,y)}\right| dy \\
\leq \; & \frac{Md}{\pi(x)}\int_{A\cap S_2}\left|\frac{f_{\lambda_{k+1}^{(2)}}(x,y)}{f_{\lambda_{k+1}^{(1)}}(x,y)} - \frac{f_{\lambda_k^{(2)}}(x,y)}{f_{\lambda_k^{(1)}}(x,y)}\right| dy.
\end{aligned}$$

It is easy to check that when $\lim_{k\to\infty}|\lambda_1^{(i)}(k+1) - \lambda_1^{(i)}(k)| = 0,\ i = 1,2$ the first term tends to zero. We know that $\lambda_j^{(i)}(k) = \frac{d_j^{(i)}(k)}{d_1^{(i)}(k)+d_2^{(i)}(k)},\ i = 1,2;\ j = 1,2$. Consider the random variable $d_n = (X_{n+1} - X_n)^2$. Since $\mathcal{S}$ is compact, we know that d_n is bounded

by some $R > 0$. Therefore

$$\begin{aligned}
& |\lambda_1^{(1)}(k+1) - \lambda_1^{(1)}(k)| \\
= & \left|\frac{d_1^{(1)}(k+1)}{d_1^{(1)}(k+1) + d_2^{(1)}(k+1)} - \frac{d_1^{(1)}(k)}{d_1^{(1)}(k) + d_2^{(1)}(k)}\right| \\
= & \left|\frac{d_1^{(1)}(k+1)d_2^{(1)}(k) - d_1^{(1)}(k)d_2^{(1)}(k+1)}{[d_1^{(1)}(k+1) + d_2^{(1)}(k+1)][d_1^{(1)}(k) + d_2^{(1)}(k)]}\right| \\
\leq & \left|\frac{(k+1)^{-1}\{[kd_1^{(1)}(k) + (x_{k+1} - x_k)^2]d_2^{(1)}(k) - d_1^{(1)}(k)[kd_2^{(1)}(k) + (x_{k+1} - x_k)^2]\}}{[d_1^{(1)}(k+1) + d_2^{(1)}(k+1)][d_1^{(1)}(k) + d_2^{(1)}(k)]}\right| \\
\leq & \left|\frac{(k+1)^{-1}\{[kd_1^{(1)}(k) + (x_{k+1} - x_k)^2]d_2^{(1)}(k) + d_1^{(1)}(k)[kd_2^{(1)}(k) + (x_{k+1} - x_k)^2]\}}{[d_1^{(1)}(k+1) + d_2^{(1)}(k+1)][d_1^{(1)}(k) + d_2^{(1)}(k)]}\right| \\
\leq & \frac{R^2}{(k+1)(d_1^{(1)}(k+1) + d_2^{(1)}(k+1))} = \frac{R^2}{\sum_{i=1}^{k+1}(x_i - x_{i-1})^2} \to 0 \; as \; k \to \infty.
\end{aligned}$$

So we have: $|II_{\gamma_{k+1}}(x, A) - II_{\gamma_k}(x, A)| \to 0$. Similarly we can prove $|I_{\gamma_{k+1}}(x, A) - I_{\gamma_k}(x, A)| \to 0$, $|III_{\gamma_{k+1}}(x, A) - III_{\gamma_k}(x, A)| \to 0$, $|IV_{\gamma_{k+1}}(x, A) - IV_{\gamma_k}(x, A)| \to 0$. Therefore, the Diminishing Adaptation holds. □

7.3.2 The Ergodicity of the Dual RAPT Algorithm

Further we will prove the ergodicity of the Dual RAPT algorithm in this subsection. As stated in section 7.2, the proposal distribution at the $t - th$ step of the Dual RAPT is

$$q^{(t)}(x, dy) = \sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^{(i)}(t)q_1^{(t)}(x, y) + \lambda_2^{(i)}(t)q_2^{(t)}(x, y)],$$

where the $q_i^{(t)},\ i = 1, 2$ are Gaussian distribution with the covariance matrices adapted using the same algorithm as [26] regionally. More precisely, $q_i^{(t)}(x, y)\ i = 1, 2$ are the Gaussian distributions with mean at the current point X_{t-1} and covariance $C_i^{(t)} = C_i^{(t)}(X_{t_0}^i, X_{t_1}^i, \cdots, X_{t_{N_i(t)}}^i)$, where $C_i^{(t)}$ is defined as below:

$$C_i^{(t)} = \begin{cases} C_{0i}, & t \leq t_0 \\ s_d cov(X_{t_0}^i, X_{t_1}^i, \cdots, X_{t_{N_i(t)}}^i) + s_d \epsilon I_d, & t > t_0 \end{cases}.$$

Here s_d is a parameter that depends only on the dimension d, $\epsilon > 0$ is a constant that we may choose very small compared to the size of S, I_d denotes the d-dimensional identity matrix and the initial covariance C_{0i} is a strictly positive definite matrix chosen in line with our prior knowledge of π.

We note that all the parameters adapted in the kernel of the Dual RAPT are made up of four parts: $\lambda_1^{(i)}(t)$, $i = 1, 2$ and $C_i^{(t)}$, $i = 1, 2$. In the proof of Theorem 1 of [26], they have proved the following inequality:

$$c_1 I_k \leq C \leq c_2 I_k,$$

for some $c_1, c_2 > 0$ (i.e., both $C - c_1 I_k$ and $c_2 I_k - C$ are non-negative-definite). If we define $\mathbb{M}(c_1, c_2) = \{M \in M_k | c_1 I_k \leq M \leq c_2 I_k\}$ where M_k is the set of all positive definite matrices of dimension k, that is, $\mathbb{M}(c_1, c_2)$ consists of all the positive definite matrix M such that both $M - c_1 I_k$ and $c_2 I_k - M$ are non-negative definite. Then the parameter space can be expressed as:

$$\mathcal{Y} = \{(\lambda_1^{(1)}, \lambda_1^{(2)}, C_1^{(t)}, C_2^{(t)}) | (\lambda_1^{(1)}, \lambda_1^{(2)}, C_1^{(t)}, C_2^{(t)}) \in [0, 1] \times [0, 1] \times \mathbb{M}(c_1, c_2) \times \mathbb{M}(c_1, c_2)\}.$$

Without loss of generality, we will consider the parameter space

$$\mathcal{Y} = [0, 1] \times [0, 1] \times \mathbb{M}(c_1, c_2) \times \mathbb{M}(c_1, c_2).$$

Then we can prove that the proposed algorithm is ergodic.

Theorem 7.2. *Suppose the state space $\mathcal{S}$ is compact, $\pi \in \mathcal{M}(\mathcal{S})$ and $q_i(x, y)$ are Gaussian distributions as described above. Then the Dual RAPT algorithm is ergodic with respect to the target distribution π.*

Proof. Using the fact that $\inf_{x,y \in \mathcal{S},\ M \in \mathbb{M}(c_1,c_2)} q_M(x, y) > 0$ (where q_M denotes the density function of Gaussian distribution with variance M), we have $\inf_{x,y \in \mathcal{S}, \gamma \in \mathcal{Y}} q_\gamma(x, y) > 0$. Then following a similar proof to that of the Lemma 7.1, one can show that there exists $0 < \rho < 1$ so that for any $\gamma \in \mathcal{Y}$

$$\|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \rho^n.$$

Using the fact that $d_n = (X_{n+1} - X_n)^2$ is still bounded and the similar proof of the Lemma 7.2, we can prove that the Diminishing Adaptation condition holds for Dual RAPT. □

7.3.3 The Ergodicity of the Mixed RAPT Algorithm

Finally we show the ergodicity of the Mixed RAPT algorithm which has one more component $q^{(t)}_{whole}$ than the Dual RAPT algorithm. We still tune the Gaussian proposal density $q^{(t)}_{whole}$ in (7.5) at the $t-th$ step by adapting its covariance matrix as in [26]. First compute the empirical covariance matrix $C^{(t)}$ of $\{X_i\}_{i=1}^t$ as:

$$C^{(t)} = \begin{cases} C_0, & t \leq t_0 \\ s_d cov(X_0, X_1, \cdots, X_t) + s_d \epsilon I_d, & t > t_0 \end{cases}.$$

Then we will use the proposal of Mixed Dual RAPT algorithm at the $t-th$ step as:

$$q^{(t)}(x,y) = (1-\beta)\sum_{i=1}^{2} 1_{\mathcal{S}_i}(x)[\lambda_1^i(t) q_1^{(t)}(x,y) + \lambda_2^i(t) q_2^{(t)}(x,y)] + \beta q^{(t)}_{whole}(x,y),$$

where $q_i^{(t)}(x,y),\ \lambda_i^{(t)},\ i = 1,2$ are the same as those in the dual adaptive kernel and $q^{(t)}_{whole}(x,y)$ is the Gaussian proposal distribution with covariance $C^{(t)}$. Similarly following the proof of Theorem 1 in [26] and the construction of the covariances $C^{(t)},\ C_i^{(t)},\ i = 1,2$, we know that:

$$c_1 I_k \leq C^{(t)},\ C_i^{(t)} \leq c_2 I_k.$$

for some $c_1, c_2 > 0$. Therefore we consider the adaption parameter space as:

$$\mathcal{Y} = [0,1] \times [0,1] \times \mathbb{M}(c_1,c_2) \times \mathbb{M}(c_1,c_2) \times \mathbb{M}(c_1,c_2).$$

Theorem 7.3. *Suppose the state space $\mathcal{S}$ is compact, $\pi \in \mathcal{M}$ and the mixed proposal distribution $q^{(t)}(x,y)$ is defined as above. Then the Mixed RAPT algorithm is ergodic with respect to the target distribution π.*

Proof. Given that $\inf_{x,y\in\mathcal{S},\ M\in\mathbb{M}(c_1,c_2)} q_M(x,y) > 0$, then $\inf_{x,y\in\mathcal{S},\gamma\in\mathcal{Y}} q_\gamma(x,y) > 0$. Following the similar proof of lemma 7.1 there exists $0 < \rho < 1$ so that for any $\gamma \in \mathcal{Y}$

$$\|P_\gamma^n(x,\cdot) - \pi(\cdot)\| \leq \rho^n.$$

The proof for showing diminishing adaptation is similar to that of the Lemma 7.2 using that $d_n = (X_{n+1} - X_n)^2$ is bounded. Then, following the Theorem 5 in Roberts and Rosenthal [[48]] (2007), we obtain the ergodicity of the Mixed RAPT. □

7.4 Real Data Example: Genetic Instability of Esophageal Cancers

Cancer cells undergo a number of genetic changes during neoplastic progression, including loss of entire chromosome sections. We call the loss of a chromosome section containing one allele by abnormal cells by the term "Loss of Heterozygosity" (LOH). When an individual patient has two different alleles, LOH can be detected using laboratory assays. Chromosome regions with high rates of LOH are hypothesized to contain genes which regulate cell behavior so that loss of these regions disables important cellular controls.
To locate "Tumor Suppressor Genes"(TSGs), the Seattle Barrett's Esophagus research project [8] has collected LOH rates from esophageal cancers for 40 regions, each on a distinct chromosome arm. A hierarchical mixture model has been constructed by [54] in order to determine the probability of LOH for both the "background" and TSG groups. The labeling of the two groups is unknown so we model the LOH frequency using a mixture model, as described by [15]. We obtain the hierarchical Binomial-BetaBinomial mixture model

$$X_i \sim \eta\text{Binomial}(N_i,\pi_1) + (1-\eta)\text{Beta-Binomial}(N_i,\pi_2,\gamma),$$

with priors

$$\eta \sim \text{Unif}[0,1],$$

$$\pi_1 \sim \text{Unif}[0,1],$$

$$\pi_2 \sim \text{Unif}[0,1],$$

$$\gamma \sim \text{Unif}[-30,30],$$

where η is the probability of a location being a member of the binomial group, π_1 is the probability of LOH in the binomial group, π_2 is the probability of LOH in the beta-binomial group, and γ controls the variability of the beta-binomial group. Here we parameterize the Beta-Binomial so that γ is a variance parameter defined on the range $-\infty \leq \gamma \leq \infty$. As $\gamma \to -\infty$ the beta-binomial becomes a binomial and as $\gamma \to \infty$ the beta-binomial becomes a uniform distribution on $[0,1]$. Similarly we also parameterized η, π_1 and π_2. This results in the unnormalized posterior density

$$\pi(\eta, \pi_1, \pi_2, \gamma | x) \propto \Pi_{i=1}^{N} f(x_i, n_i | \eta, \pi_1, \pi_2, \omega_2)$$

on the prior range, where

$$\begin{aligned} f(x,n|\eta,\pi_1,\pi_2,\omega_2) &= \eta \begin{pmatrix} n \\ x \end{pmatrix} \pi_1^x (1-\pi_1)^{n-x} + \\ &+ (1-\eta) \begin{pmatrix} n \\ x \end{pmatrix} \frac{\Gamma(\frac{1}{\omega_2})}{\Gamma(\frac{\pi_2}{\omega_2})\Gamma(\frac{1-\pi_2}{\omega_2})} \frac{\Gamma(x+\frac{\pi_2}{\omega_2})}{\Gamma(n-x+\frac{1-\pi_2}{\omega_2})\Gamma(n+\frac{1}{\omega_2})} \end{aligned}$$

and $\omega_2 = \frac{e^\gamma}{2(1+e^\gamma)}$. In order to use the random walk Metropolis we have used the logistic transformation on all the parameters with range $[0,1]$. However, all our conclusions are presented on the original scale for an easier interpretation.

Using the optimization procedures used by [54] we determine that the two modes of π are reasonably well separated by the partition made of $S_1 = \{(\eta, \pi_1, \pi_2, \gamma) \in [0,1] \times [0,1] \times [0,1] \times [-30,30] | \pi_2 \geq \pi_1\}$ and $S_2 = \{(\eta, \pi_1, \pi_2, \gamma) \in [0,1] \times [0,1] \times [0,1] \times [-30,30] | \pi_2 \leq \pi_1\}$.

Mean in	Region 1	Region 2	Whole space
η	0.897	0.079	0.838
π_1	0.229	0.863	0.275
π_2	0.714	0.237	0.679
γ	15.661	-14.796	13.435

Table 7.1: *Simulation results for the LOH data.*

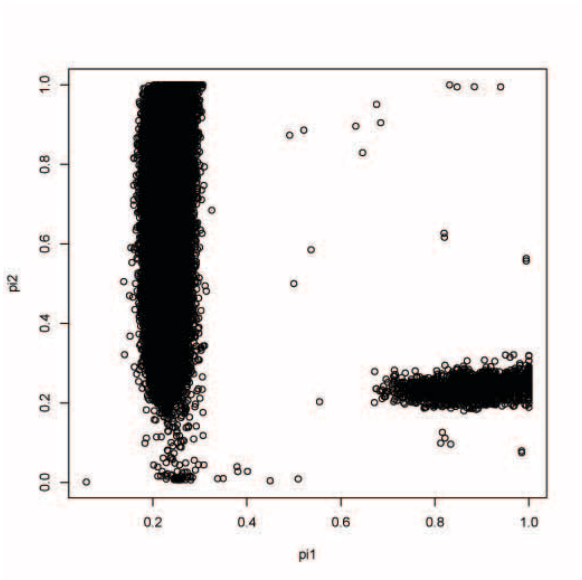

Figure 7.3: *Scatterplot of the 50,000 samples for* (π_1, π_2).

Simulation results

We will combine the parallel chain strategy with the MRAPT algorithm together in this part. For more details of the parallel chain strategy, readers can refer to R.Craiu, J.Rosenthal, and C.Yang [14]. Here we run five parallel mixed RAPT algorithms to simulate from π using the partition $S_1 \cup S_2$. After $50,000$ iterations, we obtain $\lambda_1^{(1)} = 0.923$ and $\lambda_1^{(2)} = 0.412$. Further results are shown in Table 7.4. A two dimensional scatter plot of the (π_1, π_2) samples which is similar to the findings of [54] (Figure 8) is shown in Figure 7.3.

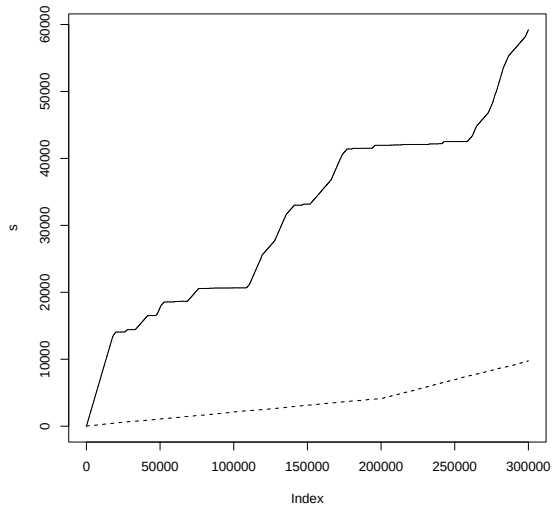

Figure 7.4: *The total number of switches times for the five parallel Mixed RAPT vs the number of switch times of a single Mixed RAPT run for 300,000 iterations.*

The most advantage to run five parallel MRAPT together is that all these parallel chains can share all the past information so that they can learn the "geography" much more quickly than a single chain, although the total iteration times are the same. To see this fact more clearly, we run a single Mixed RAPT algorithm for 300,000 iterations, and five parallel Mixed RAPT algorithms independently for 60,000 iterations each. To be fair, we plot the total number of switches for the five parallel chains up to $60,000$ iterations versus the number of switches for the single chain up to $5 \times 60,000$ iterations in Figure 7.4. One can see that the five parallel Mixed RAPT switch the models much better than a single chain.

Finally we use the BGR diagnostic statistic as a criterion to describe how these parallel chains learn from each other. More precisely, following the definition of the BGR diagnostic statistic, we can assume all chains have the same information regarding π when the BGR is close to 1. For this LOH data example, we can see that each chain has learned almost all the information from the other chains after 40,000 iterations, because

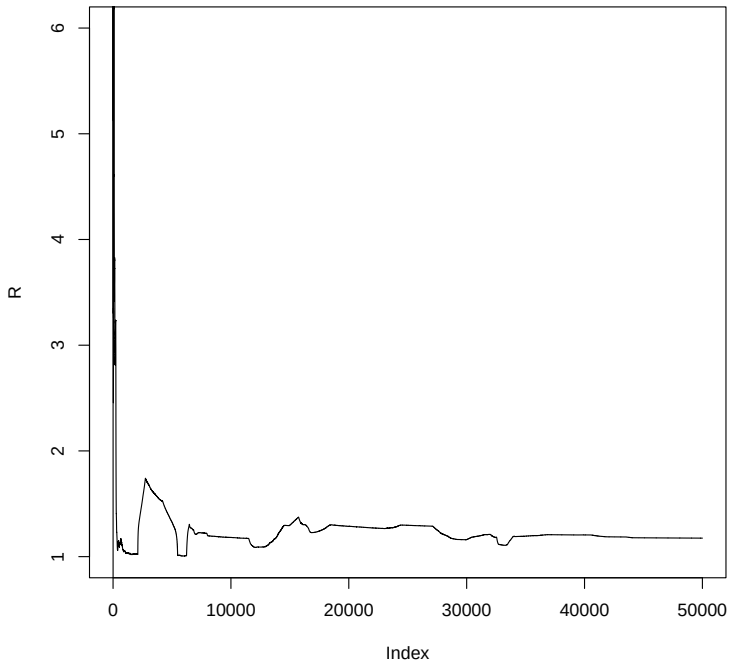

Figure 7.5: *The evolution of BGR's R statistics*

the BGR statistic becomes below 1.1 at that time (see Figure 7.5). In the practical use we only need to run a single chain to reduce the computation costs.

Chapter 8

The Ergodicity of Modified Mixed RAPT on the State Space $\mathbb{R}^k$

8.1 Introduction

In last chapter we designed the MRAPT algorithm to sample from a multi-model distribution on compact state space(see also [14]). Here we will try to construct the modified MRAPT algorithm when the state space is $\mathbb{R}^k$ and show the ergodicity of modified MRAPT algorithm under additional conditions. Furthermore, we will simulate some toy examples to discuss the complications arising when using AMCMC, especially adaptive random walk Metropolis, for sampling from multi-model targets and also when the optimal proposal distribution is regional, i.e. the optimal proposal should change across regions of the state space, and check our theoretical results.

We still suppose the state space $\mathcal{X} = \mathcal{S}_1 \cup \mathcal{S}_2$. Given the initial value $X_0 = x_0$ and $\Gamma_0 = \gamma_0$, at $t - th$ step we will run the MH algorithm with the adaptive proposal distribution:

$$q^{(t)}(x, y) = \sum_{i=1}^{2} \mathbb{I}_x(\mathcal{S}_i)\{(1-\beta)[\lambda_1^{(i)}(t) q_1^{(t)}(x, y) + \lambda_2^{(i)}(t) q_2^{(t)}(x, y)] + \beta q_{whole}^{(t)}(x, y)\}, \quad (8.1)$$

where $q_i^{(t)}$, $j = 1, 2$ are Gaussian distributions with adaptive variance covariance

matrixes $C_i^{(t)}$ as in [26] (henceforth denoted HST), but here we need to do a little change to ensure the ergodicity and avoid singular cases. Suppose $\{x_i\}_{i=0}^t$ are the samples obtained until time t, let $N_i(t)$ is the total number of sample points $\{x_{t_g}^i\}_{g=0}^{N_i(t)}$ generated up to time t that are lying in $\mathcal{S}_i$. We also define the set of time points at which the proposal is generated from Q_j and the current state is in $\mathcal{S}_i$, $W_{jt}^{(i)} = \{0 \leq s \leq t : \ x_s \in \mathcal{S}_i$ and proposal at time s is generated from $Q_j\}$. For some large $B > 0$ and $0 < \tau < \frac{1}{2}$, we let $q_{whole}^{(t)}$ be a Gaussian distribution with variance covariance matrix $C^{(t)}$. Since it is hard to estimate the bound of samples $\{x_i\}_{i=0}^t$ for any fixed t, we can not ensure the ergodicity if we still tune the C^t as in [26]. We construct new samples $\{y_i\}_{i=0}^t$ using $\{x_i\}_{i=0}^t$. Let $y_i = x_i, \ i = 1, \cdots, n_0$, when $t > n_0$, if $|x_t| \leq B + t^\tau$ for some $B > 0$ large enough and $0 < \tau < \frac{1}{2}$, we still set $y_t = x_t$, otherwise $y_t = x_{t-1}$. The we can adapt C^t as below:

1. When $t \leq n_0$, we set $C^t = C_0$, where C_0 is some fixed positive definite matrix;

2. When $t > n_0$, if $Tr(s_k cov(y_0, y_1, \cdots, y_t) + s_k \epsilon I_k) \leq L$ where $L > 0$ is large enough, we will set

$$C^{(t)} = s_k cov(y_0, y_1, \cdots, y_t) + s_k \epsilon I_k,$$

otherwise $C^{(t)} = C^{(t-1)}$.

As a basic optimal choice for scaling parameter we have adopted the value $s_k = \frac{2.4^2}{d}$ from [19]. Similarly we let $q_i^{(t)}$ be Gaussian distributions with adaptive variance covariance matrixes $C_i^{(t)}$. Let us construct new samples $\{y_{t_g}^i\}_{g=0}^{N_i(t)}$ first, Let $y_{t_g}^i = x_{t_g}^i, \ 1 \leq t_g \leq n_0$, when $t_g > n_0$, if $|x_{t_g}^i| \leq B + N_i(t)^\tau$ for some $B > 0$ large enough and $0 < \tau < \frac{1}{2}$, we still set $y_{t_g}^i = x_{t_g}^i$, otherwise $y_{t_g}^i = x_{t_{g-1}}^i$. We can adapt $C_i^{(t)}$ as below:

1. When $t \leq n_0$, we set $C_i^t = C_i, \ i = 1, 2$, where $C_i, \ i = 1, 2$ are fixed positive definite matrices;

2. When $t > n_0$, if $Tr(cov(y_{t_0}^i, y_{t_1}^i, \cdots, y_{t_{N_i(t)}}^i) + s_k \epsilon I_k) \leq L$, we will set

$$C^{(t)} = s_k cov(y_{t_0}^i, y_{t_1}^i, \cdots, y_{t_{N_i(t)}}^i) + s_k \epsilon I_k,$$

else $C^{(t)} = C^{(t-1)}$.

Since we do not know enough information to find the perfect partition $\mathcal{S}_i, i = 1, 2$ especially when both model affect each other too much, we will use the linear combination of $q_i^{(t)}, i = 1, 2$ at each region. We will adapt the coefficients $\lambda_j^{(i)}(t),\ i, j = 1, 2$ of $q_j^{(t)}(x, y)$ when $x \in \mathcal{S}_i$ respectively using the ratio of jump distance. That is

$$\lambda_j^{(i)}(t) = \frac{d_j^{(i)}(t)}{\sum_{h=1}^{2} d_h^{(i)}(t)},$$

where $d_h^{(i)}(t)$ is the average square jump distance up to time n computed when the accepted proposals are distributed with Q_h and the current state of the chain lies in $\mathcal{S}_i$. To avoid singular case, we suppose $\lambda_j^{(i)}(t) = \max\{a, \frac{d_j^{(i)}(t)}{\sum_{h=1}^{2} d_h^{(i)}(t)}\}$, where $a > 0$ will take very small value.

Recall that the average square jump distance:

$$d_j^{(i)}(t) = \frac{\sum_{s \in W_{jt}^{(i)}} |x_{t_{s+1}}^i - x_{t_s}^i|^2}{|W_{jt}^{(i)}|},$$

where $|W_{jt}^{(i)}|$ denotes the number of elements in the set $W_{jt}^{(i)}$. Since all the covariances $C^{(t)},\ C_i^{(t)},\ i = 1, 2$ satisfy the matrix inequality :

$$\epsilon I_k \leq C^{(t)},\ C_i^{(t)} \leq L I_k.$$

and $\lambda_1^{(i)}(t) = 1 - \lambda_2^{(i)}(t)$, we can see that the parameter space consists of

$$\{(\lambda_1^{(1)}(t), \lambda_1^{(2)}(t), C_1^{(t)}, C_2^{(t)}, C^{(t)}) \in [a, 1] \times [a, 1] \times \mathbb{M}(\epsilon, L) \times \mathbb{M}(\epsilon, L) \times \mathbb{M}(\epsilon, L)\},$$

where $\mathbb{M}(\epsilon, L) = \{M \in M_k | \epsilon I_k \leq M \leq L I_k\}$, M_k denotes the set of all positive definite matrices of dimension k, that is $\mathbb{M}(\epsilon, L)$ consists of all the positive definite matrix M such that both $M - \epsilon I_k$ and $L I_k - M$ are non-negative-definite. Without losing generalities, we let the parameter space

$$\mathcal{Y} = [a, 1] \times [a, 1] \times \mathbb{M}(\epsilon, L) \times \mathbb{M}(\epsilon, L) \times \mathbb{M}(\epsilon, L).$$

We hope the mixed RAPT proposal to use "better" proposals with high proportion in each region, and expect the q_{whole} is more efficient to switch the models through learning the "geography" of different regions separately than the general AMCMC algorithm.
To prove the ergodicity theorem of MRAPT algorithm when the state space is R^k, we need the target distribution π to have smoothly decreasing properties in its tail. First we suppose the target density π on $\mathbb{R}^k$ super-exponential that is it has exponential or lighter tails. More precisely, $\pi(x)$ is positive and has continuous first derivatives such that:

$$\lim_{|x|\to\infty} n(x)\cdot\nabla log\pi(x) = -\infty,$$

where $n(x)$ denotes the unit vector $\frac{x}{|x|}$. The condition implies that for any $H > 0$ there exists $R > 0$ such that:

$$\frac{\pi(x+an(x))}{\pi(x)} \leq \exp(-aH)\ (|x| \geq R,\ a \geq 0). \tag{8.2}$$

That is, $\pi(x)$ is at least exponentially decaying along any ray with rate H tending to infinity as x goes to infinity. It also implies that for ϵ small enough the contour manifold C_ϵ defined by $C_\epsilon = \{x \in \mathbb{R}^k | \pi(x) = \epsilon\}$ can be parameterized by the unit sphere S^{k-1}, that is:

$$C_\epsilon = \{r(\zeta)\zeta | \zeta \in S^{k-1}\},$$

where r is a positive continuous function on S^{k-1}, and the set enclosed by the contour manifold $C_{\pi(x)}$ through a point x is the region $A_0(x) = \{y \in \mathbb{R}^k | \pi(x) \leq \pi(y)\}$. Secondly we assume target density π is decreasing along any direction when $|x|$ is large enough. That is:

$$\limsup_{|x|\to\infty} n(x)\cdot m(x) < 0, \tag{8.3}$$

where $m(x) = \frac{\nabla\pi(x)}{|\nabla\pi(x)|}$. We suppose $\mathcal{E} = \{$all the positive density functions satisfy (8.2) and (8.3)$\}$. We also suppose ∂S_i is a hyperplane. However, Theorem 8.1 holds also when ∂S_i is any surface with good smooth properties.

Theorem 8.1. *Suppose $\pi(x) \in \mathcal{E}$. Let $S_1 \cup S_2$ be a partition of $\mathbb{R}^k$ where ∂S_i is a hyperplane in R^k. Then the above MRAPT algorithm is ergodic with respect to distribution $\pi(\cdot)$.*

8.2 Preliminary

In the MRAPT algorithm, we actually consider a family of kernels $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ generated by the adaptive MH algorithm. For any $\gamma = (\lambda_1^{(1)}(\gamma), \lambda_1^{(2)}(\gamma), C_1^{(\gamma)}, C_2^{(\gamma)}, C^{(\gamma)}) \in \mathcal{Y}$, P_γ is the transition kernel corresponding to the proposal distribution

$$q_\gamma(x,y) = \sum_{i=1}^{2} \mathbb{I}_x(\mathcal{S}_i)\{(1-\beta)[\lambda_1^{(i)}(\gamma) q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(i)}(\gamma)) q_2^{(\gamma)}(x,y)] + \beta q_{whole}^{(\gamma)}(x,y)\},$$

where $q_\gamma^{(i)}$ and q_γ^w are Gaussian distributions with variance matrix $C_\gamma^{(i)}$ and C_γ respectively and with mean x. We will apply the Theorem 4.4 to prove teh Theorem 8.1 . Before starting the proof, we introduce some notations first. Define the acceptance region for each $x \in \mathbb{R}^k$ and $\gamma \in \mathcal{Y}$ as

$$A(x;\gamma) = \{y \in \mathbb{R}^k | \pi(y) q_\gamma(y,x) \geq \pi(x) q_\gamma(x,y)\}.$$

Denote:

$$A_i(x;\gamma) = A(x;\gamma) \cap \mathcal{S}_i \ i = 1,2.$$

The acceptance rate $\alpha_\gamma(x,y) = \min\{1, \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}\}$, and the rejection region is

$$R(x;\gamma) = \{y \in \mathbb{R}^k | \pi(y) q_\gamma(y,x) < \pi(x) q_\gamma(x,y)\}.$$

Denote:

$$R_i(x;\gamma) = R(x;\gamma) \cap S_i \ i = 1,2.$$

We also denote:

$$A(x) = \{y \in \mathbb{R}^k | \pi(y) \geq \pi(x)\},$$

$$R(x) = \{y \in \mathbb{R}^k | \pi(y) < \pi(x)\}$$

and denote $A_i(x) = A(x) \cap S_i$, $R_i(x) = R(x) \cap S_i$, $i = 1, 2$.

8.3 Some Technical Results

Let us prove some lemmas first.

Lemma 8.1. *For any $\gamma \in \mathcal{Y}$, we have $\frac{q_\gamma(y,x)}{q_\gamma(x,y)}$ is uniformly bounded. That is there exist $M > m > 0$ such that $0 < m < \frac{q_\gamma(y,x)}{q_\gamma(x,y)} \leq M$ for any $\gamma \in \mathcal{Y}$.*

Proof. Without losing generalities, suppose $x \in \mathcal{S}_1$. Obviously when $y \in \mathcal{S}_1$, we have $\frac{q_\gamma(y,x)}{q_\gamma(x,y)} = 1$. It suffices to prove the result when $y \in \mathcal{S}_2$. In this case we know that

$$\begin{aligned}
\frac{q_\gamma(y,x)}{q_\gamma(x,y)} &= \frac{(1-\beta)[\lambda_1^{(2)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(2)}(\gamma))q_2^{(\gamma)}(x,y)] + \beta q_{whole}^{(\gamma)}(x,y)}{(1-\beta)[\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(1)}(\gamma))q_2^{(\gamma)}(x,y)] + \beta q_{whole}^{(\gamma)}(x,y)} \\
&= \frac{\frac{\lambda_1^{(2)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(2)}(\gamma))q_2^{(\gamma)}(x,y)}{\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(1)}(\gamma))q_2^{(\gamma)}(x,y)} + \frac{\beta}{1-\beta}\frac{q_{whole}^{(\gamma)}(x,y)}{\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(\gamma)})q_2^{(\gamma)}(x,y)}}{1 + \frac{\beta}{1-\beta}\frac{q_{whole}^{(\gamma)}(x,y)}{\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(1)}(\gamma))q_2^{(\gamma)}(x,y)}}.
\end{aligned}$$

We can denote $W = \frac{\beta}{1-\beta}\frac{q_{whole}^{(\gamma)}(x,y)}{\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(1)}(\gamma))q_2^{(\gamma)}(x,y)} > 0$. If we know that there exist $M > 1 > m > 0$ such that for any $\gamma \in \mathcal{Y}$, $m < \frac{\lambda_1^{(2)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(2)}(\gamma))q_2^{(\gamma)}(x,y)}{\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y)+(1-\lambda_1^{(1)}(\gamma))q_2^{(\gamma)}(x,y)} < M$, then we have $\frac{q_\gamma(y,x)}{q_\gamma(x,y)} \leq \frac{M+W}{1+W} < M$ and $\frac{q_\gamma(y,x)}{q_\gamma(x,y)} \geq \frac{m+W}{1+W} > m$. We have

$$\frac{\lambda_1^{(2)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(2)}(\gamma))q_2^{(\gamma)}(x,y)}{\lambda_1^{(1)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(1)}(\gamma))q_2^{(\gamma)}(x,y)} = \frac{\lambda_1^{(2)}(\gamma)\frac{q_1^{(\gamma)}(x,y)}{q_2^{(\gamma)}(x,y)} + (1-\lambda_1^{(2)}(\gamma))}{\lambda_1^{(1)}(\gamma)\frac{q_1^{(\gamma)}(x,y)}{q_2^{(\gamma)}(x,y)} + (1-\lambda_1^{(1)}(\gamma))},$$

Let $z = \frac{q_1^{(\gamma)}(x,y)}{q_2^{(\gamma)}(x,y)}$, we know that $0 < z < \infty$. Consider function $g_\gamma(z) = \frac{\lambda_1^{(2)}(\gamma)z+(1-\lambda_1^{(2)}(\gamma))}{\lambda_1^{(1)}(\gamma)z+(1-\lambda_1^{(1)}(\gamma))}$, we know that:

$$g_\gamma^{'}(z) = \frac{\lambda_1^{(2)}(\gamma) - \lambda_1^{(1)}(\gamma)}{[\lambda_1^{(1)}(\gamma)z + (1-\lambda_1^{(1)}(\gamma))]^2}$$

If $\lambda_1^{(2)}(\gamma) \geq \lambda_1^{(1)}(\gamma)$, we have $g_\gamma'(z) > 0$, then $g_\gamma(z)$ is increasing function. So we have:

$$\begin{aligned} & g_\gamma(0) \leq g_\gamma(z) \leq g_\gamma(\infty) \\ \Rightarrow \quad & \frac{1-\lambda_1^{(2)}(\gamma)}{1-\lambda_1^{(1)}(\gamma)} \leq g_\gamma(z) \leq \frac{\lambda_1^{(1)}(\gamma)}{\lambda_1^{(2)}(\gamma)} \\ \Rightarrow \quad & 1 \leq g_\gamma(z) \leq \frac{1}{a}. \end{aligned}$$

If $\lambda_1^{(2)}(\gamma) < \lambda_1^{(1)}(\gamma)$, we have $g_\gamma'(z) < 0$, then $g(z)$ is decreasing function. Similarly we have

$$a \leq g_\gamma(z) \leq 1.$$

From all above we know that: $\frac{1}{a} \geq g(z) \geq a$, therefore we can let $m = \min\{a, 1\}$ and $M = \max\{\frac{1}{a}, 1\}$. □

Since $\pi(x)$ is super-exponential, we know that for any $\gamma \in \mathcal{Y}$ there exists $\delta_1 > 0$ such that for any $y \in C^+_{\pi(x)}(\delta_1) = \{y + sn(y) | y \in C_{\pi(x)} \cap S_2, s \geq \delta_1\}$, following the Lemma 8.1 we have:

$$\frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} \leq M\frac{\pi(y)}{\pi(x)} \leq M\exp\{-aH\} < 1$$

for x large enough. That is $C^+_{\pi(x)}(\delta_1) \subset R_2(x)$. Similarly, we also can choose $\delta_1 > 0$, then consider any $y \in C^-_{\pi(x)}(\delta_1)$, where $C^-_{\pi(x)}(\delta_1) = \{y - sn(y) | y \in C_{\pi(x)} \cap S_2, s \geq \delta_1\}$. Denote y_x is the intersection point of the radius with direction $\overrightarrow{Oy}$ and the contour $C_{\pi(x)}$.

$$\frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} \geq m\frac{\pi(y)}{\pi(y_x)} \geq m\exp\{aH\} \geq 1$$

for x large enough. That is $C^-_{\pi(x)}(\delta_1) \subset A_2(x)$. Then we denote:

$$\begin{aligned} \delta_1(x) = \inf\{\delta_1 > 0 | \text{ for any } y \in C^-_{\pi(x)}(\delta_1), \ \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} \geq 1 \\ \textbf{and} \ \text{ for any } y \in C^+_{\pi(x)}(\delta_1), \ \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} < 1 \text{ for any } \gamma \in \mathcal{Y}\} \end{aligned}$$

We know that $\delta_1(x) \to 0$ as $|x| \to \infty$.

Lemma 8.2. *If $\pi \in \mathcal{E}$, then there exists $\eta > 0$ such that $Q_\gamma(x, A(x;\gamma)) > \eta$ for any $\gamma \in \mathcal{Y}$.*

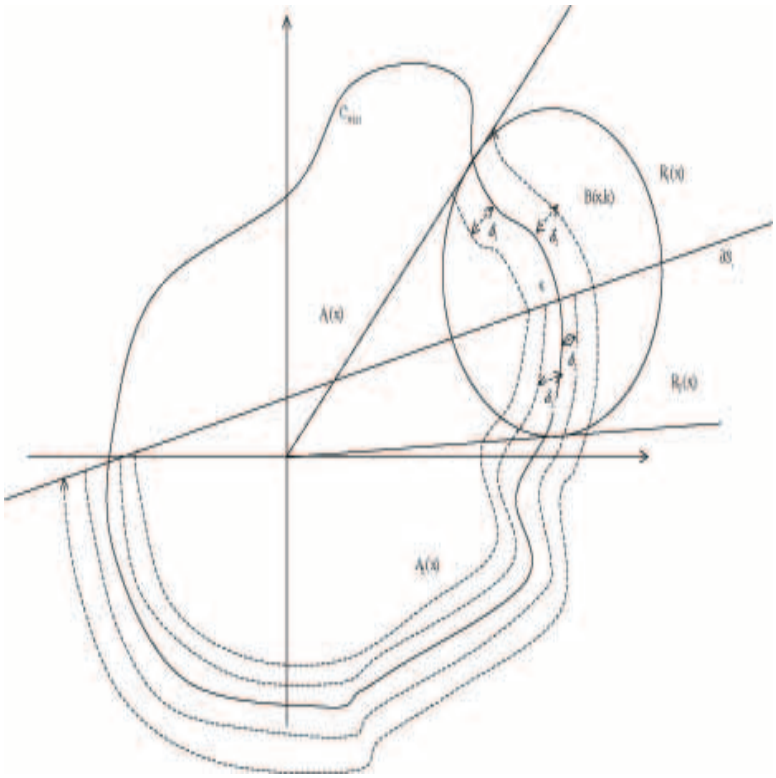

Figure 8.1: *The contour manifold $C_{\pi(x)}$ (the curved solide line), the radius δ_i-zone $C_{\pi(x)}(\delta_i)$ $i = 1, 2$ (the areas between the four curved dotted lines) and the regions $A_i(x)$ and $R_i(x)$.*

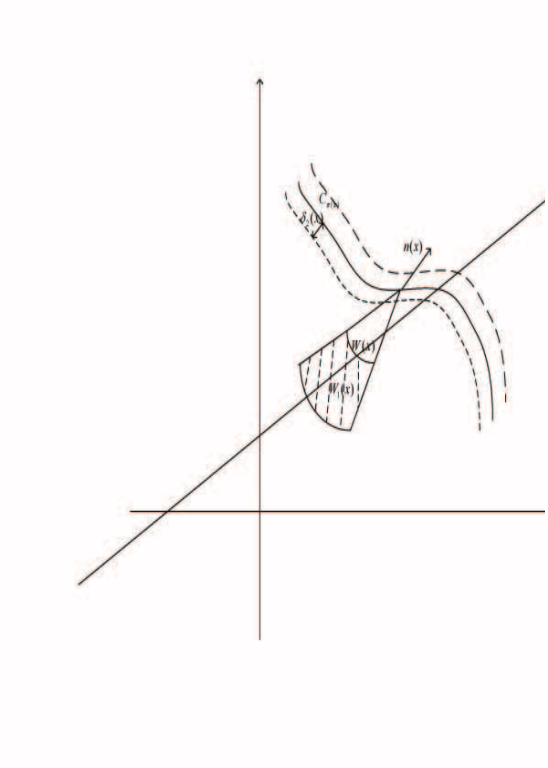

Figure 8.2: *The $\delta_2(x)-$zone and the cone $M(x)$.*

Proof. Following (8.3), we know that there exists $\beta > 0$ such that for x sufficiently large $n(x) \cdot m(x) \leq -\beta$. With this β and with fixed $K > 0$, we consider the cones (see figure 8.2):

$$W(x) = \{x - a\xi | 0 < a < K, \xi \in S^{k-1}, |\xi - n(x)| \leq \frac{\epsilon}{2}\}.$$

For x large enough that $n(y) \cdot m(y) \leq -\eta$ and $|n(x) - n(y)| < \frac{\eta}{2}$ for all $y \in W(x)$ we have for $y = x - a\xi$ in $W(x)$. Since that

$$\xi \cdot m(y) = (\xi - n(x) + n(x) - n(y) + n(y)) \cdot m(y) < \frac{\eta}{2} + \frac{\eta}{2} - \eta = 0,$$

and the Lemma 4.2 in [32], we know that $W(x) \in A(x)$. Define

$$W_1(x) = \{x - a\xi | \frac{K}{2} < a < K, \xi \in S^{k-1}, |\xi - n(x)| \leq \frac{\epsilon}{2}\}.$$

Because $\delta_2(x)$ tends to zero as $|x|$ tends to infinity, for fixed K, $W_1(x) \cap C_{\pi(x)}(\delta_2(x)) = \emptyset$.

Therefore for any $x \in S_1$ and $|x|$ large enough we can get:

$$\begin{aligned}\limsup_{\gamma\in\mathcal{Y}}\liminf_{|x|\to\infty} Q_\gamma(x, A(x;\gamma)) &\geq \limsup_{\gamma\in\mathcal{Y}}\liminf_{|x|\to\infty} Q_\gamma(x, W_1(x)) \geq \\ &\geq \min\{Q_0^{(1)}(x, W_1(x)), Q_0^{(2)}(x, W_1(x))\} = c > 0.\end{aligned}$$

The last equation is followed by the fact that $Q_0^{(2)}$, $i = 1, 2$ are both symmetric. So the $Q_0^{(i)}(x,\cdot)-$measure of $W_1(x)$ does NOT depend on x. □

Using the above lemma we can prove that:

Lemma 8.3. *Consider the kernel family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$, there exists $V : \mathbb{R} \to [1,\infty)$ such that*

$$\sup_{\gamma\in\mathcal{Y}}\limsup_{|x|\to\infty}\frac{P_\gamma V(x)}{V(x)} < 1.$$

Proof. Assume $x \in \mathcal{S}_1$, denote $p_\gamma^{(i)} = (1-\beta)[\lambda_1^{(i)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(i)}(\gamma))q_2^{(\gamma)}(x,y)] + \beta q_{whole}^{(\gamma)}(x,y)$, $i = 1, 2$. Consider $V(x) = c\pi(x)^{-\frac{1}{2}}$, where c is a constant such that $V(x) \geq 1$. and Let us compute $\frac{P_\gamma V(x)}{V(x)}$ for any $\gamma \in \mathcal{Y}$,

$$\begin{aligned}\frac{P_\gamma V(x)}{V(x)} &= \frac{\int_{\mathbb{R}^k} q_\gamma(x,y)\alpha_\gamma(x,y)\pi(y)^{-\frac{1}{2}}dy + (1-\int_{\mathbb{R}^k} q_\gamma(x,y)\alpha_\gamma(x,y)dy)c\pi(x)^{-\frac{1}{2}}}{c\pi(x)^{-\frac{1}{2}}} \\ &= \int_{A(x;\gamma)} q_\gamma(x,y)\frac{\pi(x)^{\frac{1}{2}}}{\pi(y)^{\frac{1}{2}}}dy + \int_{R(x;\gamma)} q_\gamma(x,y)[1 - \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} + \frac{\pi(y)^{\frac{1}{2}}q_\gamma(y,x)}{\pi(x)^{\frac{1}{2}}q_\gamma(x,y)}]dy \\ &= \int_{A_1(x;\gamma)} p_\gamma^{(1)}(x,y)\frac{\pi(x)^{\frac{1}{2}}}{\pi(y)^{\frac{1}{2}}}dy + \int_{A_2(x;\gamma)} p_\gamma^{(1)}(x,y)\frac{\pi(x)^{\frac{1}{2}}(p_\gamma^{(1)}(x,y))^{\frac{1}{2}}}{\pi(y)^{\frac{1}{2}}(p_\gamma^{(2)}(x,y))^{\frac{1}{2}}} \times \frac{(p_\gamma^{(2)}(x,y))^{1/2}}{(p_\gamma^{(1)}(x,y))^{1/2}}dy \\ &+ \int_{R(x;\gamma)} p_\gamma^{(1)}dy - \int_{R_1(x)} p_\gamma^{(1)}(x,y)\frac{\pi(y)}{\pi(x)}dy - \int_{R_2(x)} p_\gamma^{(1)}\frac{\pi(y)p_\gamma^{(2)}(x,y)}{\pi(x)p_\gamma^{(1)}(x,y)}dy \\ &+ \int_{R_1(x)} p_\gamma^{(1)}(x,y)\frac{\pi(y)^{\frac{1}{2}}}{\pi(x)^{\frac{1}{2}}}dy + \int_{R_2(x)} p_\gamma^{(1)}(x,y)\frac{\pi(y)^{\frac{1}{2}}p_\gamma^{(2)}(x,y)^{\frac{1}{2}}}{\pi(x)^{\frac{1}{2}}p_\gamma^{(1)}(x,y)^{\frac{1}{2}}} \times \frac{(p_\gamma^{(2)}(x,y))^{\frac{1}{2}}}{(p_\gamma^{(1)}(x,y))^{\frac{1}{2}}}dy\end{aligned}$$

$$\begin{aligned}&= \int_{A_1(x)} p_\gamma^{(1)}(x,y)\min\left\{1, \frac{\pi(x)^{\frac{1}{2}}}{\pi(y)^{\frac{1}{2}}}\right\}dy + \int_{A_2(x)} \Pi_{i=1}^2(p_\gamma^{(i)}(x,y))^{\frac{1}{2}}\min\left\{1, (\frac{\pi(x)p_\gamma^{(1)}(x,y)}{\pi(y)p_\gamma^{(2)}(x,y)})^{\frac{1}{2}}\right\}dy \\ &- \int_{R_1(x)} p_\gamma^{(1)}(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}dy - \int_{R_2(x)} p_\gamma^{(1)}(x,y)\min\left\{1, \frac{\pi(y)p_\gamma^{(2)}(x,y)}{\pi(x)p_\gamma^{(1)}(x,y)}\right\}dy \\ &+ \int_{R_1(x)} p_\gamma^{(1)}(x,y)\min\left\{1, \frac{\pi(y)^{\frac{1}{2}}}{\pi(x)^{\frac{1}{2}}}\right\}dy + \int_{R_2(x)} (\Pi_{i=1}^2 p_\gamma^{(i)})\min\left\{1, \left(\frac{\pi(y)p_\gamma^{(2)}(x,y)}{\pi(x)p_\gamma^{(1)}}\right)^{\frac{1}{2}}\right\}dy \\ &+ \int_{R(x)} p_\gamma^{(1)}(x,y)dy.\end{aligned}$$

Step 1: For any $\eta > 0$,there exists $K > 0$ which is independent with the choice of x such that each of the first six integrals outside the ball $B(x;K)$ are less than $\frac{\eta}{18}$. For example, for the sixth term

$$\begin{aligned}\int_{R_2(x)\cap B(x;K)^c}(\Pi_{i=1}^2 p_\gamma^{(i)}(x,y)^{\frac{1}{2}})\min\left\{1,\left(\frac{\pi(y)p_\gamma^{(2)}(x,y)}{\pi(x)p_\gamma^{(1)}(x,y)}\right)^{\frac{1}{2}}\right\}dy &\leq \int_{R_2(x)\cap B(x;K)^c}\Pi_{i=1}^2 p_\gamma^{(i)}(x,y)^{\frac{1}{2}}dy\\ &\leq \int_{B(0;K)^c}\max\left\{q_0^{(1)}(z),q_0^{(2)}(z)\right\}dz\\ &\leq \frac{\eta}{18} \text{ if } K \text{ is large enough.}\end{aligned}$$

Step 2: Suppose $q_\gamma(x,y) \leq E$ for $i = 1,2$ for any $y \in B(x;K)$ and $\gamma \in \mathcal{Y}$. Then for fixed η, when $|x|$ is large enough, there exists $\delta_2 > \delta_1(x)$ such that for any $y \in C^+_{\pi(x)}(\delta_2)\cap R(x)$ we have $\frac{\pi(y)p_\gamma^{(2)}(y,x)}{\pi(x)p_\gamma^{(1)}(x,y)} \leq \min\{\frac{\eta}{18E},[\frac{\eta}{18E}]^2\}$ and for any $y \in C^-_{\pi(x)}(\delta_2)\cap A(x)$ we have $\frac{\pi(x)q_\gamma(x,y)}{\pi(y)q_\gamma(y,x)} \leq \min\{\frac{\eta}{18E},\left[\frac{\eta}{18E}\right]^2\}$. Then we can define:

$$\begin{aligned}\delta_2(x) = \inf\{\delta_2 \geq \delta_1(x) | &\text{for any } y \in C^+_{\pi(x)}(\delta_2)\cap R(x) \text{ we have } \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)} \leq \min\left\{\frac{\eta}{18E},\left[\frac{\eta}{18E}\right]^2\right\}\\ &\text{for any } y \in C^-_{\pi(x)}(\delta_2)\cap A(x) \text{ we have } \frac{\pi(x)q_\gamma(x,y)}{\pi(y)q_\gamma(y,x)} \leq \min\left\{\frac{\eta}{18L},\left[\frac{\eta}{18L}\right]^2\right\} \text{ for any } \gamma \in \mathcal{Y}\}.\end{aligned}$$

Then there exists $N > 0$, such that for any x with $|x| > N$ the first six integrals which are outside the ball of radius $\delta_2(x)$ and inside in any ball $B(x,K)$ will be less than $\frac{\epsilon}{18}$. For example,

$$\begin{aligned}&\int_{R_2(x;\gamma)\cap B(x;K)\cap C^+_{\pi(x)}(\delta_2(x))}\Pi_{i=1}^2(p_\gamma^{(i)})^{\frac{1}{2}}\times\min\left\{1,\left(\frac{\pi(y)p_\gamma^{(2)}(x,y)}{\pi(x)p_\gamma^{(1)}(x,y)}\right)^{\frac{1}{2}}\right\}dy\\ \leq &\int_{R_2(x)\cap B(x;K)\cap C^+_{\pi(x)}(\delta_2(x))}\Pi_{i=1}^2(p_\gamma^{(i)})^{\frac{1}{2}}\frac{\eta}{18E}dy\\ \leq &\int_{R_2(x)\cap B(x;K)\cap C^+_{\pi(x)}(\delta_2(x))}E\frac{\eta}{18E}dy\\ \leq &\frac{\eta}{18}.\end{aligned}$$

Step 3: Since $\delta_2(x) \to 0$ as $|x| \to \infty$, for the fixed K and $\eta > 0$ in step 1, there exists N_1 large enough such that for any $|x| > N_1$ we have:

$$\mu^{Leb}(C_{\pi(x)}(\delta_2(x))\cap B(x,K)) \leq \frac{\eta}{18E}.$$

Then we have the first six integrations which are inside the radius $\delta_2(x)-$zone in any ball $B(x,K)$ will be less than $\frac{\eta}{18}$ too. Therefore following the above analysis, we know that:

$$\limsup_{|x|\to\infty}\frac{P_\gamma V(x)}{V(x)} \le \eta + \limsup_{|x|\to\infty} Q_\gamma(x, R(x;\gamma)).$$

Since η is small enough, we only need to prove that:

$$\limsup_{|x|\to\infty} Q_\gamma(x, R(x;\gamma)) = 1 - \liminf_{|x|\to\infty} Q_\gamma(x, A(x;\gamma)) < 1.$$

The last inequality is followed by the Lemma 8.3. □

Lemma 8.4. *Suppose $q(z)$ is the $k-$dimension Gaussian distribution with variance matrix Σ such that $\epsilon I \le \Sigma \le LI < (L+\rho)I$ where $\rho > 0$. Then there exists $R > 0$ such that for any $|z| > R$, we have $q(z) < q_0(z)$, where $q_0(z)$ is the $k-$dimensional Gaussian distribution with variance matrix $(L+\rho)\cdot I$.*

Proof. We need to find z such that

$$\begin{aligned}\frac{q(z)}{q_0(z)} &= \frac{\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}z^t\Sigma^{-1}z}}{\frac{1}{(2\pi)^{\frac{k}{2}}|(L+\rho)I|^{\frac{1}{2}}}e^{-\frac{1}{2}z^t((L+\rho)I)^{-1}z}} = \\ &= \frac{|(L+\rho)I|^{1/2}}{|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}z^t(\Sigma^{-1}-((L+\rho)I)^{-1})z} \le 1.\end{aligned}$$

If we denote $c = 2\frac{|(L+\rho)I|^{1/2}}{|\Sigma|^{\frac{1}{2}}}$, then we have $z^t(\Sigma^{-1}-(LI)^{-1})z < c$. Since $(\Sigma^{-1}-(LI)^{-1})$ is positive definite, $\{z : z^t(\Sigma^{-1}-(LI)^{-1})z < c\}$ is the interior of an ellipsoid and its longest semi-axis is $\frac{1}{\lambda_0\sqrt{c}}$ where λ_0 is the smallest eigenvalue of $(\Sigma^{-1}-((L+\rho)I)^{-1})$. We know that $\lambda_0 > \frac{1}{L}-\frac{1}{L+\rho} = \frac{\rho}{L(L+\rho)}$, therefore the longest semi axis less than $\frac{L(L+\rho)}{\rho\sqrt{c}}$. So we can set $R = \frac{L(L+\rho)}{\rho\sqrt{c}}$ to satisfy the conclusion of the lemma. □

Recall that the distance between two $n\times n$ matrices M_1 and M_2 can be defined as $\|M_1 - M_2\| = \max\{|(M_1)_{ij} - (M_2)_{ij}||1 \le i,\ j \le n\}$. Then we have:

Lemma 8.5. $\|C^{(t+1)} - C^{(t)}\| \to 0$ *and* $\|C_i^{(t+1)} - C_i^{(t)}\| \to 0$ *as t tends to infinity.*

Proof. To prove that $\|C^{(t+1)} - C^{(t)}\| \to 0$ as t tends to infinity, obviously we only need to check $\|cov(y_0, y_1, \cdots, y_t) - cov(y_0, y_1, \cdots, y_{t-1})\| \to 0$ as $t \to 0$. We denote $\tilde{C}^{(t)} = cov(y_0, y_1, \cdots, y_{t-1})$. Then we have the following the recursion formula:

$$\tilde{C}^{(t+1)} = \frac{t-1}{t}\tilde{C}^{(t)} + \frac{1}{t}(t\bar{y}_{t-1}\bar{y}_{t-1}^T - (t+1)\bar{y}_t\bar{y}_t^T + y_t y_t^T),$$

where $\bar{y}_k = \frac{\sum_{i=0}^k y_i}{k+1}$. So

$$\begin{aligned} & \|\tilde{C}^{(t+1)} - \tilde{C}^{(t)}\| = \\ = \ & \|\frac{1}{t}\tilde{C}^{(t)} - \frac{1}{t}(t\bar{y}_{t-1}\bar{y}_{t-1}^T - (t+1)\bar{y}_t\bar{y}_t^T + y_t y_t^T\| \leq \\ \leq \ & \|\frac{1}{t}\tilde{C}^{(t)}\| + \|\bar{y}_{t-1}\bar{y}_{t-1}^T - \frac{t+1}{t}\bar{y}_t\bar{y}_t^T\| + \frac{1}{t}\|y_t y_t^T\|. \end{aligned}$$

Recall that $\tilde{C}^{(t)} = \frac{1}{t}(\sum_{i=0}^{t-1} y_i y_i^T - (t+1)\bar{y}_{t-1}\bar{y}_{t-1}^T)$, following the fact that $|y_i| \leq B + n^\kappa,\ 0 < \kappa < \frac{1}{2}$, we have $\|\frac{1}{t}\tilde{C}^{(t)}\| \to 0$ as $t \to \infty$. Similarly $\frac{1}{t}\|y_t y_t^T\| \to 0$ as $t \to \infty$.

Regarding the second term we have:

$$\begin{aligned} & \|\bar{y}_{t-1}\bar{y}_{t-1}^T - \frac{t+1}{t}\bar{y}_t\bar{y}_t^T\| = \\ = \ & \|\bar{y}_{t-1}\bar{y}_{t-1}^T - \frac{t+1}{t}(\bar{y}_{t-1}\bar{y}_{t-1}^T + \frac{y_t\bar{y}_{t-1}^T}{t} + \frac{\bar{y}_{t-1}y_t^T}{t} + \frac{y_t y_t^T}{t^2}\| \leq \\ \leq \ & \frac{1}{t}\|\bar{y}_{t-1}\bar{y}_{t-1}^T\| + \frac{t+1}{t^2}\|y_t\bar{y}_{t-1}^T\| + \frac{t+1}{t^2}\|\bar{y}_{t-1}y_t^T\| + \frac{t+1}{t^3}\|y_t y_t^T\|. \end{aligned}$$

Using $|y_i| \leq B + n^\kappa,\ 0 < \kappa < \frac{1}{2}$, we can check that each term in above formula tends to zero as t tends infinity. Therefore we have $\|C^{(t+1)} - C^{(t)}\| \to 0$ as t tends to infinity. Similarly we have and $\|C_i^{(t+1)} - C_i^{(t)}\| \to 0$ as $t \to \infty$. □

Lemma 8.6. *The Diminishing Adaptation condition holds for the MRAPT algorithm.*

Proof. We denote $r_\gamma(x,y) = \frac{(1-\beta)[\lambda_1^{(2)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda_1^{(2)}(\gamma))q_2^{(\gamma)}(x,y)] + \beta q_{whole}^{(\gamma)}(x,y)}{(1-\beta)[\lambda_1^{(2)}(\gamma)q_1^{(\gamma)}(x,y) + (1-\lambda^{(\gamma)}{}_2)q_2^{(\gamma)}(x,y)] + \beta q_{whole}^{(\gamma)}(x,y)}$. For any $x \in S_1$ and $A \in \mathcal{B}(\mathcal{X})$, we have:

$$\begin{aligned} P_{\Gamma_k}(x,A) = & \int_{A\cap\mathcal{S}_1} p_{\gamma_k}^{(1)}(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}dy + \int_{A\cap\mathcal{S}_2} p_{\gamma_k}^{(1)}(x,y)\min\left\{1, r_{\gamma_k}(x,y)\frac{\pi(y)}{\pi(x)}\right\}dy \\ + \ & \delta_x(A)[\int_{\mathcal{S}_1} p_{\gamma_k}^{(1)}(x,y)[1 - \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}dy + \int_{\mathcal{S}_2} p_{\gamma_k}^{(1)} r_{\gamma_k}(x,y)[1 - r_{\gamma_k}(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}dy. \end{aligned}$$

Denote the first term $I_k(x,A)$, the second term $II_k(x,A)$, the third term $III_k(x,A)$ and the fourth term $IV_k(x,A)$. Then we have:

$$\begin{aligned}|P_{\Gamma_{k+1}}(x,A)-P_{\Gamma_k}(x,A)| &\leq |I_{\Gamma_{k+1}}(x,A)-I_{\Gamma_k}(x,A)|+|II_{\Gamma_{k+1}}(x,A)-II_{\Gamma_k}(x,A)|\\ &+ |III_{\Gamma_{k+1}}(x,A)-III_{\Gamma_k}(x,A)|+|IV_{\Gamma_{k+1}}(x,A)-IV_{\Gamma_k}(x,A)|.\end{aligned}$$

Since for any n, we have $\epsilon I \leq C_1^{(n)}, C_2^{(n)}, C^{(n)} \leq LI$, following the Lemma 8.2, for any $\eta > 0$, there exists $R>0$ large enough such that $\int_{B(x,R)^c\cap\mathcal{S}_2} p_{\gamma_k}^{(1)}(x,y)\min\{1, r_{\gamma_k}(x,y)\frac{\pi(y)}{\pi(x)}\}dy < \eta$, where $B(x,R)$ is the ball centered at x with radius R, and $q_{\gamma_k}^{(1)}$ is bounded inside $B(x,R)$. We denote $\alpha_2^{(k)} = \min\{1, r_{\gamma_k}(x,y)\frac{\pi(y)}{\pi(x)}\}$. Then we have

$$\begin{aligned}|II_{\Gamma_{k+1}}(x,A)-II_{\Gamma_k}(x,A)| &\leq \int_{A\cap\mathcal{S}_2\cap B(x,R)} |p_{\gamma_{k+1}}^{(1)}(x,y)\alpha_2^{(k+1)}(x,y)-p_{\gamma_k}^{(1)}(x,y)\alpha_2^{(k)}(x,y)|dy\\ &+ \int_{A\cap\mathcal{S}_2\cap B(x,R)^c} |p_{\lambda_{k+1}}^{(1)}(x,y)\alpha_2^{(k+1)}(x,y)-p_{\lambda_k}^{(1)}(x,y)\alpha_2^{(k)}(x,y)|dy\\ &\leq \int_{A\cap\mathcal{S}_2\cap B(x,R)} |p_{\gamma_{(k+1)}}^{(1)}(x,y)\alpha_2^{(k+1)}(x,y)-p_{\gamma_{k+1}}^{(1)}(x,y)\alpha_2^{(k)}(x,y)\\ &+ p_{\gamma_{k+1}}^{(1)}(x,y)\alpha_2^{(k)}(x,y)-p_{\gamma_k}^{1}(x,y)\alpha_2^{(k)}(x,y)|dy+\eta\\ &\leq \int_{A\cap\mathcal{S}_2\cap B(x,R)} p_{\gamma_{k+1}}^{1}(x,y)|\alpha_2^{(k+1)}(x,y)-\alpha_2^{(k)}(x,y)|dy\\ &+ \int_{A\cap\mathcal{S}_2\cap B(x,R)} \alpha_2^{(k)}(x,y)|p_{\gamma_{k+1}}^{(1)}(x,y)-p_{\gamma_k}^{(1)}(x,y)|dy+\eta\\ &\leq \int_{A\cap\mathcal{S}_2\cap B(x,R)} \frac{\pi(y)p_{\gamma_{k+1}}^{(1)}(x,y)}{\pi(x)}|r_{\gamma_{k+1}}(x,y)-r_{\gamma_k}(x,y)|dy\\ &+ \int_{A\cap\mathcal{S}_2\cap B(x,R)} |p_{\gamma_{k+1}}^{(1)}(x,y)-p_{\gamma_k}^{(1)}(x,y)|dy.\end{aligned}$$

For fixed x, we suppose $\frac{\pi(y)p_{\gamma_{k+1}}^{(1)}(x,y)}{\pi(x)} \leq B_x$ for any $y\in B(x,R)$, then the first term less than $B_x\int_{A\cap\mathcal{S}_2\cap B(x,R)}|r_{\gamma_{k+1}}(x,y)-r_{\gamma_k}(x,y)|dy$. It suffices to prove that $|r_{\gamma_{k+1}}(x,y)-r_{\gamma_k}(x,y)|$ tends to zero in probability as k tends infinity. Following the Lemma 8.5, we have $|q_i^{(\gamma_{n+1})}(x,y)-q_i^{(\gamma_n)}(x,y)|\to 0,\ i=1,2$ and $|q_{whole}^{(\gamma_{n+1})}(x,y)-q_{whole}^{(\gamma_n)}(x,y)|\to 0$ as n tends to infinity. Secondly consider the random variable $d_n=(X_{n+1}-X_n)^2$, based on the fact that $\epsilon I \leq C_{(n)}^{(i)}, C_{(n)} \leq LI$, we have for any $\eta>0$, there exists $M>0$ such that

$$P((X_{n+1}-X_n)^2>M) \leq \max\left\{\int_{|z|>M} N_{\epsilon I}(z)dz, \int_{|z|>M} N_{LI}(z)dz\right\} \leq \eta.$$

Using the similar proof of the Lemma 4.2 in [14], $|\lambda_1^{(i)}(n+1)-\lambda_1^{(i)}(n)| \to 0$ in probability as n tends to infinity. Therefore, it is easy to check that $|r_{\gamma_{k+1}}(x,y) - r_{\gamma_k}(x,y)|$ tends to zero in probability as k tends infinity. Similarly, we can prove $\int_{A\cap\mathcal{S}_2\cap B(x,R)} |p^{(1)}_{\gamma_{k+1}}(x,y) - p^{(1)}_{\gamma_k}(x,y)|dy$ tends to zero too. So $|II_{\Gamma_{k+1}}(x,A)-II_{\Gamma_k}(x,A)| \to 0$ in probability. Obviously same conclusions hold for terms I, III and IV. So we have proved the Diminishing Adaptation condition for the MRAPT algorithm. □

8.4 The Proof Of Theorem 8.1

Now we can prove the theorem 8.1 using the Theorem 4.4.

Proof. Following theorem 4.4 and lemma 8.6, we only need to check the simultaneously strongly aperiodically geometrically ergodic conditions for the kernel family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$. Consider any compact set $B(r) = \{x \in \mathbb{R}^k || x| \le r\}$ and denote $q_{01}(x,y)$ is the Gaussian distribution with variance matrix ϵI_k and mean x and $q_{02}(x,y)$ is the Gaussian distribution with variance matrix LI_k and mean x. Since $\pi(x)$ is continuous and positive, we can define $d_r = \sup_{x\in B_r}\pi(x) < \infty$ and $\epsilon_r = \min\{\inf_{x,y\in B(r)} q_{01}(x,y), \inf_{x,y\in B(r)} q_{02}(x,y)\} > 0$. Obviously, we have $q_\gamma(x,y) \ge \epsilon$ for any $x, y \in B(r)$ and $\gamma \in \mathcal{Y}$. Then for any $x \in B(r)$ and $E \subseteq B_r$, we have

$$\begin{aligned} P_\gamma(x,B) &\ge \int_{R(x;\gamma)\cap B(r)} \frac{\pi(y)q_\gamma(y,x)}{\pi(x)}\mu^{Leb}(dy) + \int_{A(x;\gamma)\cap B(r)} q_\gamma(x,y)\mu^{Leb}(dy) \ge \\ &\ge \frac{\epsilon}{d}\int_{R(x;\gamma)\cap B(r)} \pi(y)\mu^{Leb}(dy) + \frac{\epsilon}{d}\int_{A(x;\gamma)\cap B(r)} \pi(y)\mu^{Leb}(dy) = \\ &= \frac{\epsilon}{d}\pi(B(r)). \end{aligned}$$

Thus $B(r)$ is small and we have $P_\gamma(x,E) \ge \delta_r\nu_r(E)$, where $\delta_r = \frac{\epsilon\pi(B(r))}{d_r}$ and $\nu_r(\cdot) = \frac{\pi(\cdot)}{\pi(B(r))}$ is a probability measure on $B(r)$. Lemma 8.3 indicates that there exists r_0 and $0 < \rho < 1$such that for any $|x| > r_0$, we have $\sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < \rho$, where $V(x) = c\pi(x)^{-\frac{1}{2}}$. Now let $C = B_{r_0}$, $b = \max\{V(x)|x \in C\}$. We get kernel family $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is simultaneously strongly aperiodically geometrically ergodic. □

Remark: Based on the above construction of empirical covariance matrices from all the historical simulations and the proof of the Theorem 4.4, we actually extended the HST adaptive algorithm from compact state space to general state space. We can observe that our empirical covariance matrices up to time t do not come from all the history, but from part of them which are bounded by $B + t^\tau$ or $B + N_i(t)$. It seems that we miss some information from the samples which are out of $B + t^\tau$ or $B + N_i(t)$. However we know that $B + t^\tau$ or $B + N_i(t)$ both tend to infinity as t tends to infinity, therefore the loss will become less and less when t increases.

8.5 Examples

In this section we will simulate some toy examples to verify our analysis before and check the main theoretical results. More precisely we will make comparisons on the following three aspects:

1. The efficiency of the MRAPT, the Dual RAPT and the HST algorithms to detect different models in the case of two models being far way;

2. The number of switches between differen models of the modified MRAPT and the HST algorithms in the case of two models being close;

3. The difference between running a single modified MRAPT and several parallel modified MRAPT.

Now we consider a mixture of two Gaussian distributions with equal weights as our target distribution and the state space is the whole space $\mathbb{R}^{10}$ which is not compact. We let

$$\pi(x) = 0.5 \times N(\mu_1, \Sigma_1) + 0.5 \times N(\mu_2, \Sigma_2),$$

where μ_i are ten dimensional vectors and $\Sigma_i = \sigma_i \begin{pmatrix} 1 & \rho_i & \rho_i & \cdots & \rho_i \\ \rho_i & 1 & \rho_i & \cdots & \rho_i \\ \rho_i & \rho_i & 1 & \cdots & \rho_i \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_i & \cdots & \cdots & \rho_i & 1 \end{pmatrix}_{10\times 10}$, $i = 1, 2$. Since any Gaussian distribution lies in $\mathcal{E}$, following the Theorem 4.4 in Jarner and Hansen [32],1998, we know that $\pi \in \mathcal{E}$. Then using the Theorem 8.1 we know that the Modified MRAPT should be ergodic. The HST algorithm is ergodic too based on the main results of [32].

For fair comparison we will consider six cases of target distributions with different mean and covariance matrices. Let us consider the following scenarios:

Scenario A: $\rho_1 = 0.2,\ \rho_2 = 0.3,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 3,\ \mu_{2j} = -3,\ 1 \le j \le 10.$

Scenario B: $\rho_1 = 0.2,\ \rho_2 = 0.3,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 0.5,\ \mu_{2j} = -0.5,\ 1 \le j \le 10.$

Scenario C: $\rho_1 = -0.1,\ \rho_2 = 0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 3,\ \mu_{2j} = -3,\ 1 \le j \le 10.$

Scenario D: $\rho_1 = 0.1,\ \rho_2 = -0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 3,\ \mu_{2j} = -3,\ 1 \le j \le 10.$

Scenario E: $\rho_1 = -0.1,\ \rho_2 = 0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 1,\ \mu_{2j} = -1,\ 1 \le j \le 10.$

Scenario F: $\rho_1 = 0.1,\ \rho_2 = -0.1,\ \frac{\sigma_1}{\sigma_2} = \frac{1}{3},\ \mu_{1j} = 1.5,\ \mu_{2j} = -1.5,\ 1 \le j \le 10.$

We adjust the difficulty of the inter-model transitions by changing the distance between μ_1 and μ_2. Meanwhile we try to vary the shape of both models by using different covariance matrices, like scenarios C&D and E&F. We let the partition be $\mathcal{S}_1 = \sum_{i=1}^{10} x_i \le 0$ and $\mathcal{S}_2 = \sum_{i=1}^{10} x_i > 0$.

We first consider the scenarios A, C and D in which the two models are far away. We draw the histograms of the first two and the last two coordinates, i.e. x_1, x_2, x_9, x_{10}

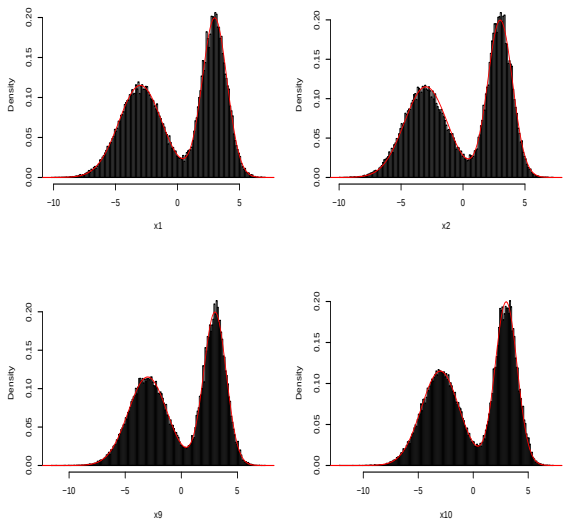

Figure 8.3: *Scenario A: The simulations of the first two coordinates and the last two coordinates with mixed RAPT after 50,000 iterations. The red curve is the true density function.*

with the true marginal density to observe the performance of the algorithms. Now let us consider Scenario A. We try the HST algorithm for 50,000 iterations and show the histograms in Figure 8.5. We notice that the performance in the second region is not good even when we choose the initial values in this region. Similarly the dual RAPT is not efficient to switch the models either, and the histograms of x_1, x_2, x_9, x_{10} are showed in Figure 8.4. However the mixed RAPT algorithm has a much better performance in Scenario A. After 50,000 iterations, the parameters are $\lambda_1^{(1)}(50,000) = 0.681$ and $\lambda_1^{(2)}(50,000) = 0.353$ and the histograms of the first two coordinations and the last two coordinations are presented in Figure 8.3. Similarly neither HST algorithm nor Dual RAPT algorithm can switch the models fluently in Scenario C and Scenario D. It seems that HST algorithm is very easy to get stuck in the $\mathcal{S}_1$ and Dual RAPT algorithm is hard to jump out of the region $\mathcal{S}_2$. Even though we vary the initial values of the covariance matrices the results do not become much better. In Scenario C the Mixed RAPT still has a very good performance as in Scenario A when we select initial values

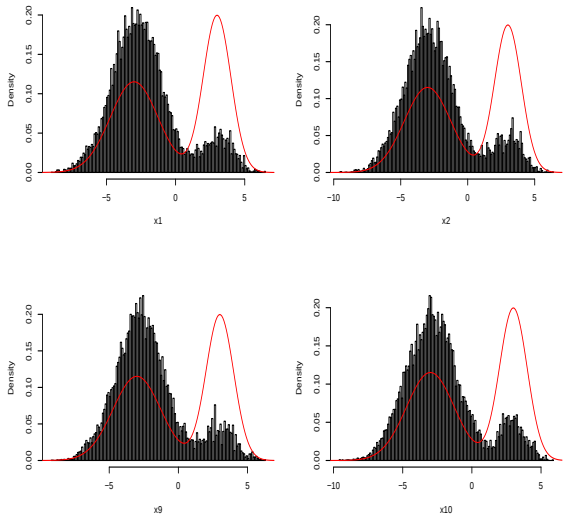

Figure 8.4: *Scenario A: The simulations of the first two coordinates and the last two coordinates using the dual RAPT algorithm after 50,000 iterations.The red curve is the true density function.*

randomly. However in scenario D running a single mixed RAPT algorithm with the starting value $x_0 = (0, \ldots, 0)^T$, $\beta = 0.3$ and $\Sigma_{whole} = diag(10, \ldots, 10)$ the algorithm does not detect both models. So we increase the "detection" log by using the initial $\Sigma_{whole} = diag(25, \ldots, 25)$, then the performance of Mixed RAPT is illustrated in Figure 8.10. We note that it is important for the initial variances of q_{whole} to be large enough so that both modes are visited during the initialization period. Another strategy to improve the detection efficiency is to run some parallel Mixed RAPT. For more details see R.Craiu, J.Rosenthal, and C.Yang [14]. Here we run five parallel chains together with 10,000 iterations. The initial value for the $i-th$ chain is $x_{i,0} = (3-i, 3-i, \ldots, 3-i)^T$ for $1 \leq i \leq 5$ and let $\beta = 0.2$. The initial values Σ_i for the Gaussian proposals q_i, $i = 1, 2$ and the covariance matrix Σ_{whole} of q_{whole} are the identity matrices, i.e. $\Sigma_1 = \Sigma_2 = \Sigma_{whole} = I$. The histograms of the first two coordinates and the last two coordinates are shown in Figure 8.9. We notice that there are much more freedom to choose the initial values of the parallel MRAPT.

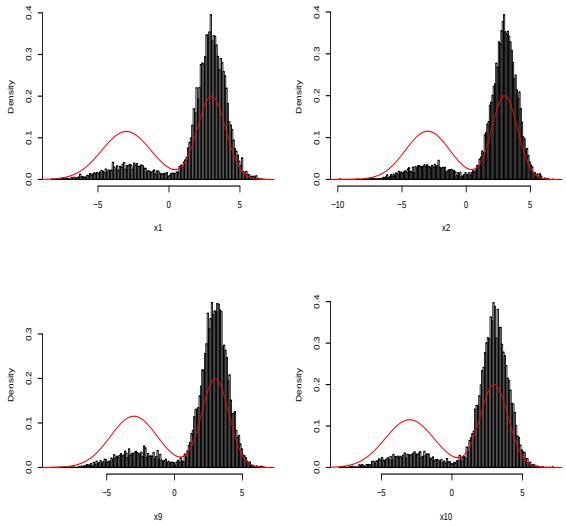

Figure 8.5: *Scenario A: The simulations of the first two coordinates and the last two coordinates with the HST algorithm after 50,000 iterations.The red curve is the true density function.*

Secondly we consider the Scenarios B, E, and F. In these cases both modes are close, therefore it is not hard to detect all the models even by the HST algorithm. Figure 8.7 shows the simulation results of Scenario B after 50,000 iterations. As anticipated the mixed RAPT simulates the Scenario B very well, which can be seen in Figure 8.6. In these cases we will compare the number of mode switches for both the modified MRAPT and HST algorithm. We show the inter-model switch times of both algorithms in Figure 8.8 in the case of Scenario B. We observe that the modified MRAPT switches modes more efficiently. Similar result also happens when we analyze the switch times (Figure 8.11) of both algorithms for Scenario E.

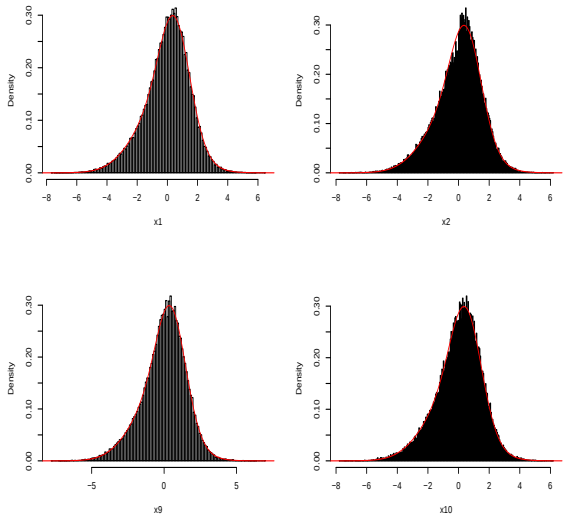

Figure 8.6: *Scenario B: Histograms of the first two coordinates and the last two coordinates using mixed RAPT after 50,000 iterations.The red curve is the true density function.*

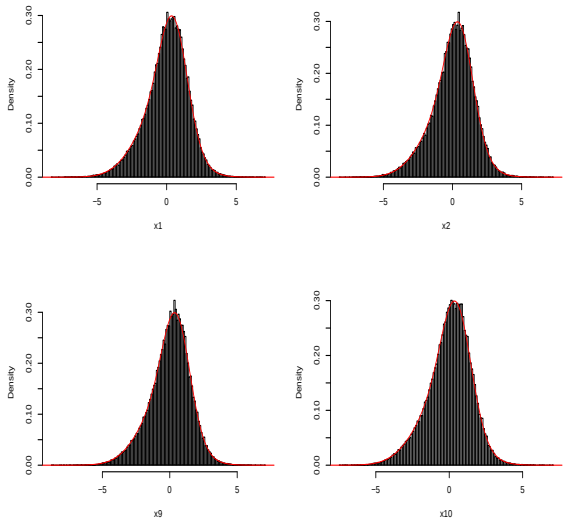

Figure 8.7: *Scenario B: Histograms of the first two coordinates and the last two coordinates using the HST agorithm after 50,000 iterations.The red curve is the true density function.*

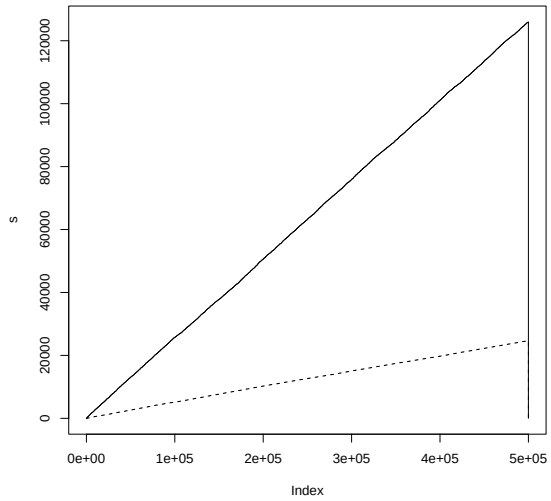

Figure 8.8: *Scenario B: Number of switches for the HST algorithm (dashed line) and for the mixed RAPT (solid line).*

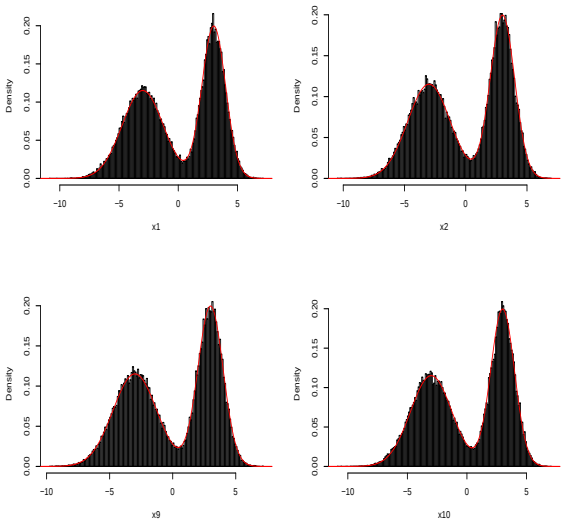

Figure 8.9: *Scenario D: The simulations of the first two coordinates and the last two coordinates with the five parallel MRAPT chain after 500,000 iterations. The red curve is the true density function.*

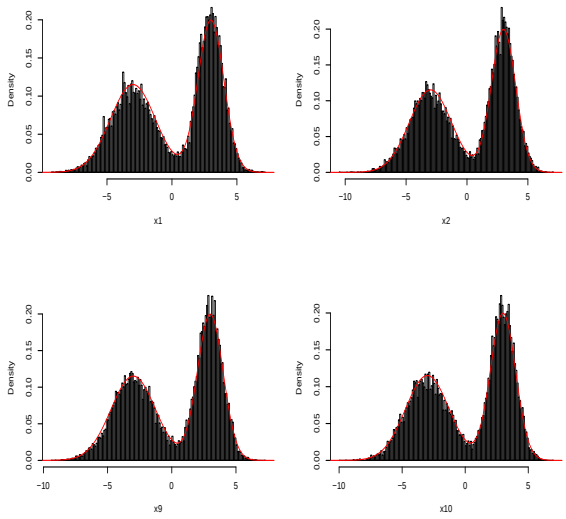

Figure 8.10: *Scenario D: The histograms of the first two coordinates and the last two coordinates using Mixed RAPT after 500,000 iterations.The red curve is the true density function.*

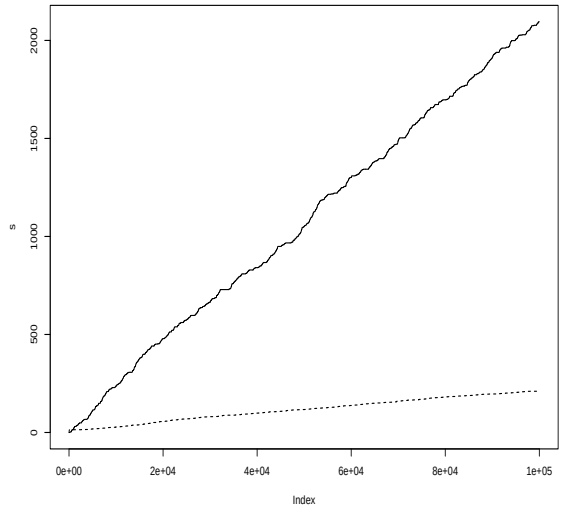

Figure 8.11: *Scenario E: The switch times of MRAPT versus HST after 100,000 iterations.*

Chapter 9

Conclusions and Further Research

Our first result focuses on the proof of CLT for uniformly ergodic Markov chain using regeneration methods. Actually under the condition of the Theorem 5.1, we can define the regeneration time for the common small set of all the kernels $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$. The future research direction is how to extend the regeneration proof to explore the CLT of adaptive MCMC under the same conditions as in Theorem 5.1.

There are quite a few conditions and conclusions in chapter 5 and chapter 6. To compare all of these, please see Table 8.1. One possible future research is to find out the sufficient and necessary condition of AMCMC's ergodicity under condition (a)(see some related reference Bai,Roberts and Rosenthal [7]).

Another possible direction is to explore some weaker conditions than those in the Theorem 5.1 to ensure the ergodiciy, or to prove the open problem 20 in in Roberts and Rosenthal [48] directly. And since the condition: $\{V(X_n)\}_{n=0}^{\infty}$ bounded in probability is hard to check in practice, we should look for other equivalent conditions which are easy to verify so that we can close some of the gaps between theory and practice.

Regarding the WLLN of AMCMC, we have also proved the WLLN for bound functions under the conditions of the Theorem 5.1. The Theorem 6.3 could be extended to non-compact state space with the super-exponential target distribution.

condition 1	condition 2	condition 3		conclusion
condition (a)	condition (b)		$\Rightarrow$	Ergodicity of AMCMC WLLN for bounded function
condition (a)	condition (b)		$\not\Rightarrow$	WLLN for unbounded function
condition (a)	condition (b')		$\Rightarrow$	WLLN for unbounded function
condition (b)	condition (d)		$\Rightarrow$	Ergodicity of AMCMC
condition (b)	Ergodicity of AMCMC		$\not\Rightarrow$	condition (d)
condition (b)	condition (d_1)		$\not\Rightarrow$	Ergodicity of AMCMC
condition (b)	condition (d_2)		$\Rightarrow$	Ergodicity of AMCMC
condition (e)	condition (b)	condition (f)	$\Rightarrow$	Ergodicity of AMCMC WLLN for bounded function

Table 9.1: *Main results of Chapter 3,4 and 5.*

Regarding the RAPT and MRAPT algorithms, intuitively we can generalize the regional adaptive algorithms to the cases with more than two regions. However it is difficult to make sure that the MRAPT algorithm visits each region often enough when there are too many regions. More precisely, we hope to visit the different regions with differen frequencies because the weight of each model may be different. How to design more efficient regionally adaptive algorithm for more regions based on our current work is one of the possible directions for our future research.

Bibliography

[1] I. A.Ibragimov and Y. V.Linnik. *Independent and stationary sequences of random variables.* Wolter-Noordhoff, Groiningen., 1971.

[2] C. Andrieu and Y. F. Atchade. On the efficiency of adaptive MCMC algorithms. *Electronic Communications In Probability*, 12:336–349, 2007.

[3] C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Annals of Applied Probability*, 16:1462–1505, 2006.

[4] C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *Siam Journal On Control and Optimization*, 44:283–312, 2005.

[5] C. Andrieu and C. P. Roberts. Conrolled MCMC for optimal sampling. Technical report, University Paris Dauphine, 2001.

[6] Y. Atchade and J. Rosenthal. On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11:815–828, 2005.

[7] Y. Bai, G. Roberts, and J. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. Preprint.

[8] M. Barrett, P. Galipeau, C. Sanchez, M. Emond, and B. Reid. Determination of the frequency of loss of heterozygosity in esophageal adeno-carcinoma nu cell sorting,

whole genome amplification and microsatellite polymorphisms. *Oncogene*, 12(1873-1878), 1996.

[9] R. C. Bradley. On the central limit question under absolute regularity. *Ann. Prob.*, 13:1314–1325, 1985.

[10] A. Brockwell and J. Kadane. Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *Journal of Computational and Graphical Statistics*, 14:436–458, 2005.

[11] X. Chen and C. Geyer. Limit theorems for functionals of ergodic Markov chains with general state space. *Mem. Amer. Math. Soc.*, 139:1747–1758, 1999.

[12] R. Cogburn. The central limit theorem for Markov processes. *Sixth Ann. Berkley Symp. Math. Statist. and Prob.*, 2:485–512, 1972.

[13] R. V. Craiu and C. Lemieux. Acceleration of the multiple-try Metropolis algorithm using antithetic and stratified sampling. *Statistics and Computing*, 17(2):109–120, 2007.

[14] R. V. Craiu, J. S. Rosenthal, and C. Yang. Learn from thy neighbor: Parallel-chain adaptive MCMC. Technical report, University of Toronto, 2008.

[15] M. Desai. *Mixture Models for Genetic changes in cancer cells*. PhD thesis, University of Washington, 2000.

[16] J. Eidsvik and H. Tjelmeland. On directional Metropolis-Hastings algorithms. *Statistics and Computing*, 16:93–106, 2006.

[17] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, Chichester, 1968.

[18] J. Gasemyr. On an adaptive version of the Metropolis-Hastings algorithm with independent proposal distribution. *Scand. J. Statist.*, 30:159–173, 2003.

[19] A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient Metropolis jumping rules. In *Bayesian statistics, 5 (Alicante, 1994)*, Oxford Sci. Publ., pages 599–607. Oxford Univ. Press, New York, 1996.

[20] W. Gilks, G. Roberts, and S. Suhu. Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, 93:1045–1054, 1998.

[21] P. Giordani and R. Kohn. Adaptive independent Metropolis-Hastings by fast estimation of mixtures of normals. Preprint, 2006.

[22] P. Green and A. Mira. Delayed rejection in reversible jump Metropolis-Hastings. *Biometrika*, 88:1035–1053, 2001.

[23] H. Haario, M. Laine, A. Mira, and E. Saksman. Dram: Efficient adaptive MCMC. *Statistics and Computing*, 16:339–354, 2006.

[24] H. Haario and E. Saksman. Simulated annealing process in general state space. *Adv. Appl. Probab.*, V:866–893, 1991.

[25] H. Haario, E. Saksman, and J. Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14:375–395, 1999.

[26] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

[27] H. Haario, E. Saksman, and J. Tamminen. Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, 20:265–273, 2005.

[28] O. Häggström. On the central limit theorem for geometrically ergodic Markov chains. *Prob. Theory and Rel.Fields*, 132:74–82, 2005.

[29] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[30] D. H. J. Besag, P.J. Green and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10:3–66, 1995.

[31] S. Jarner and G. Roberts. Polynomial convergence rates of Markov chains. *Ann. Appl. Prob.*, 12:224–247, 2002.

[32] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications*, 85:341–361, 2000.

[33] A. Jasra and C. Yang. A regeneration proof of the clt for uniformly ergodic Markov chains. *Statistics and Probability Letters*, 78:1649–1655, 2008.

[34] E. Jones. On the Markov chain central limit theorem. *Prob. Surveys*, 1:299–320, 2004.

[35] J. Liu, F. Liang, and W. Wong. The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95:121–134, 2000.

[36] N. Metroplolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J.Chem. Phys.*, 21:1087–1091, 1953.

[37] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series. Springer-Verlag, London, 1993.

[38] D. Nott and R. Kohn. Adaptive sampling for Bayesian variable selection. *Biometrika*, 92:747–763, 2005.

[39] C. Pasarica and A. Gelman. Adaptively acaling the Metropolis algorithm using the average squared jumped distance. Technical report, Department of Statistics, Columbia University, 2003.

[40] G. Roberts, A.Gelman, and W. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Prob*, 7:110–120, 1998.

[41] G. Roberts and R. Tweedie. Geometric convergence and Central Limit Theorems for multidimensional Hastings and Metropolis algorithm. *Biometrika*, 83:95–110, 1996.

[42] G. O. Roberts, A. Gelman, and W. Wilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7:110–120, 1997.

[43] G. O. Roberts and J. S. Rosenthal. Markov-chain Monte Carlo: some practical implications of theoretical results. *Canad. J. Statist.*, 26(1):5–31, 1998. With discussion by Hemant Ishwaran and Neal Madras and a rejoinder by the authors.

[44] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

[45] G. O. Roberts and J. S. Rosenthal. One-shot coupling for certain stochastic recursive sequences. *Stochastic Processes and Their Applications*, 99:195–208, 2002.

[46] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71 (electronic), 2004.

[47] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. Technical report, University of Toronto, 2006.

[48] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, 44(2):458–475, 2007.

[49] G. O. Roberts, J. S. Rosenthal, and P. O. Schwartz. Convergence properties of perturbed Markov chains. *J. Appl. Probab.*, 35(1):1–11, 1998.

[50] J. Rosenthal. *First Look at Rigorous Probability Theory.* Word Scientific, 2000.

[51] A. Smith and G.O.Roberts. Bayesian computation via the gibbs sampler and related Markov Chain Monte Carlo methods (with discussion). *J. Roy. Stat. Soc. Ser.*, B55:3–24, 1993.

[52] H. Thorisson. *Counpling, Stationary, and Regeneration*. Springer., 2000.

[53] L. Tierney. Markov chains for exploring posterior distributions(with discussion). *Ann. Stat.*, 22:1701–1762, 1994.

[54] G. Warnes. The Normal kernel coupler: An adaptive Markov chain Monte Carlo method for efficiently sampling from multi-modal distributions. Technical report, George Washington University, 2001.

[55] S. R. W.R. Gilks and D. Spiegelhalter. *Markov Chain Monte Carlo In Practice*. Chapman and Hall, London, 1996.

[56] C. Yang. On the weak law of large numbers for unbounded functionals for adaptive MCMC. Preprint.

[57] C. Yang. Recurrent and ergodic properties of adaptive MCMC. Preprint.

[58] C. Yang, R. Craiu, and J. Rosenthal. The ergodicity of mixed regional adaptive MCMC. Preprint.