CONVERGENCE OF ADAPTIVE MARKOV CHAIN MONTE CARLO ALGORITHMS

Yan Bai

Dissertation submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Department of Statistics
University of Toronto

CONVERGENCE OF ADAPTIVE MARKOV CHAIN MONTE CARLO ALGORITHMS

Yan Bai

Submitted for the Degree of Doctor of Philosophy

December 2009

Abstract

In the thesis, we study ergodicity of adaptive Markov Chain Monte Carlo methods (MCMC) based on two conditions (Diminishing Adaptation and Containment which together imply ergodicity), explain the advantages of adaptive MCMC, and apply the theoretical result for some applications.

First we show several facts: 1. Diminishing Adaptation alone may not guarantee ergodicity; 2. Containment is not necessary for ergodicity; 3. under some additional condition, Containment is necessary for ergodicity. Since Diminishing Adaptation is relatively easy to check and Containment is abstract, we focus on the sufficient conditions of Containment. In order to study Containment, we consider the quantitative bounds of the distance between samplers and targets in total variation norm. From early results, the quantitative bounds are connected with nested drift conditions for polynomial rates of convergence. For ergodicity of adaptive MCMC, assuming that all samplers simultaneously satisfy nested polynomial drift conditions, we find that either when the number of nested drift conditions is greater than or equal to two, or when the number of drift conditions with some specific form is one, the adaptive MCMC algorithm is ergodic. For adaptive MCMC algorithm with Markovian adaptation, the algorithm satisfying simultaneous polynomial ergodicity is ergodic without those restrictions. We also discuss some recent results related to this topic.

Second we consider ergodicity of certain adaptive Markov Chain Monte Carlo algorithms for multidimensional target distributions, in particular, adaptive Metropolis and adaptive Metropolis-within-Gibbs algorithms. We derive various sufficient conditions to ensure Containment, and connect the convergence rates of algorithms with the

tail properties of the corresponding target distributions. We also present a Summable Adaptive Condition which, when satisfied, proves ergodicity more easily.

Finally, we propose a simple adaptive Metropolis-within-Gibbs algorithm attempting to study directions on which the Metropolis algorithm can be run flexibly. The algorithm avoids the wasting moves in wrong directions by proposals from the full dimensional adaptive Metropolis algorithm. We also prove its ergodicity, and test it on a Gaussian Needle example and a real-life Case-Cohort study with competing risks. For the Cohort study, we describe an extensive version of Competing Risks Regression model, define censor variables for competing risks, and then apply the algorithm to estimate coefficients based on the posterior distribution.

Acknowledgements

First and foremost, I am most grateful to my advisor Professor Jeffrey S. Rosenthal for sharing his wealth of knowledge and wisdom with me. During my PhD study, I profited greatly from his personality when facing troubles. Without his guidance, encouragement and advice, the thesis can not be accomplished.

I would like to thank my thesis committee members: Professor V.Radu Craiu and Professor Michael Evans. They gave me many nice suggestions to improve my thesis. Professor Evans taught me the first statistics course during my study at the University of Toronto. Professor Craiu taught me many interesting statistical computing techniques. I would also like to thank my external examiner Professor Neal Madras in York University who spent hours reading my thesis and gave many correction comments.

I am also grateful to Professor Gareth O. Roberts in University of Warwick for some smart ideas, and to Professor Bálint Virág for useful discussions. I would thank Ms. Melania Pintilie and Mr. David Hodgson, with whom I obtained a lot about Cohort study. I owe thanks to Professor Sheldon Lin for his valuable suggestions about the presentation skills.

I also would like to thank many professors who taught me probability and statistics: David Brenner, Sebastian Jaimungal, Jeremy Quastel, Nancy Reid, and Fang Yao in University of Toronto; Alvo Mayer, Zarepour Mahmoud and David McDonald in University of Ottawa; Mohamedou Ould Haye, Majid Mojirsheibani, and Natalia Stepanova in Carleton University.

Many thanks go to the staff of Department of Statistics at the University of Toronto, Ms. Andrea Carter, Ms. Sarah Johns, Ms. Laura Kerr, Mr. Ram Mohabir, Mr. Dermot Whelan.

I would thank the Department of Statistics at the University of Toronto and Ontario Graduation Scholarship Institute for their financial support.

I also wish to thank some student colleagues in Department of Statistics with whom I profitted a lot from useful discussion: Elif Acar, Yichun Chi, Meng Du, Samuel Hikspoors, Zi Jin, Gun Ho Jang, Simon Lee, Li Li, Longhai Li, Mohammed Shakhatreh, Angelo Valov, Lizhen Xu, and Chao Yang.

I am indebted to some of my friends who gave me much emotional support when I was in the difficult time: Li Dong, Wei Fang, Shenggang Li, Cunye Qiao, Qinghong Song, Heng Wang, Sichun Wang, Biao Wu, Kun Zhang, and Jianjun Zheng.

Finally, I would thank my parents, Zongxin Bai and Aixiang Shi, and my wife, Ning Tang for their unconditional love and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The *Markov Chain Monte Carlo* (MCMC) methods are a class of simulation algorithms utilizing Markov Chain techniques to do complicated statistical computation, especially in high dimensional space. In the past half century, MCMC methods have become more and more mature and popular in the fields of statistical physics, statistics, computer science, mathematical finance, computational biology and others. More dramatically, MCMC techniques attract many practitioners who are interested in Bayesian inference, and sampling the posterior distribution of some complicated statistical models. Recently, some appealing non-ordinary MCMC algorithms called *adaptive MCMC* appear, which can also achieve the same goal as MCMC, sometimes even better than MCMC. In this section, we will give a brief description about these simulation methods, from MCMC methods to adaptive MCMC techniques.

This chapter consists of two parts. The first part attempts to give a brief introduction to the motivation and history of Monte Carlo methods (Section 1.1), Bayesian computation (Section 1.2), and some important MCMC algorithms (Section 1.3). The second part gives a description of adaptive MCMC (Section 1.4), their importance (Section 1.5), and the problems addressed in the thesis and the thesis organization (Section 1.6).

1.1 History

MCMC methods originated from Monte Carlo methods born in Los Alamos, New Mexico during World War II. At that time, many scientists were impressed by the speed and versatility of the electromechanical computers, because much tediousness and length of computation can be transformed to the burden of the electromechanical computers. With the development of the electromechanical computers, statistical computing techniques - especially Monte Carlo methods - were born.

The Metropolis algorithm published by Metropolis et al. (1953), was the first MCMC algorithm, proposed by the same group of scientists who invented Monte Carlo methods, namely the researchers of Los Alamos, mostly physicists working on mathematical physics. The specific case of the Boltzmann distribution was studied in their paper. There are N particles in a square. The potential energy of the system is

$$E = \frac{1}{2}\sum_{i \neq j} V(d_{ij}),$$

where V is the potential between molecules, and d_{ij} is the minimum distance between particle i and j. Their primary focus is to calculate the equilibrium value of any quantity of interest $f(\cdot, \cdot)$,

$$I = \frac{\int f(p,q)\exp\{-E(p,q)/kT\}dpdq}{\int \exp\{-E(p,q)/kT\}dpdq}.$$

Since p and q are $2N$-dimensional vectors, numerical integration is impossible. Standard Monte Carlo methods fails to correctly approximate I, because the $\exp\{-E(p,q)/kT\}$ is tiny for most realizations of the random configurations of the particle system. In order to improve the efficiency of Monte Carlo methods, Metropolis et al. (1953) propose a random walk modification of the N particles. For each particle i, values $x_i' = x_i + a\xi_{1i}$ and $y_i' = y_i + a\xi_{2i}$ are proposed where ξ_{ji} for $j = 1, 2$ are Unif$(-1, 1)$. The energy difference ΔE of between the new configuration and previous configuration is then computed. The new configuration is accepted with the probability

$$1 \wedge \exp(-\Delta E/kT),$$

and otherwise the previous configuration is replicated. Later Hastings (1970) generalized the Metropolis algorithm.

In the early 1970s, Hammersley, Clifford and Besag were working on the specification of joint distributions from conditional distributions, on necessary and sufficient conditions for the conditional distributions to be compatible with a joint distribution. An algorithm for extracting the marginal distributions from the full conditional distributions was studied by Geman and Geman (1984) and is known as the Gibbs sampler. The earlier articles by Metropolis et al. (1953) and Hastings (1970) developed essentially the same idea and suggested its potential for numerical problems arising in statistics. Gelfand and Smith (1990) inspired new interests in Bayesian methods, statistical computing, and stochastic processes through the use of computing algorithms such as Gibbs sampler and Metropolis-Hastings algorithm. Data argumentation algorithm was described by Tanner and Wong (1987) which has essentially the same impact as Gelfand and Smith (1990), namely the fact that simulating from conditional distributions is sufficient to simulate from the joint.

1.2 Bayesian Computation

The statistical techniques that we will be mostly concerned with are maximum likelihood and Bayesian methods. Their implementation are associated with much computation. For maximum likelihood methods, the problem is to find an estimate at which the likelihood function is maximized. For Bayesian methods, the problem is to compute posterior expectations.

In the Bayesian paradigm, the data $X_1, \cdots, X_n$ are realizations of the density function $p(x \mid \theta)$. The likelihood function $L(\theta \mid x_1, \cdots, x_n)$ for $\theta \in \Theta$ is the jointed density of $(X_1, \cdots, X_n)$ (it is viewed as a function of θ),

$$L(\theta \mid x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i \mid \theta).$$

Given some prior information specified by the prior distribution $\mu(\cdot)$, the distribution $\pi(\theta \mid x_1, \cdots, x_n)$ is called the posterior distribution. The joint distribution of

$(X_1, \cdots, X_n, \theta)$ is

$$L(\theta \mid x_1, \cdots, x_n)\mu(\theta).$$

Thus,

$$\pi(\theta \mid x_1, \cdots, x_n) = \frac{L(\theta \mid x_1, \cdots, x_n)\mu(\theta)}{\int_\Theta L(\theta' \mid x_1, \cdots, x_n)\mu(\theta')d\theta'}.$$

In general, the Bayes estimate under the loss function $R(\theta, \delta)$ and the prior μ is the solution of the minimization program

$$\min_\delta \int_\Theta R(\theta, \delta)\mu(\theta)L(\theta \mid x_1, \cdots, x_n)d\theta.$$

When the loss function is the quadratic form, the Bayes estimate will be a posterior expectation. So, we need to estimate expectation of some function $h : \Theta \longrightarrow \mathbb{R}$ with respect to $\pi(\cdot \mid x_1, \cdots, x_n)$, i.e. we want to estimate

$$E_\pi[h(Y)] = \int_\Theta h(\theta)\frac{L(\theta \mid x_1, \cdots, x_n)\mu(\theta)}{\int_\Theta L(\theta' \mid x_1, \cdots, x_n)\mu(\theta')d\theta'}d\theta.$$

$L(\theta \mid x_1, \cdots, x_n)$ could be of a complicated form. Then the direct computation of the above integration will be infeasible. The classical Monte Carlo simulation to the problem is to simulate i.i.d. random variables $Z_1, Z_2, \cdots, Z_N \sim \pi(\cdot)$, and then use the $\sum_{i=1}^N h(Z_i)/N$ to estimate $\pi(h) := \int_\Theta h(\theta)\pi(d\theta)$.

1.3 Some important MCMC algorithms

A severe drawback of Monte Carlo methods is that complete determination of the functional form of the posterior density is needed for their implementation. Situations in which the posterior distribution is indirectly specified cannot be handled. One example is a Bayesian hierarchical model where the joint distribution of a random vector is only specified by a group of conditional distributions.

In the field of MCMC methods, many critical questions related to probability the-

ory on Markov chains [1] emerge from the appearance of more complex algorithms. However, the underlying idea is very simple. Suppose that we want to generate a sample from a distribution $\pi(\cdot)$ (also called target distribution) on the state space $\mathcal{X} \subset \mathbb{R}^d$ but cannot do it directly. Assume that a Markov chain can be constructed on the space $\mathcal{X}$ with the stationary distribution $\pi(\cdot)$. Then we can run the chain for a long time. The simulated values from the chain can be viewed as a basis for exploring the feature of $\pi(\cdot)$. So, we simply need to design algorithms for constructing a Markov chain with a specified stationary distribution. The simple procedure involves the probability theory of Markov chains on the general state space, and hence some basic understanding of Markov chains is required. See some basic Markov Chain concepts and theories in Appendix A.

Many MCMC methods are related to *reversible* Markov Chain (i.e. for any sets A and B, $\int_B \int_A \pi(dx)P(x,dy) = \int_A \int_B \pi(dy)P(y,dx)$[2]), which means that given a stationary ergodic irreducible (see their definitions in Appendix A) Markov Chain $\cdots, X_{n-2}, X_{n-1}, X_n, \cdots$, the reverse process is the same Markov chain. It is easy to show the following property.

Proposition 1.3.1. *If the Markov chain* $\mathbf{X}$ *is reversible with respect to the measure* π *then* π *is stationary for the chain.*

Proof:

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

□

1.3.1 Metropolis-Hastings Algorithm

In this section, we study a very general MCMC method - Metropolis-Hastings algorithm, the discovery of which has led to very considerable progress in simulation-based inference, particular in Bayesian Analysis.

Let the state space $\mathcal{X}$ be an open set in $\mathbb{R}^d$ and a target distribution $\pi(\cdot)$ with the density $t : \mathcal{X} \to (0, \infty) \geq 0$ with $\int t(x)\mu(dx) < \infty$ where μ is d-dimensional Lebesgue

[1] The future evolution of the chain is only dependent on the current state, and independent of the past states.

[2] $P(x, dy)$ is the transition kernel of a Markov chain, see the definition in Appendix A.

measure.

Metropolis-Hastings algorithm: generate a Markov chain $\mathbf{X} = \{X_n : n \geq 0\}$ based on the target density function $t(\cdot)$ and a proposal distribution $Q(x, dy)$ with the density $q(x, y)$. At each time $n + 1$, given X_n the proposal value Y_{n+1} is obtained from the proposal distribution $Q(X_n, dy)$. X_{n+1} is assigned Y_{n+1} with the probability $\alpha(X_n, Y_{n+1})$ where

$$\alpha(x, y) := \min\left(1, \frac{t(y)q(y, x)}{t(x)q(x, y)}\right), \tag{1.1}$$

otherwise X_{n+1} is assigned X_n. For the special case $q(x, y) = q(y, x)$ implying $\alpha(x, y) = \min\left(1, \frac{t(y)}{t(x)}\right)$, call it *Metropolis algorithm*. Further when $q(x, y) = q(y, x) = q(x - y)$, call it *symmetric random-walk-based Metropolis algorithm*. From the above description, for running Metropolis-Hastings algorithm, we just need to run the proposal distribution and then accept or reject the proposal value. Thus, the procedure is quite feasible. In addition, the acceptance ratio weakens the requirement around the target distribution. The normalization factor of target distribution is not necessary.

Proposition 1.3.2. *The Metropolis-Hastings algorithm produces a reversible Markov chain* $\mathbf{X}$ *with respect to* $\pi(\cdot)$.

Proof: First the Metropolis-Hastings transition kernel is

$$P(x, dy) = \alpha(x, y)q(x, y)\mu(dy) + \delta_x(dy)\int_{\mathcal{X}}(1 - \alpha(x, z))q(x, z)\mu(dz).$$

For any measurable function $f(x, y) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, any sets $A, B \in \mathcal{X}$,

$$\int_{A\times B} f(x, y)\pi(dx)\delta_x(dy) = \int_{A\cap B} f(x, x)\pi(dx) = \int_{A\times B} f(y, x)\delta_y(dx)\pi(dy).$$

So,

$$\begin{aligned}
&\pi(dx)P(x,dy)\\
=&\pi(dx)\alpha(x,y)q(x,y)\mu(dy)+\pi(dx)\delta_x(dy)\int_{\mathcal{X}}(1-\alpha(x,z))q(x,z)\mu(dz)\\
=&(t(x)q(x,y))\wedge(t(y)q(y,x))\mu(dx)\mu(dy)+\\
&\left(\pi(dx)\delta_x(dy)\int_{\mathcal{X}}(1-\alpha(x,z))q(x,z)\mu(dz)+\right.\\
&\left.\pi(dy)\delta_y(dx)\int_{\mathcal{X}}(1-\alpha(y,z))q(y,z)\mu(dz)\right)/2\\
=&\pi(dy)P(y,dx),
\end{aligned}$$

because of the symmetry. □

1.3.2 Gibbs Sampler

The Gibbs sampler is a technique especially suitable for generating an irreducible aperiodic Markov chain that has the target distribution as its stationary distribution in a high dimensional space. It generates a sequence of values from the joint distribution of multiple random variables. The purpose of the sequence is to simulate the joint distribution. The Gibbs sampling generates an instance from the distribution of each variable in turn, conditional on the current values of other variables. So Gibbs sampling is applicable when the joint distribution is not known exactly, and the conditional distribution of each variable is known. The point is that it is simpler to sample from a conditional distribution than to sample from the joint distribution.

Consider the target density function $t(x_1,\cdots,x_d)$ in the state space $\mathcal{X}\in\mathbb{R}^d$. Gibbs sampler consists of d components, the i^{th} of which is the full conditional distribution conditioned on all the other components. Formally, let the set

$$\mathcal{S}_{i,a,b}(x):=\{y\in\mathcal{X}:y_j=x_j \text{ for } j\neq i \text{ and } y_i\in[a,b]\}. \tag{1.2}$$

Then the transition kernel is defined

$$P_i(x,\mathcal{S}_{i,a,b}(x))=\frac{\int_a^b t(x_1,\cdots,x_{i-1},u,x_{i+1},\cdots,x_d)du}{\int t(x_1,\cdots,x_{i-1},u,x_{i+1},\cdots,x_d)du}. \tag{1.3}$$

So, respectively define *deterministic-scan Gibbs sampler* as $P_{\rm DS} := P_1 \cdots P_d$, and *random-scan Gibbs sampler* as $P_{\rm RS} := \frac{1}{d}\sum_{i=1}^{d} P_i$ [3].

Here we check the invariant property of the Gibbs samplers with respect to π. For the deterministic-scan Gibbs sampler,

$$\begin{aligned}
\pi(dx) &= \int_{\mathbb{R}} \pi(dx_d \mid x_1, \cdots, x_{d-1})\pi(dx_1, \cdots, dx_{d-1}, dy_d) \\
&= \int_{\mathbb{R}^d} \prod_{j=1}^{d} \pi(dx_j \mid x_1, \cdots, x_{j-1}, y_{j+1}, \cdots, y_d)\pi(dy_1, \cdots, dy_d) \\
&= \int_{\mathbb{R}^d} P_{\rm DS}(y, dx)\pi(dy).
\end{aligned}$$

The following result will be used to show that π is invariant to $P_{\rm RS}$.

Proposition 1.3.3. *Let ν be any distribution on unit hypersurface $S^{d-1} = \{u \in \mathbb{R}^d : |u| = 1\}$. Define the specific transition kernel $P_\theta(x, \cdot)$ passing through $x \in \mathbb{R}^d$ along the direction $\theta \in S^d$ as*

$$P_\theta(x, A) = \frac{\int_{-\infty}^{\infty} \mathbb{I}_A(x + r\theta)t(x + r\theta)dr}{\int_{-\infty}^{\infty} t(x + \lambda\theta)d\lambda}.^{4} \tag{1.4}$$

Then, $\pi(\cdot)$ is invariant for $P_\nu(x, \cdot)$ where $P_\nu(x, B) = \int_{S^d} P_\theta(x, B)\nu(d\theta)$.

Proof: For any $\theta \in S^d$, by Fubini's theorem and change of variable $y = x + r\theta$ and $u = \lambda - r$,

$$\begin{aligned}
\int_A P_\theta(x, B)\pi(dx) &= \int_{\mathbb{R}^d}\int_{-\infty}^{\infty} \mathbb{I}_A(x)\mathbb{I}_B(x + r\theta)\frac{t(x + r\theta)t(x)}{\int_{-\infty}^{\infty} t(x + \lambda\theta)d\lambda} dr\mu(dx) \\
&= \int_{-\infty}^{\infty}\int_{\mathbb{R}^d} \mathbb{I}_A(y - r\theta)\mathbb{I}_B(y)\frac{t(y)t(y - r\theta)}{\int_{-\infty}^{\infty} t(y + u\theta)du} dr\mu(dy) \\
&= \int_B P_\theta(y, A)\pi(dy).
\end{aligned}$$

Integrating both sides of the above equation, we have that for any distribution ν on S^d, $P_\nu(x, \cdot)$ is reversible with respect to π. So, the result holds. □

[3] $P_{\rm DS}$ may not be reversible, and $P_{\rm RS}$ is reversible.

[4] It is a general form of Equation (1.3)

We can define ν is uniform on $\{e_1, \cdots, e_d\}$ where $e_i = (\underbrace{0, \cdots, 0}_{i-1}, 1, \underbrace{0, \cdots, 0}_{d-i})$. So, by Proposition 1.3.3, π is invariant to P_{RS}.

1.3.3 Metropolis-within-Gibbs Sampler

The Metropolis-within-Gibbs Sampler is a kind of hybrid sampler combining Metropolis algorithm and Gibbs sampler. The Gibbs sampling framework is adopted, but on each coordinate the conditional distribution is substituted by running a Metropolis algorithm. For *deterministic scan Metropolis-within-Gibbs* algorithm, sequentially select the coordinate $1, \cdots, d$, and then run Metropolis algorithm on each coordinate. But for *random scan Metropolis-within-Gibbs*, uniformly select one of the coordinates $1, \cdots, d$.

Roberts and Rosenthal (2006) studied the conditions under which the Metropolis-within-Gibbs algorithm (MwG) is Harris recurrent or not. Fort et al. (2003) presented some conditions under which the symmetric random-walk-based Metropolis-within-Gibbs algorithm is geometrically ergodic. Roberts and Rosenthal (2009) studied a certain adaptive Metropolis-within-Gibbs algorithm for hierarchical models.

For $1 \leq i \leq d$, let $q_i : \mathcal{X} \times \mathbb{R} \longrightarrow [0, \infty)$ be jointly measurable with $\int_{-\infty}^{\infty} q_i(x, z)dz = 1$ for all $x \in \mathcal{X}$ where dz is one dimensional Lebesgue measure. Let $Q_i(x, \cdot)$ be the Markov kernel on $\mathbb{R}^d$ which replaces the ith coordinate by a draw from the density $q_i(x, \cdot)$, but leaves the other coordinates unchanged. That is

$$Q_i(x, \mathcal{S}_{i,a,b}(x)) = \int_a^b q_i(x, z)dz,$$

where $\mathcal{S}_{i,a,b}$ is defined in Equation (1.2). Say $Q_i(x, \cdot)$ is symmetric if

$$q_i((x_1, \cdots, x_d), z) = q_i((x_1, \cdots, x_{i-1}, z, x_{i+1}, \cdots, x_d), x_i).$$

For $x, y \in \mathbb{R}^d$ and $1 \leq i \leq d$, let

$$\alpha_i(x, y) := \mathbb{I}\left(t(x)q_i(x, y_i) \neq 0\right) \min\left[1, \frac{t(y)q_i(y, x_i)}{t(x)q_i(x, y_i)}\right].$$

Let P_i be the kernel which proceeds as follows. Given X_n, it generates the proposal $Y_{n+1} \sim Q_i(X_n, \cdot)$. Then X_{n+1} is assigned Y_{n+1} with the probability $\alpha_i(X_n, Y_{n+1})$, and is assigned X_n with the probability $1 - \alpha_i(X_n, Y_{n+1})$.

Let I_n be a random variable on $\{1, \cdots, d\}$. Two most common schemes are *deterministic-scan Metropolis-within-Gibbs sampler* $P_{\mathrm{DS}} = P_{I_n}$ where $I_n = n \mod d$, and *random-scan Metropolis-within-Gibbs sampler* $P_{\mathrm{RS}} = P_{I_n}$ where I_n is uniform on $\{1, \cdots, d\}$. Then for $n = 0, 1, 2, \cdots$ given X_n the state $X_{n+1} \sim P_{I_n}(X_n, \cdot)$. It is straightforward to verify that the chain has stationary distribution $\pi(\cdot)$.

1.4 Adaptive MCMC

MCMC methods are widely used for approximately sampling from complicated probability distributions. However, it is often necessary to tune the scaling and other parameters before the algorithm will converge efficiently. Given some extremely complicated target distribution, it is even difficult to know how to tune the corresponding parameters. *Adaptive MCMC* methods modify the transitions on the fly, in an effort to automatically tune the parameters and improve convergence. The automatic tuning for sampler is mainly based on the historical information.

Non-adaptive MCMC algorithms are usually constructed through a fixed computing framework (a time-homogenous Markov Chain kernel). However, adaptive MCMC algorithms are generated through a collection of computing frameworks indexed by $\mathcal{Y}$. At each time n, the state X_n is chosen through some adaptation strategy automatically designed according to historical information. During the adaptation procedure, one computing framework will be selected. From the framework of two kinds of algorithms, it is clear that adaptive MCMC is more general and flexible.

Since non-adaptive MCMC algorithms have unaltered framework at each iteration, under some situations their performance may not be good, and even the convergence properties are destroyed. For instance, suppose that there is a target distribution with single unknown sized mode. We run Metropolis algorithm to sample from the target distribution. If the variance of the proposal distribution is too small, the constructed chain will be jumping very slowly although the acceptance rate is high. If the proposal variance is too big, the chain is very active but rejected at most of the

running time. Haario et al. (2001) give an adaptive Metropolis algorithm which at each iteration, studies from the obtained data, estimates the variance of the target distribution, and employs the empirical covariance estimate as the proposal variance. They provide theoretical justification for the adaptation of the covariance matrix used in the Metropolis algorithm. The method smartly and automatically adapts the proposal variance.

Another case can better explain why ordinary MCMC cannot be always used. Consider a target distribution with a high mode and a lower mode, these two modes are not very close to each other. The target has variable "local properties" (two different modes). One interesting idea is to classify the obtained sample data by some boundary on the two sides of which two modes locate, use the empirical covariance matrix generated from each side of the boundary as the local proposal variances, see details in Craiu et al. (2008). They prove ergodicity of their adaptive MCMC algorithm. We use their method to analyze the mixed uniform distribution $1/5\text{Unif}(-11,-9)+4/5\text{Unif}(9.99,10.01)$, and compare it with Metropolis algorithm using $\text{Unif}(x-23,x+23)$ as the proposal distribution through the estimate of total variation norm [5], see Figure 1.1. From the plots, it is obvious the the convergence of adaptive MCMC algorithm is more stable than Metropolis algorithm.

The above two cases explain that the performance of simulation methods is relevant to the locations of target modes, and target's local properties. We find that the relevance is also dependent on the configuration of target support's region. A simple example is a high dimensional target that is distributed on a slim needle, and at least two dimensions of the space are highly correlated. If ordinary MCMC is applied to approximately simulate from the specified target, many wasting moves in wrong directions by proposals will be generated so that their performances are not good. See details in Chapter 5.

Adaptive MCMC methods are extremely significant under some situations. For example, classical MCMC methods may not efficiently simulate a target distribution supported on a weird region. However, through long term running some useful information can be obtained. Based on the knowledge, adaptively choosing the transition scheme will be effective. In addition, Adaptive MCMC can automatically study

[5] We simply calculate the percentages of sample points falling into two intervals, $A_1 := (-11,-9)$ and $A_2 := (9.99,10.01)$, and then define $|P_{A_1}-0.2|$ ($|P_{A_2}-0.8|$) as the estimate.

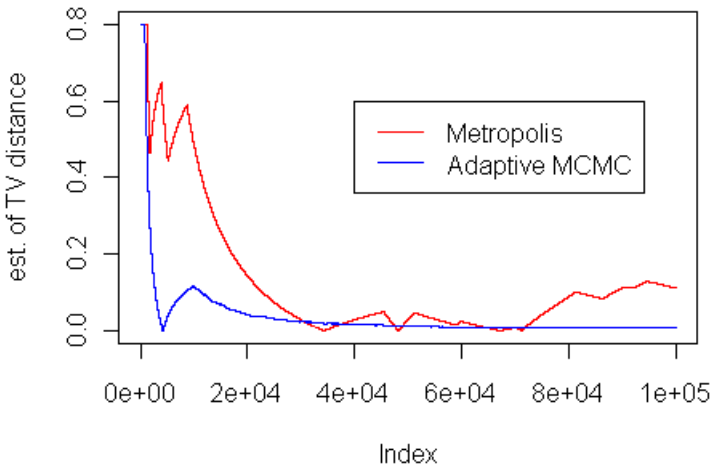

Figure 1.1: The estimates of total variation norm by Metropolis algorithm and adaptive MCMC algorithm.

statistical model parameters. If the adaptation scheme is well designed, adaptive algorithms may be better than non-adaptive algorithms. Especially for high dimensional correlated target distributions, some adaptive algorithms may perform better than non-adaptive algorithms.

Although the optimality of Metropolis algorithm was studied to some extent, there are many complicated cases where it is difficult to find the optimal MCMC algorithm. So, adaptive MCMC methods may provide an alternative approach.

Some adaptive MCMC methods use regeneration times and other somewhat complicated constructions, see Gilks et al. (1998); Brockwell and Kadane (2005). However, Haario et al. (2001) proposed an adaptive Metropolis algorithm attempting to optimise the proposal distribution, and proved that a particular version of this algorithm correctly converges strongly to the target distribution. The algorithm can be viewed as a version of the Robbins-Monro stochastic control algorithm, see Robbins and Monro (1951); Andrieu and Robert (2001). The results were then generalized through proving convergence of more general adaptive MCMC algorithms, see Atchadé and Rosenthal (2005); Andrieu and Moulines (2006). Following that, many general re-

sults were developed, see Roberts and Rosenthal (2007); Yang (2008a,b); Saksman and Vihola (2008); Bai et al. (2008); Atchadé and Fort (2008); Craiu et al. (2008); Bai (2009a,b).

1.5 Some notations for adaptive MCMC

Consider the collection $\{P_\gamma : \gamma \in \mathcal{Y}\}$ of Markov Chain transition kernels on the state space $\mathcal{X}$ and the adaptive parameter space $\mathcal{Y}$ where each Markovian transition kernel P_γ is time-homogeneous, φ_γ-irreducible and aperiodic with stationary measure $\pi(\cdot)$.

Given the $\mathcal{X}$-value random sequence $X_0, \cdots, X_n$, and the $\mathcal{Y}$-value random sequence $\Gamma_0, \cdots, \Gamma_n$, at the time n, X_{n+1} is generated by the transition kernel $P_{\Gamma_n}(X_n, \cdot)$, and then the transition kernel $P_{\Gamma_{n+1}}$ is chosen according to some adaptation scheme. Actually the adaptation scheme is to decide the selection of Γ_{n+1}.

Denote the filtration by $\mathcal{F}_n = \sigma(X_k, \Gamma_k : 0 \le k \le n)$. Formally, the *adaptive MCMC* process $\{X_n : n \ge 0\}$ is a chain which at each time $n+1$ satisfies the property:

$$\mathbf{P}(X_{n+1} \in A \mid \mathcal{F}_n) = \mathbf{P}(X_{n+1} \in A \mid X_n, \Gamma_n) = P_{\Gamma_n}(X_{n+1} \in A \mid X_n), \tag{1.5}$$

and the random kernel index Γ_{n+1} is selected through the history information with the property, i.e. Γ_{n+1} is some function of $\mathcal{F}_n$ and X_{n+1}. Obviously the joint distribution of X_{n+1} and Γ_{n+1} conditional on $\mathcal{F}_n$ is

$$\mathbf{P}\left(X_{n+1} \in dx, \Gamma_{n+1} \in d\gamma \mid \mathcal{F}_n\right) = P_{\Gamma_n}(X_{n+1} \in dx \mid X_n)\mathbf{P}(\Gamma_{n+1} \in d\gamma \mid X_{n+1} = x, \mathcal{F}_n).$$

Furthermore, if Γ_{n+1} is simply a function of X_n and Γ_n, then the adaptive MCMC algorithm is called *Markovian Adaptation*, i.e. the joint process $\{(X_n, \Gamma_n) : n \ge 0\}$ is a time-inhomogeneous Markov Chain.

Write $\mathbf{P}\left(X_n \in \cdot \mid X_0 = x_0, \Gamma_0 = \gamma_0\right) := \mathbf{P}_{(x_0,\gamma_0)}\left(X_n \in \cdot\right)$. Denote the corresponding expectation by $\mathbf{E}_{(x_0,\gamma_0)}\left[f(X_n)\right]$ for some measurable function $f : \mathcal{X} \to \mathbb{R}$.

Say the adaptive algorithm $\{X_n : n \ge 0\}$ with the adaptive scheme $\{\Gamma_n : n \ge 0\}$ is

ergodic if for any initial point $(x_0, \gamma_0) \in \mathcal{X} \times \mathcal{Y}$,

$$\lim_{n\to\infty} \left\| \mathbf{P}_{(x_0,\gamma_0)}(X_n \in \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} = 0, \tag{1.6}$$

where $\|\cdot\|_{\mathrm{TV}}$ is total variation norm, see Equation (A.5) in Appendix A.

From Theorem A.3.1, irreducibility is a natural property for studying convergence of a Markov chain, under which the chain is either transient or recurrent. However, for adaptive MCMC chains, irreducibility may not be preserved even if each kernel in $\mathcal{Y}$ is irreducible, see Example 2 in Roberts and Rosenthal (2007). Another important issue is that stationarity may not hold even if each kernel P_γ for $\gamma \in \mathcal{Y}$ does. From these two points, the coupling method in Appendix A cannot be implemented directly.

1.6 The problems addressed in the thesis

Roberts and Rosenthal (2007) use a coupling method to show that ergodicity of adaptive MCMC algorithm is implied by Containment and Diminishing Adaptation.

Definition 1.6.1 (Diminishing Adaptation)**.** $\lim_{n\to\infty} D_n = 0$ *in probability, where*

$$D_n := \sup_{x\in\mathcal{X}} \left\| P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot) \right\|_{\mathrm{TV}}, \tag{1.7}$$

is $\mathcal{F}_{n+1}$*-measurable random variable.*

The condition means that the change of adaptive kernel converges to zero. It is relatively easy to check, because the adaptive scheme is artificially designed.

Definition 1.6.2 (Containment)**.** *The stochastic process* $\{M_\epsilon(X_n, \Gamma_n) : n \geq 0\}$ *is bounded in probability given any starting point* $(x_0, \gamma_0) \in \mathcal{X} \times \mathcal{Y}$ *where*

$$M_\epsilon(x, \gamma) := \inf_n \left\{ n \in \mathbb{N}^+ : \left\| P^n_\gamma(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} < \epsilon \right\} : \mathcal{X} \times \mathcal{Y} \to \mathbb{N}^+, \tag{1.8}$$

where $\mathbb{N}^+ = \{1, 2, 3, \cdots\}$*, i.e. given any* $(x_0, \gamma_0) \in \mathcal{X} \times \mathcal{Y}$*,* $\forall \epsilon > 0$*,* $\forall \delta > 0$*,* $\exists K > 0$*, such that* $\mathbf{P}_{(x_0,\gamma_0)}(M_\epsilon(X_n, \Gamma_n) > K) < \delta$ *for all* n*.*

Indeed, Containment means that the sequence $\{M_\epsilon(X_n, \Gamma_n) : n \geq 0\}$ is tight.

However, since $M_\epsilon(x,\gamma)$ is integer-valued, so the tightness is equivalent to "bounded in probability". From intuition, the condition means that the time that transition kernels get close to the target within ϵ is bounded. Say that the adaptive parameter process Γ_n *is bounded in probability* if $\forall \epsilon > 0, \exists N > 0,\ \exists$ some compact set $B \subset \mathcal{Y},$ such that for $n > N, \mathbf{P}(\Gamma_n \in B^c) < \epsilon$.

Theorem 1.6.1 (Roberts and Rosenthal 2007)**.** *Consider an adaptive MCMC algorithm on a state space $\mathcal{X}$, with adaptation index $\mathcal{Y}$, so $\pi(\cdot)$ is stationary for each kernel P_γ for $\gamma \in \mathcal{Y}$. Assuming Containment and Diminishing Adaption, the adaptive algorithm is ergodic.*

Adaptation schemes can be artificially designed so that Diminishing Adaptation is not hard to check "relatively". Containment is considerably abstract and hard to check.

In the thesis, we will mainly study the importance of Containment and the sufficient conditions for Containment. Moreover, we find some easy-to-check conditions for adaptive Metropolis and adaptive Metropolis-within-Gibbs algorithms.

In Chapter 2, we introduce some examples which explain some relationships among Containment and Diminishing Adaptation and ergodicity, and the advantage of adaptive MCMC respectively.

In Chapter 3 we show that Simultaneously Polynomially Ergodic condition (S.P.E) implies Containment for most cases. That is either when the number of drift conditions is greater than or equal to two, or when the number of drift conditions having some specific form is one, the adaptive MCMC algorithm is ergodic.

In Chapter 4 we study Simultaneously Geometrically Ergodic condition (S.G.E.). We connect the tail properties of target densities with S.G.E., and show that when a target density is exponentially tailed, adaptive Metropolis and adaptive random-scan Metropolis-within-Gibbs algorithms are ergodic under some mild conditions. A summable adaptive condition is also given which can alone imply ergodicity.

In Chapter 5 we develop a simple adaptive directional Metropolis-within-Gibbs algorithm which avoids wasting moves in wrong directions by proposals from Metropolis-within-Gibbs sampler. We also prove its ergodicity, and test it on a Gaussian Needle example and a real-life Case-Cohort study with competing risks.

In Chapter 6 we give some conclusions and future works about the topic.

In Appendix A, we introduce some basic concepts, theorems and methodologies on Markov Chains.

Chapter 2

Some Adaptive MCMC Examples

In this chapter, we give some examples which explain some relationships among Containment and Diminishing Adaptation and ergodicity, and the advantages of adaptive MCMC.

The example in Section 2.1 is used to explain that 1. Diminishing Adaptation alone is not sufficient for ergodicity; 2. An adaptive algorithm is ergodic but both Containment and Diminishing Adaptation do not hold.

A half-Cauchy counter example is given in Section 2.2 to also show that Diminishing Adaptation alone is not sufficient for ergodicity. The example is interesting in that only two transition kernels are adaptively chosen.

In Section 2.3 an adaptive Metropolis algorithm is used to analyze a mixture model, and some simulation results are given. This algorithm is a slight variant version of Haario's algorithm.

2.1 A state-independent adaptive example

Example 2.1.1. *Let the state space* $\mathcal{X} = \{1, 2\}$ *and the transition kernel*

$$P_\theta = \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}.$$

Obviously, for each $\theta \in (0,1)$, the stationary distribution is uniform on $\mathcal{X}$.

Proposition 2.1.1. *For the target distribution and the family of transition kernels in Example 2.1.1, consider a state-independent adaptation: at each time $n \geq 1$ choose the transition kernel index $\theta_{n-1} = \frac{1}{(n+1)^r}$ for some fixed $r > 0$ (P_{θ_0} is the initial kernel). Show that*

(i) *For $r > 0$, Diminishing Adaptation holds but Containment does not;*

(ii) *For $r > 1$, $\mu_0 P_{\theta_0} P_{\theta_1} \cdots P_{\theta_n} \to \mu$ where $\mu_0 = (1,0)^\top$ and $\mu = (\frac{1+\alpha}{2}, \frac{1-\alpha}{2})^\top$ for some $\alpha \in (0,1)$;*

(iii) *For $0 < r \leq 1$ and a probability measure μ_0 on $\mathcal{X}$, $\mu_0 P_{\theta_0} P_{\theta_1} \cdots P_{\theta_n} \to \mathrm{Unif}(\mathcal{X})$.*

Remark 2.1.1. *The chain in Proposition 2.1.1 is a time inhomogeneous Markov chain. It can be suited into the framework of adaptive MCMC. Although very simple, it reflects the complexity of adaptive MCMC to some degree.*

1. *For $r > 1$, the limiting distribution of the chain is not uniform. So it shows that Diminishing Adaptation alone cannot ensure ergodicity.*

2. *For $0 < r \leq 1$, the algorithm is ergodic to an uniform distribution. So, it implies that Containment is not necessary for ergodicity.*

Proof: Since the adaptation is state-independent, the stationarity is preserved. So, the adaptive MCMC $X_n \sim \delta P_{\theta_0} P_{\theta_1} P_{\theta_2} \cdots P_{\theta_{n-1}}(\cdot)$ for $n \geq 0$ where $\delta := (\delta^{(1)}, \delta^{(2)})$ is the initial distribution.

The part (i). Consider $\left\| P_{\theta_{n+1}}(x,\cdot) - P_{\theta_n}(x,\cdot) \right\|_{\mathrm{TV}}$. For any $x \in \mathcal{X}$,

$$\left\| P_{\theta_{n+1}}(x,\cdot) - P_{\theta_n}(x,\cdot) \right\|_{\mathrm{TV}} = |\theta_{n+1} - \theta_n| \to 0.$$

Thus, for $r > 0$ Diminishing Adaptation holds.

By some algebra,

$$\|P_\theta^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} = \frac{1}{2} |1 - 2\theta|^n . \tag{2.1}$$

Hence, for any $\epsilon > 0$,

$$M_\epsilon(X_n, \theta_n) \geq \frac{\log(\epsilon) - \log(1/2)}{\log|1 - 2\theta_n|} \to +\infty \qquad \text{as } n \to \infty. \tag{2.2}$$

Therefore, the stochastic process $\{M_\epsilon(X_n, \theta_n) : n \geq 0\}$ is not bounded in probability.

The parts (ii) and (iii). Let $\mu_n := \left(\mu_n^{(1)}, \mu_n^{(2)}\right) := \delta P_{\theta_0} \cdots P_{\theta_n}$. So,

$$\mu_{n+1}^{(1)} = \mu_n^{(1)} - \theta_{n+1}\left(\mu_n^{(1)} - \mu_n^{(2)}\right) \text{ and } \mu_{n+1}^{(2)} = \mu_n^{(2)} + \theta_{n+1}\left(\mu_n^{(1)} - \mu_n^{(2)}\right).$$

Hence,

$$\mu_{n+1}^{(1)} - \mu_{n+1}^{(2)} = \left(\delta^{(1)} - \delta^{(2)}\right) \prod_{k=0}^{n+1}(1 - 2\theta_k).$$

For $r > 1$, $\prod_{k=0}^{n+1}(1-2\theta_k)$ converges to some $\alpha \in (0,1)$ as n goes to infinity. $\mu_{n+1}^{(1)} - \mu_{n+1}^{(2)} \to \left(\delta^{(1)} - \delta^{(2)}\right)\alpha$. For $0 < r \leq 1$, $\mu_{n+1}^{(1)} - \mu_{n+1}^{(2)} \to 0$. Therefore, for $r > 1$ ergodicity to Uniform distribution does not hold, and for $0 < r \leq 1$ ergodicity holds. □

Proposition 2.1.2. *For the target distribution and the family of transition kernels in Example 2.1.1, consider an independent adaptation: for $k = 1, 2, \cdots$, at each time $n = 2k-1$ choose the transition kernel index $\theta_{n-1} = 1/2$, and at each time $n = 2k$ choose the transition kernel index $\theta_{n-1} = 1/n$. Diminishing Adaptation and Containment do not hold. The chain converges to the target distribution* $\text{Unif}(\mathcal{X})$.

Proof: From Equation (2.1), for $\epsilon > 0$, $M_\epsilon(X_{2k-1}, \theta_{2k-1}) \geq \frac{\log(\epsilon)-\log(1/2)}{\log|1-1/k|} \to \infty$ as $k \to \infty$. So, Containment does not hold.
$\left\|P_{\theta_{2k}}(x,\cdot) - P_{\theta_{2k-1}}(x,\cdot)\right\|_{TV} = \left|\frac{1}{2} - \frac{1}{2k}\right| \to \frac{1}{2}$ as $k \to \infty$. So Diminishing Adaptation does not hold.
Let $\delta := (\delta^{(1)}, \delta^{(2)})$ be the initial distribution and $\mu_n := (\mu_n^{(1)}, \mu_n^{(2)}) = \delta P_{\theta_0} \cdots P_{\theta_n}$. $\mu_n^{(1)} - \mu_n^{(2)} = (\delta^{(1)} - \delta^{(2)}) 2^{-[n/2]-1} \prod_{k=1}^{[(n+1)/2]} \left(1 - \frac{1}{2k}\right) \to 0$ as n goes to infinity. So ergodicity holds. □

2.2 A Half-Cauchy Counter Example

Example 2.2.1. *Let the state space $\mathcal{X} = (0, \infty)$, and the kernel index set $\mathcal{Y} = \{-1, 1\}$. The target density $\pi(x) \propto \frac{\mathbb{I}(x>0)}{1+x^2}$ is a half-Cauchy distribution on the positive part of $\mathbb{R}$. At each time n, run the Metropolis-Hastings algorithm where the proposal*

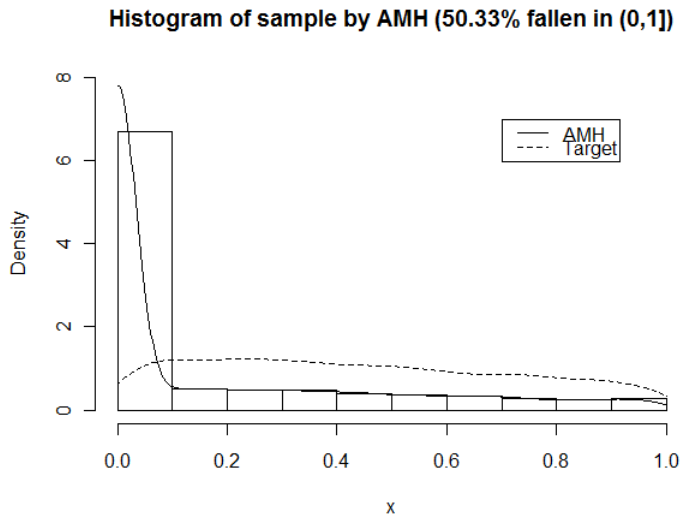

Figure 2.1: The solid line is the estimated density by adaptive Metropolis-Hastings algorithm. The dashed line is the estimated density by the sample from the target distribution.

value Y_n *is generated by*

$$Y_n^{\Gamma_{n-1}} = X_{n-1}^{\Gamma_{n-1}} + Z_n \tag{2.3}$$

with i.i.d standard normal distribution $\{Z_n\}$*, i.e. if* $\Gamma_{n-1} = 1$ *then* $Y_n = X_{n-1} + Z_n$*, while if* $\Gamma_{n-1} = -1$ *then* $Y_n = \frac{1}{(1/X_{n-1})+Z_n}$*. The adaptation is defined as*

$$\Gamma_n = -\Gamma_{n-1}\mathbb{I}(X_n^{\Gamma_{n-1}} < \frac{1}{n}) + \Gamma_{n-1}\mathbb{I}(X_n^{\Gamma_{n-1}} \geq \frac{1}{n}), \tag{2.4}$$

i.e. we change Γ *from 1 to* -1 *when* $X < 1/n$*, and change* Γ *from* -1 *to 1 when* $X > n$*, otherwise we do not change* Γ*.*

Remark 2.2.1. *From Equation (2.4), we have two implications:* $[\Gamma_n \neq \Gamma_{n-1}] = [X_n^{\Gamma_{n-1}} < 1/n]$ *and* $\mathbf{P}[X_n^{\Gamma_n} \geq 1/n] = 1$ *for* $n \geq 1$*.*

First we use R package function rcauchy() to generate $10,000$ values, and then take the absolute values of them as the initial points. Second run the algorithm in Example 2.2.1 for $100,000$ iterations. The $10,000$ values at the $100,000^{th}$ iteration are used to analyze. Withdraw the subsample from the $10,000$ values, which fall in

the interval $(0, 1]$. We plot the histogram and the estimated density of the subsample, and also plot the estimated density of the sample generated directly by the absolute value of R package function rcauchy(), see Figure 2.1[1]. Apparently, the estimated density of the sample generated by the algorithm in Example 2.2.1 stays over that generated directly from the target distribution π when $x > 0.1$.

Proposition 2.2.1. *The adaptive chain $\{X_n : n \geq 0\}$ defined in Example 2.2.1 does not converge weakly to $\pi(\cdot)$. Containment does not hold.*

First we show that Diminishing Adaptation holds.

Lemma 2.2.1. *For the adaptive chain $\{X_n : n \geq 0\}$ defined in Example 2.2.1, the adaptation is diminishing.*

Proof: For $\gamma = 1$, obviously the proposal density is $q_\gamma(x, y) = \varphi(y - x)$ where $\varphi(\cdot)$ is the density function of standard normal distribution. For $\gamma = -1$, the random variable $1/x + Z_n$ has the density $\varphi(y - 1/x)$ so the random variable $1/(1/x + Z_n)$ has the density $q_\gamma(x, y) = \varphi(1/y - 1/x)/y^2$.

The proposal density

$$q_\gamma(x, y) = \begin{cases} \varphi(y - x) & \gamma = 1 \\ \varphi(1/y - 1/x)/y^2 & \gamma = -1 \end{cases}$$

For $\gamma = 1$, the acceptance rate is $\min\left(1, \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}\right)\mathbb{I}(y \in \mathcal{X}) = \frac{1+x^2}{1+y^2}\mathbb{I}(y > 0)$. For $\gamma = -1$, the acceptance rate is $\min\left(1, \frac{\pi(y)q_\gamma(y,x)}{\pi(x)q_\gamma(x,y)}\right)\mathbb{I}(y \in \mathcal{X}) = \min\left(1, \frac{\frac{1}{1+y^2}\varphi(1/x-1/y)/x^2}{\frac{1}{1+x^2}\varphi(1/y-1/x)/y^2}\right)\mathbb{I}(y > 0) = \min\left(1, \frac{1+x^{-2}}{1+y^{-2}}\right)\mathbb{I}(y > 0)$.

So for $\gamma \in \mathcal{Y}$, the acceptance rate is

$$\alpha_\gamma(x, y) := \min\left(1, \frac{\pi(y)q_\gamma(y, x)}{\pi(x)q_\gamma(x, y)}\right)\mathbb{I}(y \in \mathcal{X}) = \min\left(1, \frac{1 + x^{2\gamma}}{1 + y^{2\gamma}}\right)\mathbb{I}(y \in \mathcal{X}). \quad (2.5)$$

From Equation (2.4), $[\Gamma_n \neq \Gamma_{n-1}] = [X_n^{\Gamma_{n-1}} < 1/n]$. Obviously the joint process

[1]The density estimate is plotted by R density function where kernel density function may generate some negative density. In this plot, the negative part is erased artificially.

$\{(X_n, \Gamma_n) : n \geq 0\}$ is a time inhomogeneous Markov chain. So

$$\begin{aligned}
&\mathbf{P}(\Gamma_n \neq \Gamma_{n-1}) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{P}(X_n^{\Gamma_{n-1}} < 1/n \mid X_{n-1} = x, \Gamma_{n-1} = \gamma)\mathbf{P}(X_{n-1} \in dx, \Gamma_{n-1} \in d\gamma) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} P_\gamma(x, [t > 0 : t^\gamma < 1/n])\mathbf{P}(X_{n-1} \in dx, \Gamma_{n-1} \in d\gamma) \\
&= \int_{[x^\gamma \geq 1/(n-1)]} P_\gamma(x, [t > 0 : t^\gamma < 1/n])\mathbf{P}(X_{n-1} \in dx, \Gamma_{n-1} \in d\gamma)
\end{aligned}$$

where the second equality is from Equation (1.5), and the last equality is from $\mathbf{P}(X_n^{\Gamma_n} \geq 1/n) = 1$ implied by Equation (2.4).
So for any $(x, \gamma) \in [(t, s) \in \mathcal{X} \times \mathcal{Y} : t^s \geq 1/(n-1)]$,

$$P_\gamma(x, [t > 0 : t^\gamma < 1/n]) = \int_0^\infty \mathbb{I}(y^\gamma < 1/n) q_\gamma(x, y) dy = \int_{-x^\gamma}^{-x^\gamma + 1/n} \varphi(z) dz.$$

Since $-x^\gamma + 1/n < 0$,

$$\frac{1}{n}\varphi(-x^\gamma) \leq P_\gamma(x, [t > 0 : t^\gamma < 1/n]) \leq \frac{\varphi(0)}{n}. \tag{2.6}$$

We have that

$$\mathbf{P}(\Gamma_n \neq \Gamma_{n-1}) \leq \frac{1}{\sqrt{2\pi n}}. \tag{2.7}$$

Therefore, for any $\epsilon > 0$,

$$\mathbf{P}\left(\sup_{x\in\mathcal{X}} \left\| P_{\Gamma_n}(x, \cdot) - P_{\Gamma_{n-1}}(x, \cdot) \right\|_{\mathrm{TV}} > \epsilon\right) \leq \mathbf{P}(\Gamma_n \neq \Gamma_{n-1}) \to 0.$$

□

From Equation (2.5), at the n^{th} iteration, the acceptance rate is $\alpha_{\Gamma_{n-1}}(X_{n-1}, Y_n) = \min\left(1, \frac{1+X_{n-1}^{2\Gamma_{n-1}}}{1+Y_n^{2\Gamma_{n-1}}}\right)\mathbb{I}(Y_n > 0)$. Let us denote $\tilde{Y}_n := Y_n^{\Gamma_{n-1}}$ and $\tilde{X}_n := X_n^{\Gamma_n}$. The acceptance rate is equal to $\min\left(1, \frac{1+\tilde{X}_{n-1}^2}{1+\tilde{Y}_n^2}\right)\mathbb{I}(\tilde{Y}_n > 0)$. From Equation (2.4), $X_n^{\Gamma_n} =$

$X_n^{-\Gamma_{n-1}}\mathbb{I}(X_n^{\Gamma_{n-1}} < 1/n) + X_n^{\Gamma_{n-1}}\mathbb{I}(X_n^{\Gamma_{n-1}} \geq 1/n)$.

$$[Y_n^{\Gamma_{n-1}} < 1/n] = [\tilde{Y}_n < 1/n] \text{ and } Y_n^{\Gamma_n} = \tilde{Y}_n^{-1}\mathbb{I}(\tilde{Y}_n < 1/n) + \tilde{Y}_n\mathbb{I}(\tilde{Y}_n \geq 1/n).$$

When Y_n is accepted, i.e. $X_n = Y_n$,

$$[X_n^{\Gamma_{n-1}} < 1/n] = [\tilde{Y}_n < 1/n] \text{ and } X_n^{\Gamma_n} = \tilde{Y}_n^{-1}\mathbb{I}(\tilde{Y}_n < 1/n) + \tilde{Y}_n\mathbb{I}(\tilde{Y}_n \geq 1/n).$$

On the other hand, from Equation (2.3), the conditional distribution $\tilde{Y}_n \mid \tilde{X}_{n-1}$ is $N(\tilde{X}_{n-1}, 1)$.

From the above discussion, the chain $\{\tilde{X}_n : n \geq 0\}$ can be constructed according to the following procedure. Define the independent random variables $Z_n \overset{\text{iid}}{\sim} N(0,1)$, $U_n \overset{\text{iid}}{\sim} \text{Bernoulli}(0.5)$, and $T_n \overset{\text{iid}}{\sim} \text{Unif}(0,1)$.

Let $\tilde{X}_0 = X_0^{\Gamma_0}$. At each time $n \geq 1$, define the variable

$$\tilde{Y}_n := \tilde{X}_{n-1} - U_n |Z_n| + (1 - U_n) |Z_n| . \tag{2.8}$$

Clearly, $-U_n |Z_n| + (1 - U_n) |Z_n| \overset{\text{d}}{=} N(0,1)$ ($\overset{\text{d}}{=}$ means equal in distribution).

If $T_n < \min\left(1, \frac{1+\tilde{X}_{n-1}^2}{1+\tilde{Y}_n^2}\right)\mathbb{I}(\tilde{Y}_n > 0)$ then

$$\tilde{X}_n = \mathbb{I}(\tilde{Y}_n < 1/n)\tilde{Y}_n^{-1} + \mathbb{I}(\tilde{Y}_n \geq 1/n)\tilde{Y}_n; \tag{2.9}$$

otherwise $\tilde{X}_n = \tilde{X}_{n-1}$.

Note that:

1. The process $\tilde{X}$ is a time inhomogeneous Markov chain.
2. $\mathbf{P}(\tilde{X}_n \geq 1/n) = 1$ for $n \geq 1$.
3. At the time n, U_n indicates the proposal direction ($U_n = 0$: try to jump towards infinity; $U_n = 1$: try to jump towards zero). $|Z_n|$ specifies the step size if the proposal value Y_n is accepted. T_n is used to check whether the proposal value Y_n is accepted or not. When $U_n = 1$ and $\tilde{Y}_n > 0$, Equation (2.9) is always run.

For two integers $0 \leq s \leq t$ and a process X and a set $A \subset \mathcal{X}$, denote $[X_{s:t} \in A] := [X_s \in A; X_{s+1} \in A; \cdots; X_t \in A]$ and $s : t := \{s, s+1, \cdots, t\}$. For a value $x \in \mathbb{R}$, denote the largest integer less than x by $[x]$.

In the following proofs for the example, we use the notation in the procedure of

constructing the process $\tilde{X}$.

Lemma 2.2.2. *Let* $a = \left(\frac{1}{2} - \frac{7\sqrt{2}}{12\sqrt{\pi}}\right)^{-2}$. *Given* $0 < r < 1$, *for* $[x] > 12^{\frac{1}{1-r}}$

$$\mathbf{P}\left(\exists i \in (k+1) : (k + [x]^{1+r}),\ \tilde{X}_i < x/2 \mid \tilde{X}_k = x\right) \leq \frac{[x]^{1+r}}{\left(\frac{[x]}{2} - \frac{7\sqrt{2}[x]^r}{\sqrt{\pi}}\right)^2} \leq \frac{a}{[x]^{1-r}}.$$

Proof: The process $\{\tilde{X}_j : j \geq 0\}$ is generated through the underlying processes $\{(\tilde{Y}_j, Z_j, U_j, T_j) : j \geq 1\}$ defined in Equation (2.8) - Equation (2.9). Conditional on $[\tilde{X}_k = x]$, we can construct an auxiliary chain $\{B_j : j \geq k\}$ that behaves like an asymmetric random walk until $\tilde{X}$ reaches below $x/2$, and B is always dominated from above by $\tilde{X}$.

It is defined as that $B_k = \tilde{X}_k$; For $j > k$, if $\tilde{X}_{j-1} < x/2$ then $B_j := \tilde{X}_j$, otherwise
1. If proposing towards zero ($U_j = 1$) then B also jumps in the same direction with the step size $|Z_j|$ (in this case, the acceptance rate $\min\left(1, \frac{1+\tilde{X}_{j-1}^2}{1+\tilde{Y}_j^2}\right)$ is equal to 1);
2. If proposing towards infinity ($U_j = 0$), then B_j is assigned the value $B_{j-1} + |Z_j|$ (the jumping direction of B at the time j is same as $\tilde{X}$) with the acceptance rate $\frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}$ (independent of $\tilde{X}_{j-1}$), i.e. for $j > k$,

$$B_j := \mathbb{I}(\tilde{X}_{j-1} < x/2)\tilde{X}_j + \mathbb{I}(\tilde{X}_{j-1} \geq x/2)\,(B_{j-1} - I_j(x)) \tag{2.10}$$

where

$$I_j(x) := U_j\,|Z_j| - (1 - U_j)\,|Z_j|\,\mathbb{I}\left(T_j < \frac{1 + (x/2)^2}{1 + (x/2 + |Z_j|)^2}\right). \tag{2.11}$$

Note that
1. $\{Z_j, U_j, T_j : j > k\}$ are independent so $\{I_j(x) : j > k\}$ are independent.
2. When $\tilde{X}_{j-1} > x/2$ and $U_j = 0$ (proposing towards infinity), the acceptance rate $1 > \frac{1+\tilde{X}_{j-1}^2}{1+\tilde{Y}_j^2} \geq \frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}$, so that $\left[T_j < \frac{1+(x/2)^2}{1+(x/2+|Z_j|)^2}\right] \subset \left[T_j < \frac{1+\tilde{X}_{j-1}^2}{1+\tilde{Y}_j^2}\right]$ which is equivalent to $[B_j - B_{j-1} = |Z_j|] \subset [\tilde{X}_j - \tilde{X}_{j-1} = |Z_j|]$. Therefore, B is always dominated from above by $\tilde{X}$.

Conditional on $[\tilde{X}_k = x]$,

$$[\exists i \in (k+1) : (k + [x]^{1+r}),\ \tilde{X}_i < x/2] \subset [\exists i \in (k+1) : (k + [x]^{1+r}),\ B_i < x/2]$$

and for $i \in (k+1) : (k + [x]^{1+r})$,

$$\begin{aligned}
&[B_{k:(i-1)} \geq x/2; B_i < x/2] \\
\subset &[B_k \geq x/2; B_k - \sum_{l=k+1}^{t-1} I_l(x) \geq x/2 \text{ for all } t \in (k+1) : i; B_k - \sum_{l=k+1}^{i} I_l(x) < x/2].
\end{aligned}$$

So,

$$\begin{aligned}
&\mathbf{P}\left(\exists i \in (k+1) : (k + [x]^{1+r}),\ \tilde{X}_i < x/2 \mid \tilde{X}_k = x\right) \\
\leq &\mathbf{P}\left(\exists i \in (k+1) : (k + [x]^{1+r}),\ B_k - \sum_{j=k+1}^{i} I_j(x) < x/2 \mid B_k = x\right) \\
\leq &\mathbf{P}(\max_{l \in 1:[x]^{1+r}} \tilde{S}_l > x/2) \\
= &\mathbf{P}(\max_{l \in 1:q} \tilde{S}_l > q^{1/(1+r)}/2)
\end{aligned}$$

where $\tilde{S}_0 = 0$ and $\tilde{S}_l = \sum_{j=1}^{l} I_{k+j}(x)$ and $q = [x]^{1+r}$. $\{I_j(x) : k < j \leq k+l\}$ and B_k are independent so that the right hand side of the above equation is independent of k.

By some algebra,

$$0 \leq \mathbf{E}[I_i(x)] = \frac{1}{2}\mathbf{E}\left[\frac{|Z_i|^2 (x + |Z_i|)}{1 + (x/2 + |Z_i|)^2}\right] \leq \frac{2}{x}\mathbf{E}\left[|Z_i|^2 (1 + |Z_i|)\right] < \frac{7\sqrt{2}}{\sqrt{\pi}x},$$

$$\mathrm{Var}[I_i(x)] = \frac{1}{2} + \frac{1}{2}\mathbf{E}\left[|Z_i|^2 \frac{1 + (x/2)^2}{1 + (x/2 + |Z_i|)^2}\right] - \frac{1}{4}\left(\mathbf{E}\left[\frac{|Z_i|^2 (x + |Z_i|)}{1 + (x/2 + |Z_i|)^2}\right]\right)^2 \in [0, 1].$$

Let $\mu_l = \mathbf{E}[\tilde{S}_l]$ and $S_l = \tilde{S}_l - \mu_l$ and note that μ_l is increasing as l increases, and $\mu_q \in [0, \frac{7\sqrt{2}q}{\sqrt{\pi}}]$. So $\{S_i : i = 1, \cdots, q\}$ is a Martingale. By Kolmogorov Maximal

Inequality,

$$\begin{aligned}\mathbf{P}(\max_{l\in 1:q}\tilde{S}_l > q^{1/(1+r)}/2) \leq & \mathbf{P}(\max_{l\in 1:q} S_l > q^{1/(1+r)}/2 - \mu_q) \\ \leq & \frac{q\mathrm{Var}[I_k(x)]}{(q^{1/(1+r)}/2-\mu_q)^2} \\ \leq & \frac{[x]^{1+r}}{\left(\frac{[x]}{2}-\frac{7\sqrt{2}[x]^r}{\sqrt{\pi}}\right)^2} < \frac{a}{[x]^{1-r}}.\end{aligned}$$

The last second inequality is from $[x] > 12^{\frac{1}{1-r}} > \left(\frac{14\sqrt{2}}{\sqrt{\pi}}\right)^{\frac{1}{1-r}}$ implying $\frac{[x]}{2} > \frac{7\sqrt{2}[x]^r}{\sqrt{\pi}}$. □

Proof of Proposition 2.2.1: Assume that X_n converges weakly to $\pi(\cdot)$. Take some $c > 1$ such that for the set $D = (1/c, c)$, $\pi(D) = 9/10$. Taking a $r \in (0,1)$, there exists $N > 2c \vee 12^{\frac{1}{1-r}} \vee \frac{a}{0.5}^{\frac{1}{1-r}} \vee 2^{1/r}\exp(\frac{1}{0.8\varphi(-c)r})$ (a is defined in Lemma 2.2.2) such that for any $n > N+1$, $\mathbf{P}(X_n \in D) > 0.8$. Since $[X_n \in D] = [X_n^{\Gamma_n} \in D]$ and $X^\Gamma \stackrel{\mathrm{d}}{=} \tilde{X}$, $\mathbf{P}(\tilde{X}_n \in D) > 0.8$. So, $\mathbf{P}(\tilde{X}_n > \frac{n}{2}) < 0.2$ for $n > N$.

Let $m = \exp(\frac{1}{0.8\varphi(-c)})(n+1) - 1$ that implies $m > n$, $m - n < n^{1+r}$ (because $n > 2^{1/r}\exp(\frac{1}{0.8\varphi(-c)r})$), and $\log(\frac{m+1}{n+1}) = \frac{1}{0.8\varphi(-c)}$. Then

$$0.2 > \mathbf{P}(\tilde{X}_m > \frac{n}{2}) \geq \sum_{j=n}^{m-1}\mathbf{P}(\tilde{X}_j \in D; \tilde{Y}_{j+1} < \frac{1}{j+1}; \tilde{X}_{(j+1):m} > \frac{n}{2}). \tag{2.12}$$

From Equation (2.8) and Equation (2.9), $[\tilde{Y}_{i+1} < \frac{1}{i+1}] = [\tilde{X}_{i+1} = \frac{1}{\tilde{Y}_{i+1}} > i+1]$ for any $i > 1$. Consider $j \in n : (m-1)$. Since $\tilde{X}$ is a time inhomogeneous Markov chain,

$$\begin{aligned}&\mathbf{P}\left(\tilde{X}_j \in D; \tilde{Y}_{j+1} < \frac{1}{j+1}; \tilde{X}_{(j+1):m} > n/2\right) \\ =& \mathbf{P}(\tilde{X}_j \in D)\mathbf{P}\left(\tilde{X}_{j+1} = \tilde{Y}_{j+1} < \frac{1}{j+1} \mid \tilde{X}_j \in D\right) \\ &\mathbf{P}\left(\tilde{X}_{(j+2):m} > \frac{n}{2} \mid \tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j+1\right) \\ =& \mathbf{P}(\tilde{X}_j \in D)\mathbf{P}\left(\tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j+1 \mid \tilde{X}_j \in D\right) \\ &\left(1 - \mathbf{P}\left(\tilde{X}_t \leq n/2 \text{ for some } t \in (j+1) : m \mid \tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j+1\right)\right).\end{aligned}$$

From Equation (2.6), for any $x \in D$,

$$\mathbf{P}(\tilde{Y}_{j+1} < \frac{1}{j+1} \mid \tilde{X}_j = x) = P_1(x, \{t \in \mathcal{X} : t < 1/(j+1)\}) \in \left[\frac{\varphi(-c)}{j+1}, \frac{\varphi(0)}{j+1}\right].$$

So,

$$\mathbf{P}(\tilde{Y}_{j+1} < \frac{1}{j+1} \mid \tilde{X}_j \in D) \geq \frac{\varphi(-c)}{j+1}.$$

Hence, for $x > j+1$,

$$\begin{aligned}
&\mathbf{P}\left(\tilde{X}_t \leq n/2 \text{ for some } t \in (j+1) : m \mid \tilde{X}_{j+1} = x\right) \\
\leq &\mathbf{P}\left(\tilde{X}_t \leq x/2 \text{ for some } t \in (j+1) : m \mid \tilde{X}_{j+1} = x\right) \\
\leq &\mathbf{P}\left(\tilde{X}_t \leq x/2 \text{ for some } t \in (j+1) : (j + [x]^{1+r}) \mid \tilde{X}_{j+1} = x\right) \\
\leq &\frac{a}{[x]^{1-r}} \leq \frac{a}{n^{1-r}},
\end{aligned}$$

because of $x/2 > n/2$, $m - n < n^{1+r}$, and Lemma 2.2.2. Thus,

$$\mathbf{P}\left(\tilde{X}_t \leq n/2 \text{ for some } t \in (j+1) : m \mid \tilde{X}_{j+1} = \frac{1}{\tilde{Y}_{j+1}} > j+1\right) \leq \frac{a}{n^{1-r}}.$$

Therefore,

$$\begin{aligned}
\mathbf{P}(\tilde{X}_m > \frac{n}{2}) \geq &0.8\varphi(-c)(1 - \frac{a}{n^{1-r}}) \sum_{j=n}^{m-1} \frac{1}{j+1} \\
\geq &0.8\varphi(-c)(1 - \frac{a}{n^{1-r}}) \log((m+1)/(n+1)) = (1 - \frac{a}{n^{1-r}}) > 0.5.
\end{aligned}$$

Contradiction! By Lemma 2.2.1, Containment does not hold. □

2.3 An Adaptive Metropolis Algorithm

For an adaptive MCMC algorithm, say that it is an adaptive Metropolis algorithm if at each iteration, one Metropolis sampler is chosen to do sampling. Many works had

been developed to analyze ergodicity of the adaptive Metropolis algorithm introduced by Haario et al. (2001). In the section, we use the adaptation in Haario's algorithm for a mixture target distribution.

First let us to define the proposal distribution at each iteration. Given the chain $X_0, \cdots, X_n \in \mathbb{R}^d$, the matrix

$$\Sigma_n = \frac{1}{n}\left(\sum_{i=0}^{n} X_i X_i^\top - (n+1)\overline{X}_n \overline{X}_n^\top\right), \tag{2.13}$$

where $\overline{X}_n = \frac{1}{n+1}\sum_{i=0}^{n} X_i$ is the current modified empirical estimate of the covariance structure of the target distribution based on the run so far. Then if $n \leq 2d$ then the proposal distribution $Q_n(x, \cdot) = N(x, (0.1)^2 I_d/d)$; For $n > 2d$, if Σ_n is positive definite then $Q_n(x, \cdot)$ is mixed by two multivariate normal distributions $N(x, (2.38)^2\Sigma_n/d)$ and $N(x, (0.1)^2 I_d/d)$ respectively with weights $1-\theta$ and θ, i.e.

$$Q_n(x, \cdot) = (1-\theta)N(x, (2.38)^2\Sigma_n/d) + \theta N(x, (0.1)^2 I_d/d), \tag{2.14}$$

otherwise $Q_n(x, \cdot) = N(x, (0.1)^2 I_d/d)$. The scaling parameter $(2.38)^2/d$ is adopted 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 were proved by Roberts et al. (1997); Roberts and Rosenthal (2001).

Consider a mixture of two normal distributions as the target distribution on $\mathbb{R}^2$ with the density function

$$\begin{aligned} t(x) =& \frac{1}{2\sqrt{|2\pi\Sigma_1|}} \exp(-(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)) + \\ & \frac{1}{2\sqrt{|2\pi\Sigma_2|}} \exp(-(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)), \end{aligned} \tag{2.15}$$

where $\mu_1 = (0,0)'$, $\mu_2 = (5,5)'$, $\Sigma_1 = diag(1,1)$, $\Sigma_2 = diag(0.01, 0.01)$.

The $t(x)$ has two modes respectively at μ_1 and μ_2. The mode at μ_2 is much taller than that at μ_1, see Figure 2.2.

Run the adaptive Metropolis algorithm with a little adjustment: after the first $2d$

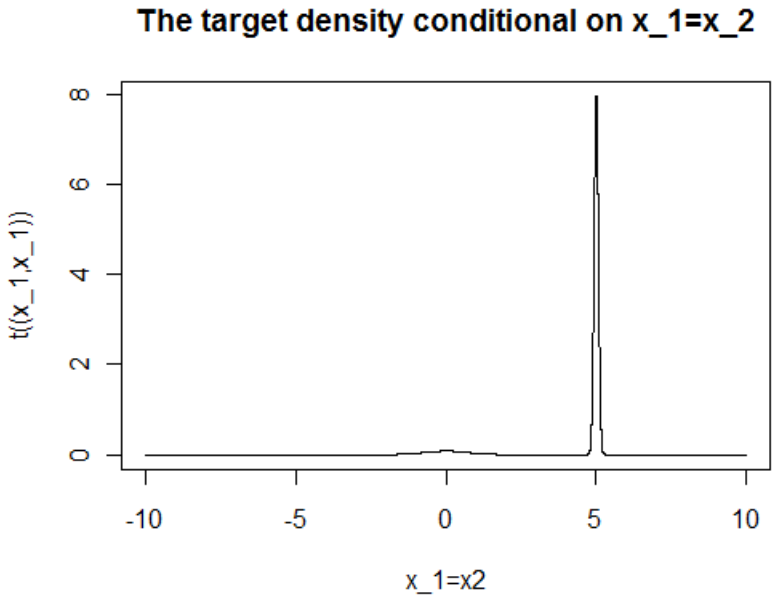

Figure 2.2: The marginal target density function on $x_1 = x_2$.

steps, the proposal distribution is

$$Q_n(x, \cdot) = (1 - \theta)N(x, (2.38)^2\Sigma_n/d) + \theta N(x, 4I_d/d) \tag{2.16}$$

where the parameter θ can be arbitrary in $(0, 1)$. For our implementation, $\theta = 1/3$. The variance of the fixed distribution (the second term of the right hand side in Equation (2.14)) in the mixture proposal is changed to $4I_d$. The reason is that at each proposal, there are some possibility to detect relatively large region where some modes may be hidden. After $1,000,000$ iterations, we got the sample data concentrating on two balls, see the left plot in Figure 2.3. See the estimated marginal density function on $x_1 = x_2$, the right plot in Figure 2.3. The average acceptance rate over every 50 steps is not stable, disturbing between 0.00 and 0.45, see the left plot in Figure 2.4. In the right plot in Figure 2.4, if the sample state is located in the vicinity of the high mode then 1 is evaluated; if one sample sate is located in the vicinity of the low mode then 0 is evaluated. From the plot, we can see the sample chain frequently jumps in the two modes.

When target distributions are defined in the high dimensional space, the adap-

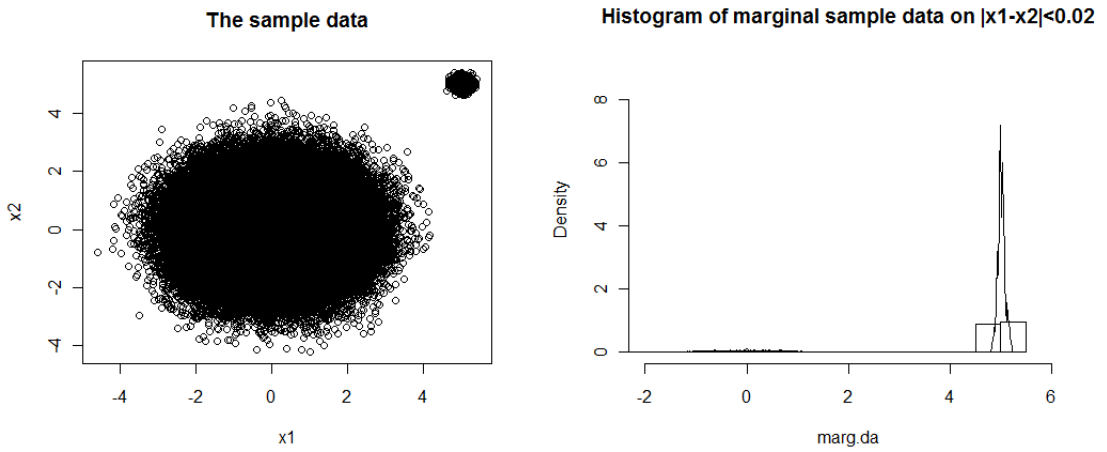

Figure 2.3: Left: The sample data over $1,000,000$ iterations; Right: the estimate marginal density on $X_1 = X_2$.

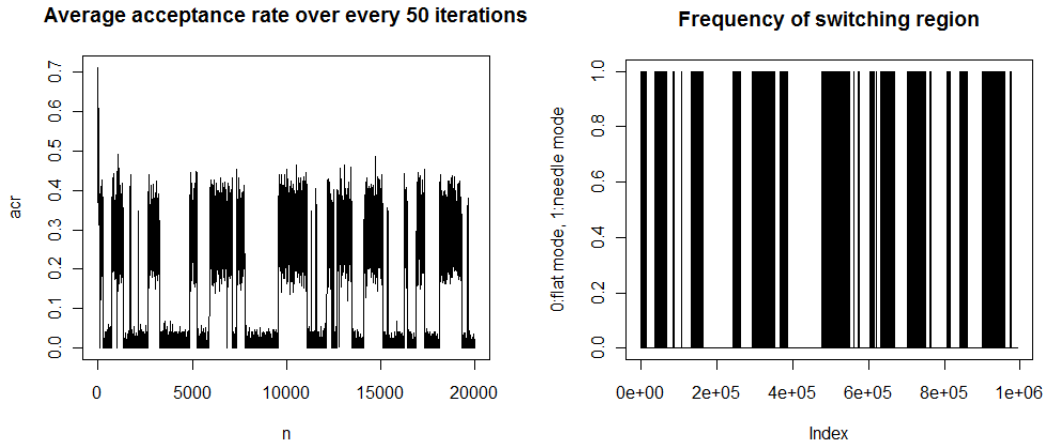

Figure 2.4: Left: the average acceptance rate over every 50 iterations; Right: The frequency of switching regions.

tation part (the first term of the right hand side in Equation (2.14)) of the mixture proposal will play a significant role. This part can learn the true variance of target distributions. Relatively, artificially adjusting parameters of the proposal distribution used in non-adaptive MCMC algorithms will be awkward.

Chapter 3

Simultaneous Polynomial Ergodicity

In the section will study *Simultaneous Polynomial Ergodicity*. We assume that under some regular conditions about target distributions, all the transition kernels in $\{P_\gamma : \gamma \in \mathcal{Y}\}$ simultaneously satisfy a group of drift conditions, and have the uniform small set C in the sense of the m-step transition.

Suppose that Diminishing Adaptation and simultaneous polynomial ergodicity hold. We find that when either the number of drift conditions is greater than or equal to two, or the number of drift conditions having certain specific form is one, the adaptive MCMC algorithm is ergodic. For adaptive MCMC algorithms with Markovian Adaptation (the joint process $\{(X_n, \Gamma_n) : n \geq 0\}$ is Markovian), the algorithm satisfying Diminishing Adaptation and simultaneous polynomial ergodicity is ergodic without those restrictions, thanks to the results in Atchadé and Fort (2008). We also discuss some recent results related to this topic, and show that under certain additional condition, Containment is necessary for ergodicity of adaptive MCMC algorithms.

Yang (2008b) and Atchadé and Fort (2008) (AF) respectively tackle the Open problem 21 in Roberts and Rosenthal (2007). Yang assumes that all the transition kernels simultaneously satisfy the drift condition $P_\gamma V - V \leq -1 + b\mathbb{I}_C$, and the adaptive parameter space is compact under certain metric, and connects it with the regener-

ation decomposition to find the uniform bound of the distance $\left\|P_\gamma^n(x,\cdot)-\pi(\cdot)\right\|_{\mathrm{TV}}$ for all γ. Once this condition given, the distance $\left\|P_\gamma^n(x,\cdot)-\pi(\cdot)\right\|_{\mathrm{TV}}$ can be uniformly bounded by the test function. The boundedness of the test function sequence $\{V(X_n): n\geq 0\}$ can ensure Containment.

Under some situations, to directly check Containment may be quite hard. AF use the similar coupling method as that in Roberts and Rosenthal (2007) to prove an attractive result when an adaptive MCMC algorithm is restricted to Markovian Adaptation. They also assume that uniformly strongly aperiodicity, simultaneously drift condition in the weakest form $P_\gamma V - V \leq -1 + b\mathbb{I}_C$, and uniform convergence on any sublevel set of the test function $V(\cdot)$. The idea is that after the chain comes into some "big" sublevel set of the test function $V(x)$, apply the coupling method for ergodicity.

In Section 3.1 we discuss Yang's, and AF's conditions (respectively (Y1)-(Y4) and (M1)-(M3)) and results. In Section 3.2 we provide a necessary condition of ergodicity conditional on an additional condition. In Section 3.3 we show our main result.

3.1 Simultaneous Drift Conditions

Roberts and Rosenthal (2007) gave one condition: the Simultaneously Strongly Aperiodically Geometrically Ergodic condition which can ensure Containment. In the definition, the simultaneous drift conditions have the form: $P_\gamma V(x) \leq \lambda V(x) + b\mathbb{I}_C(x)$ for all $\gamma \in \mathcal{Y}$. However, if $\{\Gamma_n : n \geq 0\}$ is bounded in probability, Containment can be implied by that in each compact subset B of $\mathcal{Y}$, the drift conditions have the same form: $P_\gamma V_B(x) \leq \lambda_B V_B(x) + b_B\mathbb{I}_C(x)$. More generally, we give the the following result (a corollary of (Roberts and Rosenthal, 2007, Theorem 13)).

Corollary 3.1.1. *Suppose that the parameter space $\mathcal{Y}$ is a metric space, and the adaptive parameter $\{\Gamma_n : n \geq 0\}$ is bounded in probability; for any compact set $K \subset \mathcal{Y}$, for any $\epsilon > 0$, the local ϵ-convergence time*

$$\left\{\widetilde{M}_\epsilon(X_n) := \inf_m\{m \in \mathbb{N}^+ : \sup_{\gamma\in K}\left\|P_\gamma^m(X_n,\cdot)-\pi(\cdot)\right\|_{\mathrm{TV}} < \epsilon\} : n \geq 0\right\}$$

is bounded in probability. Diminishing Adaptation implies ergodicity of the adaptive

MCMC algorithm $\{X_n : n \geq 0\}$.

The proof is trivial and omitted.

Roberts and Rosenthal (2007) propose one open problem in Roberts and Rosenthal (2007). Yang (2008b) gives the following conditions to tackle the problem:

Y1: There exist a constant $\delta > 0$, and a set $C \in \mathcal{F}$, and a probability measure $\nu_\gamma(\cdot)$ for $\gamma \in \mathcal{Y}$, such that $P_\gamma(x, \cdot) \geq \delta \mathbb{I}_C(x)\nu_\gamma(\cdot)$ for $\gamma \in \mathcal{Y}$;

Y2: all kernels simultaneously satisfy the weakest drift condition: $P_\gamma V \leq V - 1 + b\mathbb{I}_C$, where $V : \mathcal{X} \to [1, \infty)$ and $\pi(V) < \infty$;

Y3: $\mathcal{Y}$ is compact under the metric $d(\gamma_1, \gamma_2) = \sup_{x \in \mathcal{X}} \|P_{\gamma_1}(x, \cdot) - P_{\gamma_2}(x, \cdot)\|_{TV}$;

Y4: the stochastic process $\{V(X_n) : n \geq 0\}$ is bounded in probability.

Theorem 3.1.1 (Yang (2008b))**.** *Suppose Diminishing Adaptation holds. The conditions (Y1)-(Y4) ensure ergodicity of adaptive MCMC algorithms.*

Remark 3.1.1.

1. *In Yang's proof, both (Y1) and (Y2) can ensure that each transition kernel is ergodic to* π*. Both (Y3) and (Y4) imply that the total variation distance between* P_γ *and* π *converges to zero uniformly on* $\mathcal{Y}$*.*

2. *The condition* $\pi(V) < \infty$ *is a relatively strong condition. For each* P_γ*, suppose that the chain* $\{X_n^{(\gamma)} : n \geq 0\}$ *is a time homogeneous Markov chain with the transition kernel* P_γ*. For any recurrent set* $A \subset \mathcal{X}$ *with* $\pi(A) > 0$*, by Meyn and Tweedie (1993) (MT) Proposition 10.4.9,* $\pi(V) = \int_A \pi(dy) E_\gamma \left[\sum_{i=0}^{\tau_A - 1} V(X_i^{(\gamma)}) | X_0^{(\gamma)} = y\right]$*. Assuming that there exists a small set* $C_1 \subset \mathcal{X}$ *with* $\sup_{x \in C_1} E_\gamma \left[\sum_{i=0}^{\tau_{C_1} - 1} V(X_i^{(\gamma)}) | X_0^{(\gamma)} = x\right] < \infty$*, denote* $U_\gamma(x) = E_\gamma \left[\sum_{i=0}^{\sigma_{C_1}} V(X_i^{(\gamma)}) | X_0^{(\gamma)} = x\right]$*. Hence, by MT Theorem 11.3.5,* $P_\gamma U_\gamma - U_\gamma \leq -V(x) + b_1 \mathbb{I}_{C_1}$ *where* $b_1 = \sup_{x \in C_1} E_\gamma \left[\sum_{i=0}^{\tau_{C_1} - 1} V(X_i^{(\gamma)}) | X_0^{(\gamma)} = x\right]$*. Suppose that there is a test function* V_1 *satisfying* $P_\gamma V_1 - V_1 \leq -V + b\mathbb{I}_{C_1}$ *and* $V_1(x)\mathbb{I}_{C_1}(x) \geq V(x)$*. By MT Proposition 11.3.2,* $U_\gamma(x) \leq V_1(x)$*. So,* $P_\gamma U_\gamma - U_\gamma \leq -1 + b\mathbb{I}_{C_1}$*. We will study the simultaneous drift condition with the form* $P_\gamma V_1 - V_1 \leq -V_0 + b\mathbb{I}_C$ *instead where the test functions* $V_0(x)$ *and* $V_1(x)$ *are uniform for every* P_γ*. Under this situation, the condition (Y3) are unnecessary, and the condition (Y4) is implied (See Theorem 3.3.2, Remark 3.3.2).*

AF also give the following conditions to study the ergodicity of adaptive MCMC with Markovian Adaptation:

M1: there exists a probability measure $\nu(\cdot)$, a constant $\delta > 0$, and set $C \in \mathcal{F}$ such that $P_\gamma(x, \cdot) \geq \delta \mathbb{I}_C(x)\nu(\cdot)$ for $\gamma \in \mathcal{Y}$;

M2: there exists a measurable function $V : \mathcal{X} \to [1, \infty)$ and a positive constant $b > 0$ such that for any $\gamma \in \mathcal{Y}$, $(P_\gamma V)(x) - V(x) \leq -1 + b\mathbb{I}_C(x)$;

M3: for any sublevel set $\mathcal{D}_l = \{x \in \mathcal{X} : V(x) \leq l\}$ of V,

$$\lim_{n\to\infty} \sup_{\mathcal{D}_l \times \mathcal{Y}} \left\| P_\gamma^n(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} = 0.$$

Theorem 3.1.2 (Atchadé and Fort (2008))**.** *Suppose Diminishing Adaptation holds. The conditions (M1)-(M3) imply ergodicity of adaptive MCMC algorithms with Markovian Adaptation.*

Remark 3.1.2.

1. *Since*

$$\left| P_{(x_0,\gamma_0)}(V(X_n) > M) - \pi(D_M^c) \right| \leq \left\| P_{(x_0,\gamma_0)}(X_n \in \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}},$$

M can be taken extremely large such that $\pi(D_M^c) < \epsilon$. (M1-M3) and Diminishing Adaptation imply that R.H.S. of the above equation converges to zero. So, $\{V(X_n) : n \geq 0\}$ is bounded in probability.

2. *In Section 3.2 we show that under certain condition, Containment is a necessary condition of ergodicity of adaptive MCMC provided that (M3) holds. From another view, AF's proof does apply the coupling method to check Containment by using Diminishing Adaptation and simultaneous drift conditions.*

3.2 The necessary condition for ergodicity

In this section, we study the necessary condition for ergodicity of adaptive algorithms. The half-Cauchy example shows that Diminishing Adaptation alone can not

ensure ergodicity. In that example, Containment is not satisfied. Example 2.1.1 shows that Containment is also not necessary. In the following theorem, we prove that under certain additional condition similar to (M3), Containment is necessary for ergodicity of adaptive algorithms.

Theorem 3.2.1 (The necessity of Containment)**.** *Suppose that there exists an increasing sequence of sets $\mathcal{D}_k \uparrow \mathcal{X}$ on the state space $\mathcal{X}$, such that for any $k > 0$,*

$$\lim_{n\to\infty} \sup_{\mathcal{D}_k \times \mathcal{Y}} \left\| P_\gamma^n (x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} = 0. \tag{3.1}$$

If the adaptive MCMC algorithm is ergodic then Containment holds.

Proof: Fix $\epsilon > 0$. For any $\delta > 0$, take $K > 0$ such that $\pi(\mathcal{D}_K^c) < \delta/2$. For the set $\mathcal{D}_K$, there exists M such that

$$\sup_{\mathcal{D}_K \times \mathcal{Y}} \left\| P_\gamma^M (x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} < \epsilon.$$

Hence, for any $(x_0, \gamma_0) \in \mathcal{X} \times \mathcal{Y}$, by the ergodicity of the adaptive MCMC $\{X_n : n \geq 0\}$, there exists some $N > 0$ such that $n > N$,

$$\left| P_{(x_0,\gamma_0)}(X_n \in \mathcal{D}_K^c) - \pi(\mathcal{D}_K^c) \right| < \delta/2.$$

So, for $(X_n, \Gamma_n) \in (\mathcal{D}_K, \mathcal{Y})$,

$$[X_n \in \mathcal{D}_K] = [(X_n, \Gamma_n) \in \mathcal{D}_K \times \mathcal{Y}] \subset [M_\epsilon(X_n, \Gamma_n) \leq M] .$$

Hence,

$$\begin{aligned} & P_{(x_0,\gamma_0)} \left(M_\epsilon(X_n, \Gamma_n) > M \right) \\ \leq\ & P_{(x_0,\gamma_0)} \left((X_n, \Gamma_n) \in (\mathcal{D}_K \times \mathcal{Y})^c \right) \\ =\ & P_{(x_0,\gamma_0)} \left(X_n \in \mathcal{D}_K^c \right) \\ \leq\ & \left| P_{(x_0,\gamma_0)} \left(X_n \in \mathcal{D}_K^c \right) - \pi(\mathcal{D}_K^c) \right| + \pi(\mathcal{D}_K^c) < \delta. \end{aligned}$$

Therefore, Containment holds. □

Corollary 3.2.1. *Suppose that the parameter space $\mathcal{Y}$ is a metric space, and the adaptive scheme $\{\Gamma_n : n \geq 0\}$ is bounded in probability. Suppose that there exists an increasing sequence of sets $(\mathcal{D}_k, \mathcal{Y}_k) \uparrow \mathcal{X} \times \mathcal{Y}$ such that any $k > 0$,*

$$\lim_{n\to\infty} \sup_{\mathcal{D}_k \times \mathcal{Y}_k} \left\| P_\gamma^n (x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} = 0.$$

If the adaptive MCMC algorithm is ergodic then Containment holds.

Proof: Using the same technique in Theorem 3.2.1, for large enough $M > 0$,

$$\begin{aligned} & P_{(x_0,\gamma_0)} \left(M_\epsilon(X_n, \Gamma_n) > M \right) \\ \leq\ & P_{(x_0,\gamma_0)} \left((X_n, \Gamma_n) \in (D_k \times \mathcal{Y}_k)^c \right) \\ \leq\ & P_{(x_0,\gamma_0)} \left(X_n \in D_k^c \right) + P_{(x_0,\gamma_0)} \left(\Gamma_n \in \mathcal{Y}_k^c \right) \\ \leq\ & \left| P_{(x_0,\gamma_0)} \left(X_n \in \mathcal{D}_K^c \right) - \pi(\mathcal{D}_K^c) \right| + \pi(\mathcal{D}_K^c) + P_{(x_0,\gamma_0)} \left(\Gamma_n \in \mathcal{Y}_k^c \right). \end{aligned}$$

Since $\{\Gamma_n : n \geq 0\}$ is bounded in probability, the result holds. □

Example 2.1.1 is a counter example to explain that Containment is not necessary. It is easy to check that the additional conditions in Theorem 3.2.1 and Corollary 3.2.1 are not satisfied.

3.3 Simultaneous Polynomial Ergodicity

Although ergodicity of adaptive MCMC algorithms, to some degree, is solved in Yang (2008b) and Atchadé and Fort (2008), there are still some properties unknown about simultaneous polynomial ergodicity. In the section, we find that the conditions (Y4) and (M3) are implied for the adaptive MCMC with simultaneous polynomial ergodicity. Before studying it, let us recall the result about a quantitative bound for a time-homogeneous Markov chain with polynomial convergence rate by Fort and Moulines (2000b) (FM).

Theorem 3.3.1 (Fort and Moulines (2000b))**.** *Suppose that the time-homogeneous transition kernel P satisfies the following conditions:*

- *P is π-irreducible for an invariant probability measure π;*

- *There exist some sets $C \in \mathcal{B}(\mathcal{X})$ and $D \in \mathcal{B}(\mathcal{X})$, $C \subset D$, $\pi(C) > 0$ and an integer $m \geq 1$, such that for any $(x, x') \in \triangle := C \times D \cup D \times C$, $A \in \mathcal{B}(\mathcal{X})$,*

$$P^m(x, A) \wedge P^m(x', A) \geq \rho_{x,x'}(A) \tag{3.2}$$

where $\rho_{x,x'}$ is some measure on $\mathcal{X}$ for $(x, x') \in \triangle$, and $\epsilon^- := \inf_{(x,x') \in \triangle} \rho_{x,x'}(\mathcal{X}) > 0$.

- *Let $q \geq 1$. There exist some measurable functions $0 < V_0 \leq V_1 \leq \cdots \leq V_q : \mathcal{X} \to \mathbb{R}^+ \setminus \{0\}$, and for $k \in \{0, 1, \ldots, q-1\}$, for some constants $0 < a_k < 1$, $b_k < \infty$, and $c_k > 0$ such that*

$$PV_{k+1}(x) \leq V_{k+1}(x) - V_k(x) + b_k \mathbb{I}_C(x), \inf_{x \in \mathcal{X}} V_k(x) \geq c_k > 0,$$

$$V_k(x) - b_k \geq a_k V_k(x), x \in D^c, \tag{3.3}$$

$$\sup_D V_q < \infty.$$

- *$\pi(V_q^\beta) < \infty$ for some $\beta \in (0, 1]$.*

Then, for any $x \in \mathcal{X}$, $n \geq m$,

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq \min_{1 \leq l \leq q} B_l^{(\beta)}(x, n), \tag{3.4}$$

with

$$B_l^{(\beta)}(x, n) = \frac{\epsilon^+ \left\langle (I - A_m^{(\beta)})^{-1} \delta_x \otimes \pi(W^\beta), e_l \right\rangle}{S(l, n+1-m)^\beta + \sum_{j \geq n+1-m} (1 - \epsilon^+)^{j-(n-m)} (S(l, j+1)^\beta - S(l, j)^\beta)},$$

where $< \cdot, \cdot >$ denotes the inner product in $\mathbb{R}^{q+1}$, $\{e_l\}$, $0 \leq l \leq q$ is the canonical basis on $\mathbb{R}^{q+1}$, I is the identity matrix;

$$\delta_x \otimes \pi(W^\beta) := \int \delta_x(dy) \pi(dy') W^\beta(y, y')$$

where $W^\beta(x,x') := \left(W_0^\beta(x,x'),\cdots,W_q^\beta(x,x')\right)^T$ *with* $W_0(x,x') := 1$ *and*

$$W_l(x,x') = \mathbb{I}_\triangle(x,x') + \mathbb{I}_{\triangle^c}(x,x')\left(\prod_{k=0}^{l-1} a_k\right)^{-1}(m(V_0))^{-1}\left(V_l(x)+V_l(x')\right) \text{ for } 1\le l\le q$$

where $m(V_0) := \inf_{(x,x')\in\triangle^c}\{V_0(x)+V_0(x')\}$*;*

$$S(0,k) := 1 \text{ and } S(i,k) := \sum_{j=1}^{k} S(i-1,j), i\ge 1;$$

$$A_m^{(\beta)} := \begin{pmatrix} A_m^{(\beta)}(0) & 0 & \cdots & 0 & 0 \\ A_m^{(\beta)}(1) & A_m^{(\beta)}(0) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_m^{(\beta)}(q-1) & A_m^{(\beta)}(q-2) & \cdots & A_m^{(\beta)}(0) & 0 \\ A_m^{(\beta)}(q) & A_m^{(\beta)}(q-1) & \cdots & A_m^{(\beta)}(1) & A_m^{(\beta)}(0) \end{pmatrix},$$

where $A_m^{(\beta)}(l) := \sup_{(x,x')\in\triangle}\sum_{i=0}^{l} S(i,m)^\beta\left(1-\rho_{x,x'}(\mathcal{X})\right)\int R_{x,x'}(x,dy)R_{x,x'}(x',dy')W_{l-i}^\beta(y,y')$*, where the residual kernel*

$$R_{x,x'}(u,dy) := (1-\rho_{x,x'}(\mathcal{X}))^{-1}\left(P_\gamma^m(u,dy)-\rho_{x,x'}(dy)\right);$$

and $\epsilon^+ := \sup_{(x,x')\in\triangle}\rho_{x,x'}(\mathcal{X})$*.*

Remark 3.3.1. *In the* $B_l^{(\beta)}(x,n)$*,* ϵ^+ *depends on the set* $\triangle$ *and the measure* $\rho_{x,x'}$*; the matrix* $(I-A_m^{(\beta)})^{-1}$ *depends on the set* $\triangle$*, the transition kernel* P*,* $\rho_{x,x'}$ *and the test functions* V_k*;* $\delta_x\otimes\pi(W^\beta)$ *depends on the set* $\triangle$ *and the test functions* V_k*.*
Consider the special case of the theorem: $\rho_{x,x'}(dy) = \delta\nu(dy)$ *where* ν *is a probability measure with* $\nu(C)>0$*, and* $\triangle := C\times C$*.*
1. $\epsilon^+ = \epsilon^- = \delta$*.*
2. $I-A_m^{(\beta)}$ *is a lower triangle matrix so* $(I-A_m^{(\beta)})^{-1} = \left(b_{ij}^{(\beta)}\right)_{i,j=1,\ldots,q+1}$ *is also a lower triangle matrix, and fixing* $k\ge 0$ *all* $b_{i,i-k}^{(\beta)}$ *are equal.* $b_{ii}^{(\beta)} = \frac{1}{1-A_m^{(\beta)}(0)}$*. For* $i>j$*,* $b_{ij}^{(\beta)}$ *is the polynomial combination of* $A_m^{(\beta)}(0),\cdots,A_m^{(\beta)}(i+1)$ *divided by* $(1-A_m^{(\beta)}(0))^i$*. By some algebra, we can obtain that* $b_{21}^{(\beta)} = \frac{A_m^{(\beta)}(1)}{(1-A_m^{(\beta)}(0))^2}$*. So, by calculating* $B_1^{(\beta)}(x,n)$*, we can get the quantitative bound with a simple form.* $B_1^{(\beta)}(x,n)$ *only involves two test*

functions $V_0(x)$ and $V_1(x)$.

Remark 3.3.2. *From Equation (3.3), $V_0(x) \geq b_0/(1-\alpha_0) > b_0$ because $0 < \alpha_0 < 1$. Consider the drift condition: $PV_1 - V_1 \leq -V_0 + b_0\mathbb{I}_C$. Since $\pi P = \pi$, $\pi(V_0) \leq b_0\pi(C) \leq b_0$. Hence, the V_0 in the theorem cannot be constant.*

Remark 3.3.3. *Without the condition $\pi(V_q^\beta) < \infty$, the bound in Equation (3.4) can also be obtained. However, the bound is possibly infinity. The subscript l of $B_l^{(\beta)}(x,n)$ and β can explain the polynomial rate ($S(l, n+1-m)^\beta = \mathrm{O}((n+1-m)^{l\beta})$). It can be observed that $B_l^{(\beta)}(x,n)$ involves test functions $V_0(x), \cdots, V_l(x)$, and $\limsup_n n^{\beta l} B_l^{(\beta)}(x,n) < \infty$. Given $x \in \mathcal{X}$, the decaying rate of $B_l^{(\beta)}(x,n)$ is less than $\mathrm{O}(n^{-q\beta})$.*

3.3.1 Conditions

The following conditions are derived from Theorem 3.3.1, and some changes are added to apply for adaptive MCMC algorithms. Say that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is *simultaneously polynomially ergodic* (S.P.E.) if the conditions (A1)-(A4) are satisfied.

A1: each P_γ is ψ_γ-irreducible with stationary distribution $\pi(\cdot)$;

A2: there is a set $C \subset \mathcal{X}$, some integer $m \in \mathbb{N}$, some constant $\delta > 0$, and some probability measure $\nu_\gamma(\cdot)$ on $\mathcal{X}$ such that:

$$\pi(C) > 0, \text{ and } P_\gamma^m(x,\cdot) \geq \delta\mathbb{I}_C(x)\nu_\gamma(\cdot) \text{ for } \gamma \in \mathcal{Y}; \tag{3.5}$$

A3: there is $q \in \mathbb{N}$ and measurable functions: $V_0, V_1, \ldots, V_q : \mathcal{X} \to (0,\infty)$ where $V_0 \leq V_1 \leq \cdots \leq V_q$, such that for $k = 0, 1, \ldots, q-1$, there are $0 < \alpha_k < 1$, $b_k < \infty$, and $c_k > 0$ such that:

$$P_\gamma V_{k+1}(x) \leq V_{k+1}(x) - V_k(x) + b_k\mathbb{I}_C(x),\ V_k(x) \geq c_k \text{ for } x \in \mathcal{X} \text{ and } \gamma \in \mathcal{Y}; \tag{3.6}$$

$$V_k(x) - b_k \geq \alpha_k V_k(x) \text{ for } x \in C^c; \tag{3.7}$$

$$\sup_{x\in C} V_q(x) < \infty. \tag{3.8}$$

A4: $\pi(V_q^\beta) < \infty$ for some $\beta \in (0, 1]$.

Remark 3.3.4. *From MT Proposition 10.1.2, if P_γ is φ-irreducible, then P_γ is π-irreducible and the invariant measure π is a maximal irreducibility measure.*

Remark 3.3.5. *In Theorem 3.3.1, there is one condition (Equation (3.2)) ensuring the splitting technique. Here we consider the special case of that condition: $\rho_{x,x'}(dy) = \delta\nu_\gamma(dy)$ and $\triangle = C \times C$. Thus, by Remark 3.3.1, the bound of $\left\|P_\gamma^n(x,\cdot) - \pi(\cdot)\right\|_{TV}$ depends on C, m, the minorization constant δ, $\pi(\cdot)$, ν_γ, and test functions $V_l(x)$ so we assume that C, m and δ are uniform for all the transition kernels.*

Remark 3.3.6. *For $x \in C$, $\nu_\gamma(V_l) \le \frac{1}{\delta} P_\gamma^m V_l(x) \le \frac{1}{\delta} \sup_{x\in C} V_l(x) + \frac{mb_{l-1}}{\delta}$.*

3.3.2 Main Result

Before showing the main result, we give one lemma used in the proof of the main result.

Lemma 3.3.1. *Suppose that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.P.E.. If the stochastic process $\{V_l(X_n) : n \ge 0\}$ is bounded in probability for some $l \in \{1, \ldots, q\}$, then Containment is satisfied.*

The proof is in Section 3.3.3.

Theorem 3.3.2. *Suppose an adaptive MCMC algorithm satisfies Diminishing Adaptation. Then, the algorithm is ergodic under any of the following cases:*

(i) *S.P.E., and the number q of simultaneous drift conditions is strictly greater than two;*

(ii) *S.P.E., and when the number q of simultaneous drift conditions is greater than or equal to two, there exists an increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $V_1(x) \le f(V_0(x))$;*

(iii) *Under the conditions* (A1) *and* (A2)*, there exist some positive constants $c > 0$, $b' > b > 0$, $\alpha \in (0, 1)$, and a measurable function $V(x) : \mathcal{X} \to \mathbb{R}+$ with $V(x) \ge 1$ and $\sup_{x\in C} V(x) < \infty$ such that*

$$P_\gamma V(x) - V(x) \le -cV^\alpha(x) + b\mathbb{I}_C(x) \text{ for } \gamma \in \mathcal{Y}, \tag{3.9}$$

$cV^{\alpha}(x)\mathbb{I}_{C^c}(x) \geq b'$;
(iv) *Under the condition* (A1), (A2), *and* (A4), *there exist some constant* $b' > b > 0$, *two measurable functions* $V_0 : \mathcal{X} \to \mathbb{R}^+$ *and* $V_1 : \mathcal{X} \to \mathbb{R}^+$ *with* $1 \leq V_0(x) \leq V_1(x)$ *and* $\sup_{x\in C} V_1(x) < \infty$ *such that*

$$P_{\gamma}V_1(x) - V_1(x) \leq -V_0(x) + b\mathbb{I}_C(x) \text{ for } \gamma \in \mathcal{Y}, \tag{3.10}$$

$V_0(x)\mathbb{I}_{C^c}(x) \geq b'$, *and the process* $\{V_1(X_n) : n \geq 0\}$ *is bounded in probability.*
Remark 3.3.7. *For the part (iii), (A4) is implied by MT Theorem 14.3.7 with* $\beta = \alpha$.

The theorem consists of Theorem 3.3.3, Theorem 3.3.4, Theorem 3.3.5, and Lemma 3.3.1. Theorem 3.3.5 shows that $\{V(X_n) : n \geq 0\}$ in the case (iii) is bounded in probability. The case (iii) is a special case of S.P.E. with $q = 1$ so that the uniform quantitative bound of $\left\|P^n_{\gamma}(x,\cdot) - \pi(\cdot)\right\|_{TV}$ for $\gamma \in \mathcal{Y}$ exists.

3.3.3 Proof of Theorem 3.3.2

Proof of Lemma 3.3.1: We use the notation in Theorem 3.3.1.
From S.P.E., for $\gamma \in \mathcal{Y}$, let $\rho_{x,x'}(dy) = \delta\nu_{\gamma}(dy)$ (so $\rho_{x,x'}(\mathcal{X}) = \delta$) and $\triangle := C \times C$. So, $\epsilon^+ = \epsilon^- = \delta$.

Note that the matrix $I - A^{(\beta)}_m$ is a lower triangle matrix. Denote $(I - A^{(\beta)}_m)^{-1} := (b^{(\beta)}_{ij})_{i,j=0,\cdots,q}$.

By the definition of $B^{(\beta)}_l(x,n)$,

$$\begin{aligned} B^{(\beta)}_l(x,n) =& \frac{\epsilon^+ \sum_{k=0}^{l} b^{(\beta)}_{lk} \int \pi(dy) W^{\beta}_k(x,y)}{S(l,n+1-m)^{\beta} + \sum_{j\geq n+1-m}(1-\epsilon^+)^{j-(n-m)}(S(l,j+1)^{\beta} - S(l,j)^{\beta})} \\ \leq& \frac{\epsilon^+}{S(l,n+1-m)^{\beta}} \sum_{k=0}^{l} b^{(\beta)}_{lk} \int \pi(dy) W^{\beta}_k(x,y). \end{aligned}$$

By some algebra, for $k = 1, \cdots, q$,

$$\int \pi(dy) W^{\beta}_k(x,y) \leq 1 + \left(m(V_0)\prod_{i=0}^{k-1} a_i\right)^{-\beta} \left[V^{\beta}_k(x) + \pi(V^{\beta}_k)\right] \tag{3.11}$$

because $\beta \in (0,1]$. In addition, $m(V_0) \geq c_0$ so the coefficient of the second term on the right hand side is finite.

By induction, we obtain that $b_{10}^{(\beta)} = \frac{A_m^{(\beta)}(1)}{(1-A_m^{(\beta)}(0))^2}$, and $b_{11}^{(\beta)} = \frac{1}{1-A_m^{(\beta)}(0)}$. It is easy to check that $0 < b_{11}^{(\beta)} \leq \frac{1}{\delta}$.

By some algebra,

$$\begin{aligned} A_m^{(\beta)}(1) \leq & m^\beta + \sup_{(x,x')\in C\times C} \int R_{x,x'}(x,dy) R_{x,x'}(x',dy') W_1^\beta(y,y') \\ \leq & m^\beta + \sup_{(x,x')\in C\times C} \left[1 + (a_0 m(V_0))^{-\beta} (P_\gamma^m V_1^\beta(x) + P_\gamma^m V_1^\beta(x'))\right] \\ \leq & m^\beta + 1 + 2(a_0 m(V_0))^{-\beta} (\sup_{x\in C} V_1(x) + m b_0) \end{aligned}$$

because $P_\gamma^m V_1^\beta(x) \leq P_\gamma^m V_1(x) \leq V_1(x) + m b_0$. Therefore, $b_{10}^{(\beta)}$ is bounded from the above by some value independent of γ.

Thus,

$$\begin{aligned} B_1^{(\beta)}(x,n) \leq & \frac{\delta}{S(1,n+1-m)^\beta} \left(b_{10}^{(\beta)} \int \pi(dy) W_0^\beta(x,y) + b_{11}^{(\beta)} \int \pi(dy) W_1^\beta(x,y) \right) \\ \leq & \frac{\delta}{(n+1-m)^\beta} \left(b_{10}^{(\beta)} \pi(C) + b_{11}^{(\beta)} \left[1 + (a_0 m(V_0))^{-\beta} (V_1^\beta(x) + \pi(V_1^\beta))\right] \right). \end{aligned}$$

Therefore, the boundedness of the process $\{V_1(X_k) : k \geq 0\}$ implies that the random sequence $B_1^{(\beta)}(X_n, n)$ converges to zero in probability. Containment holds. □

Let $\{Z_j : j \geq 0\}$ be an adaptive sequence of positive random variables. For each j, Z_j will denote a fixed positive Borel measurable function of X_j. τ_n will denote a stopping time starting from the time n of the process $\{X_i : i \geq 0\}$ i.e. $[\tau_n = i] \subset \sigma(X_k : k = 1, \cdots, n+i)$ and $\mathbf{P}(\tau_n < \infty) = 1$.

Lemma 3.3.2 (Dynkin's Formula for adaptive MCMC)**.** *For $m > 0$, and $n > 0$,*

$$\mathbf{E}[Z_{\tilde{\tau}_{m,n}} \mid X_m, \Gamma_m] = Z_m + \mathbf{E}[\sum_{i=1}^{\tilde{\tau}_{m,n}} (\mathbf{E}[Z_{m+i} \mid \mathcal{F}_{m+i-1}] - Z_{m+i-1}) \mid X_m, \Gamma_m]$$

where $\tilde{\tau}_{m,n} := \min(n, \tau_m, \inf(k \geq 0 : Z_{m+k} \geq n))$.

Proof:

$$Z_{\tilde{\tau}_{m,n}} = Z_m + \sum_{i=1}^{\tilde{\tau}_{m,n}} (Z_{m+i} - Z_{m+i-1}) = Z_m + \sum_{i=1}^{n} \mathbb{I}(\tilde{\tau}_{m,n} \geq i)(Z_{m+i} - Z_{m+i-1})$$

Since $\tilde{\tau}_{m,n} \geq i$ is measurable to $\mathcal{F}_{m+i-1}$,

$$\begin{aligned}\mathbf{E}[Z_{\tilde{\tau}_{m,n}} \mid X_m, \Gamma_m] =& Z_m + \mathbf{E}[\sum_{i=1}^{n} \mathbf{E}[Z_{m+i} - Z_{m+i-1} \mid \mathcal{F}_{m+i-1}]\mathbb{I}(\tilde{\tau}_{m,n} \geq i) \mid X_m, \Gamma_m] \\ =& Z_m + \mathbf{E}[\sum_{i=1}^{\tilde{\tau}_{m,n}} (\mathbf{E}[Z_{m+i} \mid \mathcal{F}_{m+i-1}] - Z_{m+i-1}) \mid X_m, \Gamma_m].\end{aligned}$$

□

Lemma 3.3.3 (Comparison Lemma for adaptive MCMC)**.** *Suppose that there exist two sequences of positive functions* $\{s_j, f_j : j \geq 0\}$ *on* $\mathcal{X}$ *such that*

$$\mathbf{E}[Z_{j+1} \mid \mathcal{F}_j] \leq Z_j - f_j(X_j) + s_j(X_j). \tag{3.12}$$

Then for a stopping time τ_n *starting from the time* n *of the adaptive MCMC* $\{X_i : i \geq 0\}$*,*

$$\mathbf{E}[\sum_{j=0}^{\tau_n - 1} f_{n+j}(X_{n+j}) \mid X_n, \Gamma_n] \leq Z_n(X_n) + \mathbf{E}[\sum_{j=0}^{\tau_n - 1} s_{n+j}(X_{n+j}) \mid X_n, \Gamma_n].$$

Proof: From Lemma 3.3.2 and Equation (3.12), the result can be obtained. □

The following proposition shows the relations between the moments of the hitting time and the test function V-modulated moments for adaptive MCMC algorithms with S.P.E., which is derived from the result for Markov chain in (Jarner and Roberts, 2002, Theorem 3.2). Define the *first return time* and the ith *return time* to the set C from the time n respectively:

$$\tau_{n,C} := \tau_{n,C}(1) := \min\{k \geq 1 : X_{n+k} \in C\} \tag{3.13}$$

and

$$\tau_{n,C}(i) := \min\{k > \tau_{n,C}(i-1) : X_{n+k} \in C\} \text{ for } n \geq 0 \text{ and } i > 1. \tag{3.14}$$

Proposition 3.3.1. *Consider an adaptive MCMC* $\{X_i : i \geq 0\}$ *with the adaptive parameter* $\{\Gamma_i : i \geq 0\}$. *If the family* $\{P_\gamma : \gamma \in \mathcal{Y}\}$ *is* S.P.E., *then there exist some constants* $\{d_i : i = 0, \cdots, q-1\}$ *such that at the time* n, *for* $k = 1, \cdots, q$,

$$\begin{aligned}\frac{c_{q-k}\mathbf{E}[\tau_{n,C}^k \mid X_n, \Gamma_n]}{k} \leq & \mathbf{E}[\sum_{i=0}^{\tau_{n,C}-1} (i+1)^{k-1} V_{q-k}(X_{n+i}) \mid X_n, \Gamma_n] \\ \leq & d_{q-k}(V_q(X_n) + \sum_{i=1}^{k} b_{q-i}\mathbb{I}_C(X_n))\end{aligned}$$

where the test functions $\{V_i(\cdot) : i = 0, \cdots, q\}$, *the set* C, $\{c_i : i = 0, \cdots, q-1\}$, *and* $\{b_i : i = 0, \cdots, q-1\}$ *are defined in the* S.P.E..

Proof:

$$\sum_{i=0}^{\tau_{n,C}-1} (i+1)^{k-1} \geq \int_0^{\tau_{n,C}} x^{k-1} dx = k^{-1}\tau_{n,C}^k.$$

Since $V_{q-k}(x) \geq c_{q-k}$ on $\mathcal{X}$,

$$\mathbf{E}[\sum_{i=0}^{\tau_{n,C}-1} (i+1)^{k-1} V_{q-k}(X_{n+i}) \mid X_n, \Gamma_n] \geq \frac{c_{q-k}}{k}\mathbf{E}[\tau_{n,C}^k \mid X_n, \Gamma_n]. \tag{3.15}$$

So, the first inequality holds.

Consider $k = 1$. By S.P.E. and Lemma 3.3.3,

$$\mathbf{E}[\sum_{i=0}^{\tau_{n,C}-1} V_{q-1}(X_{n+i}) \mid X_n, \Gamma_n] \leq V_q(X_n) + b_{q-1}\mathbb{I}_C(X_n). \tag{3.16}$$

So, the case $k = 1$ of the second inequality of the result holds.

For $i \geq 0$, by S.P.E.,

$$\begin{aligned}
&\mathbf{E}[(i+1)^{k-1}V_{q-k+1}(X_{n+i+1}) \mid X_{n+i}, \Gamma_{n+i}] - i^{k-1}V_{q-k+1}(X_{n+i}) \\
\leq &(i+1)^{k-1}\left(V_{q-k+1}(X_{n+i}) - V_{q-k}(X_{n+i}) + b_{q-k}\mathbb{I}_C(X_{n+i})\right) - i^{k-1}V_{q-k+1}(X_{n+i}) \\
\leq &-(i+1)^{k-1}V_{q-k}(X_{n+i}) + \tilde{d}\left(i^{k-2}V_{q-k+1}(X_{n+i}) + (i+1)^{k-1}b_{q-k}\mathbb{I}_C(X_{n+i})\right)
\end{aligned}$$

for some positive $\tilde{d}$ independent of i.

By Lemma 3.3.3,

$$\begin{aligned}
&\mathbf{E}[\sum_{i=0}^{\tau_{n,C}-1}(i+1)^{k-1}V_{q-k}(X_{n+i}) \mid X_n, \Gamma_n] \leq \\
&\qquad \tilde{d}\mathbf{E}[\sum_{i=0}^{\tau_{n,C}-1} i^{(k-1)-1}V_{q-(k-1)}(X_{n+i}) \mid X_n, \Gamma_n] + b_{q-k}\mathbb{I}_C(X_n).
\end{aligned} \tag{3.17}$$

From the above equation, by induction, the second inequality of the result holds. □

Theorem 3.3.3. *Suppose that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.P.E. for $q > 2$. Then, Containment holds.*

Proof: For $k = 1, \ldots, q$, take large enough $M > 0$ such that $C \subset \{x : V_{q-k}(x) \leq M\}$,

$$\begin{aligned}
\mathbf{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M\right) = &\sum_{i=0}^{n}\mathbf{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{i,C} > n-i, X_i \in C\right) + \\
&\mathbf{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{0,C} > n, X_0 \notin C\right).
\end{aligned}$$

By Proposition 3.3.1, for $i = 0, \cdots, n$,

$$\begin{aligned}
&\mathbf{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{i,C} > n-i \mid X_i \in C\right)\\
\leq &\mathbf{P}_{(x_0,\gamma_0)}\left(\sum_{j=0}^{\tau_{i,C}-1}(j+1)^{k-1}V_{q-k}(X_{i+j}) > (n-i)^{k-1}M + \right.\\
&\qquad \left. c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1}, \tau_{i,C} > n-i \mid X_i \in C\right)\\
\leq &\mathbf{P}_{(x_0,\gamma_0)}\left(\sum_{j=0}^{\tau_{i,C}-1}(j+1)^{k-1}V_{q-k}(X_{i+j}) > (n-i)^{k-1}M + \right.\\
&\qquad \left. c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1} \mid X_i \in C\right)\\
\leq &\frac{\sup_{x\in C}\mathbf{E}_{(x_0,\gamma_0)}\left[\mathbf{E}_{(x_0,\gamma_0)}\left[\sum_{j=0}^{\tau_{i,C}-1}(j+1)^{k-1}V_{q-k}(X_{i+j}) \mid X_i, \Gamma_i\right] \mid X_i = x\right]}{(n-i)^{k-1}M + c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1}}\\
\leq &\frac{d_{q-k}\left(\sup_{x\in C}V_q(x) + \sum_{j=1}^{k}b_{q-j}\mathbb{I}_C(x)\right)}{(n-i)^{k-1}M + c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1}},
\end{aligned}$$

and

$$\mathbf{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M, \tau_{0,C} > n \mid X_0 \notin C\right) \leq \frac{d_{q-k}\left(V_q(x_0) + \sum_{j=1}^{k}b_{q-j}\mathbb{I}_C(x_0)\right)}{n^{k-1}M + c_{q-k}\sum_{j=0}^{n-1}(j+1)^{k-1}}.$$

By simple algebra,

$$(n-i)^{k-1}M + c_{q-k}\sum_{j=0}^{n-i-1}(j+1)^{k-1} = \mathrm{O}\left((n-i)^{k-1}\left(M + c_{q-k}(n-i)\right)\right).$$

Therefore,

$$\begin{aligned}
&\mathbf{P}_{(x_0,\gamma_0)}\left(V_{q-k}(X_n) > M\right) \\
&\leq d_{q-k}\left(\sup_{x\in C\cup\{x_0\}} V_q(x) + \sum_{j=1}^{k} b_{q-j}\right) \\
&\quad\left(\sum_{i=0}^{n} \frac{\mathbf{P}_{(x_0,\gamma_0)}(X_i \in C)}{(n-i)^{k-1}\left(M + c_{q-k}(n-i)\right)} + \frac{\delta_{C^c}(x_0)}{n^{k-1}\left(M + c_{q-k}n\right)}\right).
\end{aligned} \tag{3.18}$$

Whenever $q \geq 2$, k can be chosen as 2. While $k \geq 2$, the summation of L.H.S. of Equation (3.18) is finite given M. But if $q = 2$ then just the process $\{V_0(X_n) : n \geq 0\}$ is bounded probability so that $q > 2$ is required for the result. Hence, taking large enough $M > 0$, the probability will be small enough. So, the sequence $\{V_{q-2}(X_n) : n \geq 0\}$ is bounded in probability. By Lemma 3.3.1, Containment holds. □

Remark 3.3.8. *In the proof, only* (A3) *is used.*

Remark 3.3.9. *If $V_0(\cdot)$ is a "nice" function (non-decreasing) of $V_1(\cdot)$, then the sequence $\{V_1(X_n) : n \geq 0\}$ is bounded in probability. In Theorem 3.3.5, we discuss this situation for certain simultaneously single polynomial drift condition.*

Theorem 3.3.4. *Suppose that $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is S.P.E. for $q = 2$. Suppose that there exists a strictly increasing function $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that $V_1(x) \leq f(V_0(x))$ for all $x \in \mathcal{X}$. Then, Containment is implied.*

Proof: From Equation (3.18), we have that $\{V_0(X_n) : n \geq 0\}$ is bounded in probability. Since $V_1(x) \leq f(V_0(x))$,

$$\mathbf{P}_{(x_0,\gamma_0)}\left(V_1(X_n) > f(M)\right) \leq \mathbf{P}_{(x_0,\gamma_0)}\left(f(V_0(X_n)) > f(M)\right) = \mathbf{P}_{(x_0,\gamma_0)}\left(V_0(X_n) > M\right),$$

because $f(\cdot)$ is strictly increasing. By the boundedness of $V_0(X_n)$, for any $\epsilon > 0$, there exists $N > 0$ and some $M > 0$ such that for $n > N$, $\mathbf{P}_{(x_0,\gamma_0)}\left(V_1(X_n) > f(M)\right) \leq \epsilon$. Therefore, $\{V_1(X_n) : n \geq 0\}$ is bounded in probability. By Lemma 3.3.1, Containment is satisfied. □

Consider the single polynomial drift condition, see Jarner and Roberts (2002): $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + b\mathbb{I}_C(x)$ where $0 \leq \alpha < 1$. Because the moments of the hitting time to the set C is (see details in Jarner and Roberts (2002)), for any

$1 \leq \xi \leq 1/(1-\alpha)$,

$$E_x\left[\sum_{i=0}^{\tau_C-1}(i+1)^{\xi-1}V(X_i)\right] < V(x) + b\mathbb{I}_C(x).$$

The polynomial rate function $r(n) = n^{\xi-1}$. If $\alpha = 0$, then $r(n)$ is a constant. Under this situation, it is difficult to utilize the technique in Theorem 3.3.3 to prove $\{V(X_n) : n \geq 0\}$ is bounded in probability. Thus, we assume $\alpha \in (0,1)$.

Proposition 3.3.2. *Consider an adaptive MCMC $\{X_n : n \geq 0\}$ with an adaptive scheme $\{\Gamma_n : n \geq 0\}$. Suppose that (A1) holds, and there exist some positive constants $c > 0$, $> b > 0$, $\alpha \in (0,1)$, and a measurable function $V(x) : \mathcal{X} \to \mathbb{R}+$ with $V(x) \geq 1$ and $\sup_{x\in C} V(x) < \infty$ such that*

$$P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + b\mathbb{I}_C(x) \text{ for } \gamma \in \mathcal{Y}. \tag{3.19}$$

Then for $1 \leq \xi \leq 1/(1-\alpha)$,

$$\mathbf{E}_{(x_0,\gamma_0)}\left[\sum_{i=0}^{\tau_{n,C}-1}(i+1)^{\xi-1}V^{1-\xi(1-\alpha)}(X_{n+i}) \mid X_n, \Gamma_n\right] \leq c_\xi(C)(V(X_n)+1). \tag{3.20}$$

Proof: The proof applies the techniques in Lemma 3.5 and Theorem 3.6 of Jarner and Roberts (2002). □

Theorem 3.3.5. *Suppose that (A2) and the conditions in Proposition 3.3.2 are satisfied, and there exists some constant $b' > b$ such that $cV^\alpha(x)\mathbb{I}_{C^c} > b'$. Then, Containment is implied.*

Proof: Using the same techniques in Theorem 3.3.3, we have that

$$\begin{aligned} & \mathbf{P}_{(x_0,\gamma_0)}\left(V^{1-\xi(1-\alpha)}(X_n) > M\right) \\ \leq\ & c_\xi\left(\sup_{x\in C\cup\{x_0\}} V(x)+1\right)\left(\sum_{i=0}^{n}\frac{P_{(x_0,\gamma_0)}(X_i\in C)}{(n-i)^{\xi-1}(M+n-i)} + \frac{\delta_{C^c}(x_0)}{n^{\xi-1}(M+n)}\right). \end{aligned} \tag{3.21}$$

Therefore, for $\xi \in [1, 1/(1-\alpha))$, the sequence $\left\{V^{1-\xi(1-\alpha)}(X_n) : n \geq 0\right\}$ is bounded in probability. Since $1-\xi(1-\alpha) > 0$, the process $\{V(X_n) : n \geq 0\}$ is bounded in probability. By Lemma 3.3.1, Containment holds. □

Chapter 4

Some Applicable Ergodicity Conditions for Multidimensional Targets

This chapter considers ergodicity properties of certain adaptive Markov chain Monte Carlo (MCMC) algorithms for multidimensional target distributions, in particular adaptive Metropolis and adaptive Metropolis-within-Gibbs algorithms. We derive various sufficient conditions to ensure Containment, and connect the convergence rates of algorithms with the tail properties of the target distributions. We also present a Summable Adaptive Condition which, when satisfied, proves ergodicity more easily.

When designing adaptive algorithms, it is not difficult to ensure that Diminishing Adaptation holds. However, Containment may be more challenging, which raises the questions. How can Containment be verified in specific examples? Roberts and Rosenthal (2007) prove that an adaptive MCMC satisfying Diminishing Adaptation satisfies Containment if the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ is *simultaneously strongly aperiodically geometrically ergodic*. We study a weaker condition: Simultaneous Geometrical Ergodicity which is also sufficient for ergodicity of adaptive MCMC, but this may be difficult to check in practice. In this section, we give some simpler criteria related to proposals to check Containment, more easily.

First we discuss simultaneous geometric ergodicity in Section 4.1. Then we show in Section 4.2 that a stronger version of the Diminishing Adaptation alone implies ergodicity of adaptive algorithm. We then give some results which ensure ergodicity for certain adaptive Metropolis algorithms in Section 4.3 and adaptive Metropolis-within-Gibbs algorithms in Section 4.4.

4.1 Simultaneous Geometric Ergodicity

Following standard results about geometric ergodicity and polynomial ergodicity, Roberts and Rosenthal (2007) also considered certain "simultaneous" ergodicity conditions.

Definition 4.1.1 (simultaneously strongly aperiodically geometrically ergodic)**.** *Consider the family* $\{P_\gamma : \gamma \in \mathcal{Y}\}$*. Suppose that there is* $C \in \mathcal{F}$*, a measurable function* $V : \mathcal{X} \to [1, \infty), \delta > 0, \lambda < 1$*, and* $b < \infty$*, such that* $\sup_C V = v < \infty$*, and*

(i) $\exists$ *a probability measure* $\nu(\cdot)$ *on* C *with* $P_\gamma(x, \cdot) \geq \delta\nu_\gamma(\cdot)$ *for* $x \in C$*; and*

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

We say that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ *is Simultaneously Strongly Aperiodically Geometrically Ergodic (S.S.A.G.E.).*

Theorem 4.1.1 (Roberts and Rosenthal (2007))**.** *Consider an adaptive MCMC algorithm with Diminishing Adaptation. Suppose that the family* $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ *is simultaneously strongly aperiodically geometrically ergodic. Then the adaptive algorithm is ergodic.*

Before we study Simultaneous Geometric Ergodicity for adaptive MCMC algorithms, let us review Rosenthal (1995, Theorem 5).

Proposition 4.1.1. *Suppose a time homogeneous Markov chain* $P(x, dy)$ *on the state space* $\mathcal{X}$*. Let* $\{X_n : n \geq 0\}$ *and* $\{Y_n : n \geq 0\}$ *be two realizations of* $P(x, dy)$*. There are a set* $C \subset \mathcal{X}$*,* $\delta > 0$*, some integer* $m > 0$*, and a probability measure* ν_m *on* $\mathcal{X}$ *such that*

$$P^m(x, \cdot) \geq \delta\nu_m(\cdot) \text{ for } x \in C.$$

Suppose further that there exist $0 < \lambda < 1$*,* $b > 0$*, and a function* $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$

such that

$$\mathbf{E}\left[h(X_1, Y_1) \mid X_0 = x, Y_0 = y\right] \leq \lambda h(x, y) + b\mathbb{I}_{C\times C}((x, y)).$$

Let $A := \sup_{(x,y)\in C\times C} \mathbf{E}[h(X_m, Y_m) \mid X_0 = x, Y_0 = y]$, μ *be the initial distribution, and* π *be the stationary distribution. Then for any* $j > 0$,

$$\|\mathcal{L}(X_n) - \pi\|_{\mathrm{TV}} \leq (1-\delta)^{[j/m]} + \lambda^{n-jm+1} A^{j-1} \mathbf{E}_{\mu\times\pi}[h(X_0, Y_0)].$$

To make use of Proposition 4.1.1, we consider the *Simultaneously Geometrically Ergodic* condition (S.G.E.) also studied by Roberts et al. (1998):

Definition 4.1.2 (S.G.E.)**.** *Suppose that there is* $C \in \mathcal{F}$, *some integer* $m \geq 1$, *a function* $V : \mathcal{X} \to [1, \infty), \delta > 0, \lambda < 1$, *and* $b < \infty$, *such that* $\sup_{x\in C} V(x) < \infty$, $\pi(V) < \infty$, *and*

(i) C *is an uniform* ν_m*-small set, i.e., for each* γ, $\exists$ *a probability measure* $\nu_\gamma(\cdot)$ *on* C *with* $P_\gamma^m(x, \cdot) \geq \delta\nu_\gamma(\cdot)$ *for* $x \in C$*;*

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

We say that the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ *is Simultaneously Geometrically Ergodic.*

Note that the difference between S.G.E. and S.S.A.G.E. is that the uniform minorization set C for all P_γ is assumed in S.S.A.G.E., however the uniform small set C is assumed in S.G.E.. Obviously S.G.E. is a special case of S.P.E.. Here we use the quantitative bound in Proposition 4.1.1 to show the following theorem.

Theorem 4.1.2. *S.G.E. implies Containment.*

Proof: Let $\{X_n^{(\gamma)} : n \geq 0\}$ and $\{X_n^{(\gamma)} : n \geq 0\}$ be two realizations of P_γ for $\gamma \in \mathcal{Y}$. Define $h(x, y) := (V(x) + V(y))/2$. From (ii) of S.G.E., $\mathbf{E}[h(X_1^{(\gamma)}, Y_1^{(\gamma)}) \mid X_0^{(\gamma)} = x, Y_0^{(\gamma)} = y] \leq \lambda h(x, y) + b\mathbb{I}_{C\times C}((x, y))$. It is not difficult to get $P_\gamma^m V(x) \leq \lambda^m V(x) + bm$ so $A := \sup_{(x,y)\in C\times C} \mathbf{E}[h(X_m^{(\gamma)}, Y_m^{(\gamma)}) \mid X_0^{(\gamma)} = x, Y_0^{(\gamma)} = y] \leq \lambda^m \sup_C V + bm =: B$.

Consider $\mathcal{L}(X_0^{(\gamma)}) = \delta_x$ and $j := \sqrt{n}$. By Proposition 4.1.1,

$$\left\|P_\gamma^n(x, \cdot) - \pi(\cdot)\right\|_{\mathrm{TV}} \leq (1-\delta)^{[\sqrt{n}/m]} + \lambda^{n-\sqrt{n}m+1} B^{\sqrt{n}-1}(V(x) + \pi(V))/2. \qquad (4.1)$$

Note that the quantitative bound is dependent of x, n, δ, m, C, V and π, and independent of γ. Given $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, the uniform quantitative bound of

$\left\|P_\gamma^n(x,\cdot)-\pi(\cdot)\right\|_{\text{TV}}$ tends to zero as n goes to infinity.

Let $\{X_n : n \geq 0\}$ be the adaptive MCMC satisfying S.G.E.. From (ii) of S.G.E., $\sup_n \mathbf{E}[V(X_n) \mid X_0 = x, \Gamma_0 = \gamma_0] < \infty$ so the process $\{V(X_n) : n \geq 0\}$ is bounded in probability. Therefore, for any $\epsilon > 0$, $\{M_\epsilon(X_n, \Gamma_n) : n \geq 0\}$ is bounded in probability given any $X_0 = x_0$ and $\Gamma_0 = \gamma_0$. □

Corollary 4.1.1. *Consider the family $\{P_\gamma : \gamma \in \mathcal{Y}\}$ of Markov chains on $\mathcal{X}$. Suppose that for any compact set $C \in \mathcal{F}$, there exist some integer $m > 0$, $\delta > 0$ and a probability measure $\nu_\gamma(\cdot)$ on C for $\gamma \in \mathcal{Y}$ such that $P_\gamma^m(x,\cdot) \geq \delta\nu_\gamma(\cdot)$ for $x \in C$. Suppose that there is a function $V : \mathcal{X} \to (1,\infty)$ such that $\sup_{x\in C} V(x) < \infty$, $\pi(V) < \infty$, and*

$$\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1. \tag{4.2}$$

Then for any adaptive strategy using only $\{P_\gamma : \gamma \in \mathcal{Y}\}$, Containment holds.

Proof: From Equation (4.2), letting $\lambda = \limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1$, there exists some positive constant K such that $\sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < \frac{\lambda+1}{2}$ for $|x| > K$. By $V > 1$, $P_\gamma V(x) < \frac{\lambda+1}{2} V(x)$ for $|x| > K$. $P_\gamma V(x) \leq \frac{\lambda+1}{2} V(x) + b\mathbb{I}_{\{z\in\mathcal{X}:|z|\leq K\}}(x)$ where $b = \sup_{x\in\{z\in\mathcal{X}:|z|\leq K\}} V(x)$. □

4.2 Summable Adaptive Condition

In Chapter 2, we give two examples to explain that Diminishing Adaptation alone is not sufficient for ergodicity. Yang (2008a) assumes a summable adaptive condition and Simultaneous Uniform Ergodicity[1] that imply ergodicity. Here we present a summable adaptive condition (Equation (4.3)) to show ergodicity of adaptive MCMC without assuming simultaneous uniform ergodicity. We also will present a modification of Example 2.2.1 which is ergodic.

Proposition 4.2.1. *Consider an adaptive MCMC $\{X_n : n \geq 0\}$ on the state space $\mathcal{X}$ with the kernel index space $\mathcal{Y}$. Under the following conditions:*
(i) *$\mathcal{Y}$ is finite. For every $\gamma \in \mathcal{Y}$, P_γ is ergodic with the stationary distribution π;*

[1]Simultaneous Uniform Ergodicity: For all $\epsilon > 0$, there is a $N > 0$ such that $\left\|P_\gamma^N(x,\cdot)-\pi(\cdot)\right\|_{\text{TV}} \leq \epsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.

(ii) *At each time* n, Γ_n *is a deterministic measurable function of* $X_0, \cdots, X_n$;
(iii) *For every initial state* $x_0 \in \mathcal{X}$ *and initial kernel index* $\gamma_0 \in \mathcal{Y}$,

$$\sum_{n=1}^{\infty} \mathbf{P}(\Gamma_n \neq \Gamma_{n-1} \mid X_0 = x_0, \Gamma_0 = \gamma_0) < \infty. \tag{4.3}$$

Then the adaptive MCMC $\{X_n : n \geq 0\}$ *is ergodic with the stationary distribution* π.

Proof: Fix $x_0 \in \mathcal{X}$, $\gamma_0 \in \mathcal{Y}$. By the condition (iii) and the Borel-Cantelli Lemma, $\forall \epsilon > 0$, $\exists N_0(x_0, \gamma_0, \epsilon) > 0$ such that $\forall n > N_0$,

$$\mathbf{P}\left(\Gamma_n = \Gamma_{n+1} = \cdots \mid X_0 = x_0, \Gamma_0 = \gamma_0\right) > 1 - \epsilon/2. \tag{4.4}$$

Construct a new chain $\{\tilde{X}_n : n \geq 0\}$ which satisfies that for $n \leq N_0$, $\tilde{X}_n = X_n$, and for $n \geq N_0$, $\tilde{X}_n \sim P^{n-N_0}_{\Gamma_{N_0}}(\tilde{X}_{N_0}, \cdot)$. So, for any $n > N_0$ and any set $A \in \mathcal{F}$, by the condition (ii),

$$\begin{aligned}
&\mathbf{P}(X_n \in A, \Gamma_{N_0} = \Gamma_{N_0+1} = \cdots = \Gamma_{n-1} \mid X_0 = x_0, \Gamma_0 = \gamma_0) \\
&= \int_{\mathcal{X}^{N_0} \cap [\gamma_{N_0} = \cdots = \gamma_{n-1}]} P_{\gamma_0}(x_0, dx_1) \cdots P_{\gamma_{N_0-1}}(x_{N_0-1}, dx_{N_0}) P^{n-N_0}_{\gamma_{N_0}}(x_{N_0}, A)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{P}(\tilde{X}_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0) \\
&= \int_{\mathcal{X}^{N_0}} P_{\gamma_0}(x_0, dx_1) \cdots P_{\gamma_{N_0-1}}(x_{N_0-1}, dx_{N_0}) P^{n-N_0}_{\gamma_{N_0}}(x_{N_0}, A)
\end{aligned}$$

So,

$$\begin{aligned}
|\mathbf{P}\left(X_n \in A, \Gamma_{N_0} = \cdots = \Gamma_{n-1} \mid X_0 = x_0, \Gamma_0 = \gamma_0\right) - \\
\mathbf{P}(\tilde{X}_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0)\,| \leq \epsilon/2.
\end{aligned}$$

Since the condition (i) holds, suppose that for some $K > 0$, $\mathcal{Y} = \{y_1, \cdots, y_K\}$. Denote $\mu_i(\cdot) = \mathbf{P}(\tilde{X}_{N_0} \in \cdot \mid X_0 = x_0, \Gamma_0 = \gamma_0, \Gamma_{N_0} = y_i)$ for $i = 1, \cdots, K$. Because of

the condition (ii), for $n > N_0$,

$$\begin{aligned}
&\mathbf{P}(\tilde{X}_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0) \\
=&\sum_{i=1}^{K} \mathbf{P}(\tilde{X}_n \in A, \Gamma_{N_0} = y_i \mid X_0 = x_0, \Gamma_0 = \gamma_0) \\
=&\sum_{i=1}^{K} \int_{\mathcal{X}^{N_0} \cap [\gamma_{N_0} = y_i]} P_{\gamma_0}(x_0, dx_1) \cdots P_{\gamma_{N_0-1}}(x_{N_0-1}, dx_{N_0}) P_{y_i}^{n-N_0}(x_{N_0}, A) \\
=&\sum_{i=1}^{K} \mathbf{P}(\Gamma_{N_0} = y_i \mid X_0 = x_0, \Gamma_0 = \gamma_0) \mu_i P_{y_i}^{n-N_0}(A).
\end{aligned}$$

By the condition (i), there exists $N_1(x_0, \gamma_0, \epsilon, N_0) > 0$ such that for $n > N_1$,

$$\sup_{i \in \{1, \cdots, K\}} \left\| \mu_i P_{y_i}^n(\cdot) - \pi(\cdot) \right\|_{TV} < \epsilon/2.$$

So, for any $n > N_0 + N_1$, any $A \in \mathcal{F}$,

$$\begin{aligned}
&|\mathbf{P}(X_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0) - \pi(A)| \\
\leq &\left| \mathbf{P}(X_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0) - \mathbf{P}(\tilde{X}_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0) \right| + \\
&\left| \mathbf{P}(\tilde{X}_n \in A \mid X_0 = x_0, \Gamma_0 = \gamma_0) - \pi(A) \right| \\
\leq &(\epsilon/2 + \epsilon/2) + \epsilon/2 = 3\epsilon/2.
\end{aligned}$$

Therefore, the adaptive MCMC $\{X_n : n \geq 0\}$ is ergodic with the target distribution π. □

Example 4.2.1. *Consider again the Metropolis-Hastings algorithm of Example 2.2.1, with $\mathcal{X} = (0, \infty)$ and $\mathcal{Y} = \{-1, 1\}$, and $\pi(x) \propto \frac{\mathbb{I}(x \geq 0)}{1+x^2}$, and is $Y_n^{\Gamma_{n-1}} = X_{n-1}^{\Gamma_{n-1}} + Z_n$ where $\{Z_n\}$ are i.i.d. standard normal. Assume now that the adaptive parameters $\{\Gamma_n\}$ are updated according to $\Gamma_n = -\Gamma_{n-1}\mathbb{I}(X_n^{\Gamma_{n-1}} < \frac{1}{n^{1+r}}) + \Gamma_{n-1}\mathbb{I}(X_n^{\Gamma_{n-1}} \geq \frac{1}{n^{1+r}})$ for some $r \geq 0$, so the case $r = 0$ corresponds to Example 2.2.1 (which was shown to be non-ergodic), while the case $r > 0$ is new.*

Proposition 4.2.2. *If $r > 0$, then the adaptive algorithm of Example 4.2.1 is ergodic, i.e. X_n converges to π.*

Proof: From the calculation in Example 2.2.1, we have that

$$P(\Gamma_n \neq \Gamma_{n-1} \mid X_{n-1} = x, \Gamma_{n-1} = \gamma) = \int_{-x^\gamma}^{\frac{1}{n^{1+r}} - x^\gamma} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) dz \leq \mathrm{O}(\frac{1}{n^{1+r}}).$$

Therefore, $\sum_{n=1}^{\infty} P(\Gamma_n \neq \Gamma_{n-1}) < \infty$. Hence, from Proposition 4.2.1, the adaptive algorithm is ergodic to π. □

4.3 Adaptive Metropolis Algorithms

The target density $\pi(\cdot)$ is defined on the state space $\mathcal{X} \subset \mathbb{R}^d$. In what follows, we shall write $\langle \cdot, \cdot \rangle$ for the usual scalar product on $\mathbb{R}^d$, $|\cdot|$ for the Euclidean and the operator norm, $n(z) := z/|z|$, ∇ for the usual differential (gradient) operator, $m(x) := \nabla\pi(x)/|\nabla\pi(x)|$, $\mathrm{B}^d(x, r) := \{y \in \mathbb{R}^d : |y - x| < r\}$ for the hyperball in $\mathbb{R}^d$ with the center x and the radius r and its closure $\bar{\mathrm{B}}^d(x, r)$, and $\mathrm{Vol}(A)$ for the volume of the set $A \subset \mathbb{R}^d$.

Say an adaptive MCMC is an *Adaptive Metropolis-Hastings algorithm* if each kernel P_γ is from a Metropolis-Hastings algorithm

$$P_\gamma(x, dy) = \alpha_\gamma(x, y) Q_\gamma(x, dy) + \left[1 - \int_{\mathcal{X}} \alpha_\gamma(x, z) Q_\gamma(x, dz)\right] \delta_x(dy) \tag{4.5}$$

where $Q_\gamma(x, dy)$ is the proposal distribution, $\alpha_\gamma(x, y) := \left(\frac{\pi(y) q_\gamma(y,x)}{\pi(x) q_\gamma(x,y)} \wedge 1\right) \mathbb{I}(y \in \mathcal{X})$, and μ_d is Lebesgue measure. Say an adaptive Metropolis-Hastings algorithm is a *random-walk-based Adaptive Metropolis algorithm* if each $q_\gamma(x, y)$ is symmetric for all $\gamma \in \mathcal{Y}$, i.e. $q_\gamma(x, y) = q_\gamma(x - y) = q_\gamma(y - x)$.

Jarner and Hansen (2000) give conditions which imply geometric ergodicity of symmetric random-walk-based Metropolis algorithm on $\mathbb{R}^d$ for target distribution with lighter-than-exponential tails, (see other related results Mengersen and Tweedie, 1996; Roberts and Tweedie, 1996). Here, we extend their result a little to target distribution with exponential tails.

Definition 4.3.1 (Lighter-than-exponential tail)**.** *The density $\pi(\cdot)$ on $\mathbb{R}^d$ is lighter-*

than-exponentially tailed if it is positive and has continuous first derivatives such that

$$\limsup_{|x|\mapsto\infty} \langle n(x), \nabla \log \pi(x)\rangle = -\infty. \tag{4.6}$$

Remark 4.3.1. 1. *The definition implies that for any $r > 0$, there exists $R > 0$ such that*

$$\frac{\pi(x + \alpha n(x)) - \pi(x)}{\pi(x)} \leq -\alpha r, \text{ for } |x| \geq R, \alpha > 0.$$

It means that $\pi(x)$ is exponentially decaying along any ray, but with the rate r tending to infinity as x goes to infinity.

2. *While a target distribution has finite modes and $|x|$ is sufficiently large, the normed gradient $m(x)$ will point towards the origin. The direction $n(x)$ points away from the origin. For Definition 4.3.1, $\langle n(x), \nabla \log \pi(x)\rangle = \frac{|\nabla \pi(x)|}{\pi(x)} \langle n(x), m(x)\rangle$. Even $\limsup_{|x|\mapsto\infty} \langle n(x), m(x)\rangle < 0$, Equation (4.6) might not be true. E.g. $\pi(x) \propto \frac{1}{1+x^2}$, $x \in \mathbb{R}$. $m(x) = -n(x)$ so that $\langle n(x), m(x)\rangle = -1$. $\langle n(x), \nabla \log \pi(x)\rangle = -\frac{2|x|}{1+x^2}$ so $\lim_{|x|\mapsto\infty} \langle n(x), \nabla \log \pi(x)\rangle = 0$.*

Definition 4.3.2 (Exponential tail)**.** *The density function $\pi(\cdot)$ on $\mathbb{R}^d$ is exponentially tailed if it is a positive, continuously differentiable function on $\mathbb{R}^d$, and*

$$\eta_2 := -\limsup_{|x|\to\infty} \langle n(x), \nabla \log \pi(x)\rangle > 0. \tag{4.7}$$

Remark 4.3.2. *There exists $\beta > 0$ such that for x sufficiently large,*

$$\langle n(x), \nabla \log \pi(x)\rangle = \langle n(x), m(x)\rangle \, |\nabla \log \pi(x)| \leq -\beta.$$

Further, if $0 < -\langle n(x), m(x)\rangle \leq 1$, then $|\nabla \log \pi(x)| \geq \beta$.

Define the *symmetric proposal density family* $\mathfrak{C} := \{q : q(x, y) = q(x - y) = q(y - x), x, y \in \mathbb{R}^d\}$. Our ergodic results for adaptive Metropolis algorithm are based on the following assumptions.

Assumption 4.3.1. *The target distribution is absolutely continuous w.r.t. Lebesgue measure μ_d with a density π bounded away from zero and infinity on compact sets, and $\sup_{x\in\mathcal{X}} \pi(x) < \infty$.*

Assumption 4.3.2 (Strongly decreasing)**.** *The target density π has continuous first derivatives and satisfies*

$$\eta_1 := -\limsup_{|x|\mapsto\infty} \langle n(x), m(x)\rangle > 0. \tag{4.8}$$

Assumption 4.3.3 (Uniform Local Positivity)**.** *Assume that $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$. There exists $\zeta > 0$ such that*

$$\iota := \inf_{\gamma\in\mathcal{Y}} \inf_{|z|\le\zeta} q_\gamma(z) > 0. \tag{4.9}$$

Given $0 < p < q < \infty$, for $u \in S^{d-1}$ (S^{d-1} is the unit hypersphere in $\mathbb{R}^d$.) and $\theta > 0$, define

$$C_{p,q}(u,\theta) := \left\{z = a\xi \mid p \le a \le q,\, \xi \in S^{d-1},\, |\xi - u| < \theta/3\right\}. \tag{4.10}$$

Assumption 4.3.4. *Suppose the target density π is exponentially tailed. Under Assumptions 4.3.2, assume that there are $\epsilon \in (0, \eta_1)$, $\beta \in (0, \eta_2)$, δ, and Δ with $0 < \frac{3}{\beta\epsilon} \le \delta < \Delta \le \infty$ such that*

$$\inf_{(u,\gamma)\in S^{d-1}\times\mathcal{Y}} \int_{C_{\delta,\Delta}(u,\epsilon)} |z|\, q_\gamma(z)\mu_d(dz) > \frac{3(e+1)}{\beta\epsilon(e-1)}. \tag{4.11}$$

Remark 4.3.3. *Under Assumption 4.3.3, let $\tilde{P}(x, dy)$ be the transition kernel of Metropolis-Hastings algorithm with the proposal distribution $\tilde{Q}(x,\cdot) \sim \mathrm{Unif}(\bar{\mathrm{B}}^d(x, \zeta/2))$. For any $\gamma \in \mathcal{Y}$, $P_\gamma(x, dy) \ge \iota \mathrm{Vol}(\bar{\mathrm{B}}^d(0, \zeta/2))\tilde{P}(x, dy)$. By Assumptions 4.3.1 and Roberts and Tweedie (1996, Theorem 2.2), any compact set is a small set for $\tilde{P}$ so that any compact set is an uniform small set for all P_γ.*

Remark 4.3.4. 1. *Assumption 4.3.4 means that the proposal family has uniform lower bound of the first moment on some local cone around the origin. It shows that the tails of all proposal distributions can not be too light, and the quantity of the lower bound is given and dependent on the decaying rate η_1 of and strongly decreasing rate η_2 of the target distribution.*

2. *If every proposal distribution in $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$ is a mixture distribution with*

one fixed part, then Assumption 4.3.4 is relatively easy to check, because the integral in Equation (4.11) can be estimated by the fixed part distribution. Especially for the lighter-than-exponentially tailed target, Assumption 4.3.4 can be reduced. We will give a sufficient condition for Assumption 4.3.4, see Lemma 4.3.1.

Now, we consider a particular class of target densities with tails which are heavier than exponential tails. It was previously shown by Fort and Moulines (2000a) that the Metropolis algorithm converges at any polynomial rate when the proposal distribution is compact supported and the log density decreases hyperbolically at infinity, $\log \pi(x) \sim -|x|^s$, for $0 < s < 1$, as $|x| \to \infty$.

Definition 4.3.3 (Hyperbolic tail)**.** *The density function $\pi(\cdot)$ is twice continuously differentiable, and there exist $0 < m < 1$ and some finite positive constants d_i, D_i, $i = 1, 2$ such that for large enough $|x|$,*

$0 < d_0 |x|^m \leq -\log \pi(x) \leq D_0 |x|^m$;

$0 < d_1 |x|^{m-1} \leq |\nabla \log \pi(x)| \leq D_1 |x|^{m-1}$;

$0 < d_2 |x|^{m-2} \leq |\nabla^2 \log \pi(x)| \leq D_2 |x|^{m-2}$.

Assumption 4.3.5 (Proposal's Uniform Compact Support)**.** *Under Assumption 4.3.3, there exists some $M > \zeta$ such that all $q_\gamma(\cdot)$ with $\gamma \in \mathcal{Y}$ are uniformly supported on $\bar{\mathrm{B}}^d(0, M)$.*

Theorem 4.3.1. *Adaptive Metropolis algorithm with Diminishing Adaptation is ergodic, under either condition of the following:*

(i). *Target density π is lighter-than-exponentially tailed, and Assumptions 4.3.1 - 4.3.3;*

(ii). *Target density π is exponentially tailed, and Assumptions 4.3.1 - 4.3.4;*

(iii). *Target density π is hyperbolically tailed, and Assumptions 4.3.1 - 4.3.3 and 4.3.5.*

4.3.1 Applications

Here we discuss two examples. The first one (Example 4.3.1) is from Roberts and Rosenthal (2009) where the proposal density is a fixed distribution of two multivariate normal distributions, one with a fixed small variance, another using the estimate of empirical covariance matrix from historical information as its variance. It is a

slight variant of the famous adaptive Metropolis algorithm of Haario et al. (2001). In the example, the target density has lighter-than-exponential tails. The second (Example 4.3.2) concerns with target densities with truly exponential tails.

Proposition 4.3.1. *If the target density π on $\mathbb{R}^d$ is normal (i.e. $N(\mu, \Sigma)$, Σ is positive definite), then π is strongly decreasing and lighter-than-exponentially tailed.*

Proof: Without loss of generality assume that $\mu = 0$.

Since $\pi(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \frac{1}{|\Sigma|^{1/2}} \exp(-x^\top \Sigma^{-1} x/2)$,

$$\langle n(x), m(x) \rangle = \left\langle \frac{x}{|x|}, \frac{-\Sigma^{-1}x}{|\Sigma^{-1}x|} \right\rangle = -\frac{x^\top \Sigma^{-1} x}{|x|\,|\Sigma^{-1}x|}.$$

Since Σ is a real symmetric and positive definite matrix, suppose that $\Sigma = A^\top D A$ where A is orthogonal, and D is diagonal with positive diagonal elements. Hence,

$$\frac{x^\top \Sigma^{-1} x}{|x|\,|\Sigma^{-1}x|} = \frac{yD^{-1}y}{|y|\,|D^{-1}y|} = \frac{\sum_{i=1}^d y_i^2 d_i^{-1}}{\sqrt{\sum_{i=1}^d y_i^2 \sum_{i=1}^d d_i^{-2} y_i^2}} \geq \frac{\min\left(d_i^{-1}\right)}{\max\left(d_i^{-1}\right)}.$$

where $y = Ax$.

$$\langle n(x), \nabla \log \pi(x) \rangle = \frac{|\nabla \pi(x)|}{\pi(x)} \left\langle \frac{x}{|x|}, \frac{-\Sigma^{-1}x}{|\Sigma^{-1}x|} \right\rangle = -\frac{x\Sigma^{-1}x}{|x|} \underset{|x|\to\infty}{\to} -\infty.$$

So, the result holds. □

Example 4.3.1. *Consider a d-dimensional target distribution $\pi(\cdot)$ satisfying Assumptions 4.3.1 - 4.3.2. We perform a Metropolis algorithm with proposal distribution given at the n^{th} iteration by $Q_n(x, \cdot) = N(x, (0.1)^2 I_d/d)$ for $n \leq 2d$; For $n > 2d$,*

$$Q_n(x,\cdot) = \begin{cases} (1-\theta)N(x, (2.38)^2\Sigma_n/d) + \theta N(x, (0.1)^2 I_d/d), & \Sigma_n \text{ is positive definite,} \\ N(x, (0.1)^2 I_d/d), & \Sigma_n \text{ is not positive definite,} \end{cases} \tag{4.12}$$

for some fixed $\theta \in (0,1)$, I_d is $d \times d$ identity matrix, and the empirical covariance matrix

$$\Sigma_n = \frac{1}{n}\left(\sum_{i=0}^n X_i X_i^\top - (n+1)\overline{X}_n \overline{X}_n^\top\right), \tag{4.13}$$

where $\overline{X}_n = \frac{1}{n+1}\sum_{i=0}^{n} X_i$*, is the current modified empirical estimate of the covariance structure of the target distribution based on the run so far.*

Remark 4.3.5. *The proposal* $N(x, (2.38)^2\Sigma/d)$ *is optimal in a particular large-dimensional context, (see Roberts et al., 1997; Roberts and Rosenthal, 2001). Thus the proposal* $N(x, (2.38)^2\Sigma_n/d)$ *is an effort to approximate this.*

Remark 4.3.6. *Commonly, the iterative form of Equation (4.13) is more useful,*

$$\Sigma_n = \frac{n-1}{n}\Sigma_{n-1} + \frac{1}{n+1}\left(X_n - \bar{X}_{n-1}\right)\left(X_n - \bar{X}_{n-1}\right)^\top. \tag{4.14}$$

Proposition 4.3.2. *Suppose that the target density* π *is exponentially tailed. Under Assumptions 4.3.1-4.3.4,* $\left|\overline{X}_n - \overline{X}_{n-1}\right|$ *and* $\|\Sigma_n - \Sigma_{n-1}\|_M$ *converge to zero in probability where where* $\|\cdot\|_M$ *is matrix norm.*

Proof: Note that in the proof of Theorem 4.3.1, some test function $V(x) = c\pi^{-s}(x)$ for some $s \in (0,1)$ and some $c > 0$ is found such that S.G.E. holds.

By some algebras,

$$\begin{aligned}
&\Sigma_n - \Sigma_{n-1} \\
=&\frac{1}{n+1}X_nX_n^\top - \frac{1}{n-1}\left(\frac{1}{n}\sum_{i=0}^{n-1}X_iX_i^\top\right) + \frac{2n}{n^2-1}\overline{X}_{n-1}\overline{X}_{n-1}^\top - \\
&\frac{1}{n+1}\left(X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\Sigma_n - \Sigma_{n-1}\|_M \\
\leq\ & \tfrac{1}{n+1}\left\|X_nX_n^\top\right\|_M + \tfrac{1}{n-1}\left\|\tfrac{1}{n}\textstyle\sum_{i=0}^{n-1}X_iX_i^\top\right\|_M + \tfrac{2}{n}\left\|\overline{X}_{n-1}\overline{X}_{n-1}^\top\right\|_M + \\
& \tfrac{1}{n+1}\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M.
\end{aligned} \tag{4.15}$$

To prove $\Sigma_n - \Sigma_{n-1}$ converges to zero in probability, it is sufficient to check that $\left\|X_nX_n^\top\right\|_M$, $\left\|\frac{1}{n}\sum_{i=0}^{n-1}X_iX_i^\top\right\|_M$, $\left\|\overline{X}_{n-1}\overline{X}_{n-1}^\top\right\|_M$ and $\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M$ are bounded in probability.

Since $\limsup_{|x|\to\infty}\langle n(x), \nabla\log\pi(x)\rangle < 0$, there exist some $K > 0$ and some $\beta > 0$ such

that

$$\sup_{|x|\geq K} \langle n(x), \nabla \log \pi(x) \rangle \leq -\beta.$$

For $|x| \geq K$, $\frac{\log \pi(y) - \log \pi(x)}{(r-1)|x|} \leq -\beta$ where $r > 1$ and $y = rx$, i.e. $\left(\frac{\pi(y)}{\pi(x)}\right)^{-s} \geq e^{s\beta \frac{r-1}{r}|y|}$. Taking $x_0 \in \mathbb{R}^d$ with $|x_0| = K$, $V(x) = c\pi^{-s}(x_0)\left(\frac{\pi(x)}{\pi(x_0)}\right)^{-s} \geq cae^{s\beta\frac{r-1}{r}|x|}$ for $x = rx_0$, $r > 1$, and $a := \inf_{|y|\leq K} \pi^{-s}(y) > 0$, because of Assumption 4.3.1. If $r \geq 2$ then $\frac{r-1}{r} \geq 0.5$. Therefore, as $|x|$ is extremely large, $V(x) \geq |x|^2$. We know that $\sup_n \mathbf{E}[V(X_n)] < \infty$ (See Theorem 18 in Roberts and Rosenthal (2007)).

Since $\left\|X_n X_n^\top\right\|_M := \sup_{|u|=1} u^\top X_n X_n^\top u \leq \sup_{|u|=1} |u|^2 |X_n|^2 \leq |X_n|^2$, $\left\|X_n X_n^\top\right\|_M$ is bounded in probability.

Obviously,

$$\left\|\frac{1}{n}\sum_{i=0}^{n-1} X_i X_i^\top\right\|_M \leq \frac{1}{n}\sum_{i=0}^{n-1} \left\|X_i X_i^\top\right\|_M.$$

Then, for $K > 0$,

$$\begin{aligned}\mathbf{P}\left(\frac{1}{n}\sum_{i=0}^{n-1}\left\|X_i X_i^\top\right\|_M > K\right) \leq & \frac{1}{K}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{E}\left[\left\|X_i X_i^\top\right\|_M\right] \leq \frac{1}{K}\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{E}\left[|X_i|^2\right] \\ \leq & \frac{1}{K}\sup_n \mathbf{E}[V(X_n)].\end{aligned}$$

Hence, $\left\|\frac{1}{n}\sum_{i=0}^{n-1} X_i X_i^\top\right\|_M$ is bounded in probability.

$\left|\overline{X}_n\right| \leq \frac{1}{n+1}\sum_{i=0}^{n} |X_i|$. So,

$$\mathbf{P}(\left|\overline{X}_n\right| > K) \leq \frac{1}{K}\frac{1}{n+1}\sum_{i=0}^{n}\mathbf{E}[|X_i|] \leq \frac{1}{K}\sup_n \mathbf{E}[V(X_n)].$$

$\left|\overline{X}_n\right|$ is bounded in probability. Hence, $\left\|\overline{X}_{n-1}\overline{X}_{n-1}^\top\right\|_M$ is bounded in probability.

Finally,

$$\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M \leq 2\,|X_n|\,\left|\overline{X}_{n-1}\right|.$$

Therefore, $\left\|X_n\overline{X}_{n-1}^\top + \overline{X}_{n-1}X_n^\top\right\|_M$ is bounded in probability. □

Theorem 4.3.2. *Suppose that the target density π in Example 4.3.1 is lighter-than-*

exponentially tailed. The algorithm in Example 4.3.1 is ergodic.

Proof: Obviously, the proposal densities have uniformly lower bound function. By Theorem 4.3.1 and Proposition 4.3.2, the adaptive Metropolis algorithm is ergodic. □

The following lemma is used to check Assumption 4.3.4.

Lemma 4.3.1. *Suppose that the target density π is exponentially tailed and the proposal density family $\{q_\gamma : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}$. Suppose further that there is a function $q^-(z) := g(|z|)$, $q^- : \mathbb{R}^d \to \mathbb{R}^+$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$, some constants $M \geq 0$, $\epsilon \in (0, \eta_1)$, $\beta \in (0, \eta_2)$ and $\frac{3}{\beta\epsilon} \vee M < \delta < \Delta$ such that for $|z| \geq M$ with the property that $q_\gamma(z) \geq q^-(z)$ for $\gamma \in \mathcal{Y}$ and*

$$\frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2}, \frac{1}{2}\right)\int_\delta^\Delta g(t)t^d dt > \frac{3(e+1)}{\beta\epsilon(e-1)}, \tag{4.16}$$

where η_1 is defined in Equation (4.7), η_2 is defined in Equation (4.8), $r := \frac{\epsilon}{18}\sqrt{36-\epsilon^2}$, and the incomplete beta function $\mathrm{Be}_x(t_1, t_2) := \int_0^x t^{t_1-1}(1-t)^{t_2-1}dt$, then Assumption 4.3.4 holds.

Proof: For $u \in S^{d-1}$,

$$\int_{C_{\delta,\Delta}(u,\epsilon)} |z|\, g(|z|)\mu_d(dz) = \int_\delta^\Delta g(t)t^d dt \int_{\left\{\xi \in S^{d-1} : |\xi - u| < \epsilon/3\right\}} \omega(d\xi).$$

where $\omega(\cdot)$ denotes the surface measure on S^{d-1}.

By the symmetry of $u \in S^{d-1}$, let $u = e_d := (\underbrace{0, \cdots, 0}_{d-1}, 1)$. So, the projection from the piece $\left\{\xi \in S^{d-1} : |\xi - u| < \epsilon/3\right\}$ of the hypersphere S^{d-1} to the subspace $\mathbb{R}^{d-1}$ generated by the first $d-1$ coordinates is $d-1$ hyperball $\mathrm{B}^{d-1}(0, r)$ with the center 0 and the radius $r = \frac{\epsilon}{18}\sqrt{36-\epsilon^2}$. Define $f(z) = \sqrt{1 - (z_1^2 + \cdots + z_{d-1}^2)}$.

$$\begin{aligned}
&\omega\left(\left\{\xi \in S^{d-1} : |\xi - u| < \epsilon/3\right\}\right) = \int_{\mathrm{B}^{d-1}(0,r)} \sqrt{1 + |\nabla f|^2}\, dz_1 \cdots dz_{d-1} \\
&= \frac{(d-1)\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d+1}{2})}\int_0^r \frac{\rho^{d-2}}{\sqrt{1-\rho^2}}d\rho = \frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2}, \frac{1}{2}\right).
\end{aligned}$$

Hence,

$$\int_{C_{\delta,\Delta}(u,\epsilon)} |z|\, g(|z|)\mu_d(dz) = \frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\int_\delta^\Delta g(t)t^d dt. \tag{4.17}$$

Therefore, the result holds. □

Example 4.3.2. *Consider the standard multivariate exponential distribution* $\pi(x) = c\exp(-\lambda|x|)$ *on* $\mathbb{R}^d$ *where* $\lambda > 0$. *We perform a Metropolis algorithm with proposal distribution in the family* $\{Q_\gamma(\cdot) : \gamma \in \mathcal{Y}\}$ *at the* n^{th} *iteration where*

$$Q_n(x,\cdot) = \begin{cases} \mathrm{Unif}\left(\mathrm{B}^d(x,\Delta)\right), & n \le 2d, \text{ or } \Sigma_n \text{ is nonsingular,} \\ (1-\theta)N(x,(2.38)^2\Sigma_n/d) + \theta\,\mathrm{Unif}\left(\mathrm{B}^d(x,\Delta)\right), & n > 2d, \text{ and } \Sigma_n \text{ is singular,} \end{cases} \tag{4.18}$$

for a predetermined parameter $\theta \in (0,1)$, $\mathrm{Unif}\left(\mathrm{B}^d(x,\Delta)\right)$ *is an uniform distribution on the hyperball* $\mathrm{B}^d(x,\Delta)$ *with the center* x *and the radius* Δ, *and* Σ_n *is as defined in Equation (4.13). The problem is: how to choose* Δ *such that the adaptive Metropolis algorithm is ergodic?*

Proposition 4.3.3. *There exists a large enough* $\Delta > 0$ *such that the adaptive Metropolis algorithm of Example 4.3.2 is ergodic.*

Proof: We compute that $\nabla\pi(x) = -\lambda n(x)\pi(x)$. So, $\langle n(x), \nabla\log\pi(x)\rangle = -\lambda$ and $\langle n(x), m(x)\rangle = -1$. So, the target density is exponentially tailed, and Assumptions 4.3.1 and 4.3.2 hold. Obviously, each proposal density is locally positive. Now, let us check Assumption 4.3.4 by using Lemma 4.3.1. Let $M = 0$. Because

$$\mathrm{Vol}(\mathrm{B}^d(x,\Delta)) = \frac{\Delta^d\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2}+1)},$$

the function $g(t)$ defined in Lemma 4.3.1 is equal to $\frac{\theta\mathbb{I}(|t|\le\Delta)}{\mathrm{Vol}(\mathrm{B}^d(x,\Delta))}$. η_2 defined in Equation (4.7) and η_1 defined in Equation (4.8) are respectively λ and 1. Now, fix any

$\epsilon \in (0,1)$ and any $\delta \in (\frac{3}{\lambda}, \infty)$. The left hand side of Equation (4.16) is

$$\begin{aligned}&\frac{(d-1)\pi^{\frac{d-1}{2}}}{2\Gamma(\frac{d+1}{2})}\mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\int_{\delta}^{\Delta} g(t)t^d dt\\ =&\frac{d(d-1)}{2(d+1)\mathrm{Be}(\frac{d+1}{2},1/2)}\cdot \mathrm{Be}_{r^2}\left(\frac{d-1}{2},\frac{1}{2}\right)\cdot\Delta\left(1-\frac{\delta^{d+1}}{\Delta^{d+1}}\right),\end{aligned}$$

where $\mathrm{Be}(x,y)$ and $\mathrm{Be}_r(x,y)$ are beta function and incomplete beta function, r is a function of ϵ defined in Lemma 4.3.1.

Once ϵ and δ are fixed, the first two terms in the right hand side of the above equation is fixed. Then, as Δ goes to infinity, the whole equation tends to infinity. So, there exists a large enough $\Delta > 0$ such that Equation (4.16) holds. By Lemma 4.3.1, Assumption 4.3.4 holds. Then, by Proposition 4.3.4, Containment holds. By Proposition 4.3.2, Diminishing Adaptation holds. By Theorem 1.6.1, the adaptive Metropolis algorithm is ergodic. □

4.3.2 Some Technical Arguments

Before we show that Theorem 4.3.1, we state (Jarner and Hansen, 2000, Lemma 4.2).

Lemma 4.3.2. *Let x and z be two distinct points in $\mathbb{R}^d$, and let $\xi = n(x-z)$. If $\langle \xi, m(y)\rangle \neq 0$ for all y on the line from x to z, then z does not belong to $\{y \in \mathbb{R}^d : \pi(y) = \pi(x)\}$.*

Consider the test function $V(x) = c\pi^{-s}(x)$ for some $c > 0$ and $s \in (0,1)$ such that $V(x) \geq 1$. Note that it is not difficult to check that for $s \in (0,1)$, $\pi(V) < \infty$ by utilizing Definition 4.3.2.

By some algebra,

$$\begin{aligned}P_\gamma V(x)/V(x) = &\int_{A(x)-x}\left(\frac{\pi^s(x)}{\pi^s(x+z)}\right)q_\gamma(z)\mu_d(dz)+\\ &\int_{R(x)-x}\left(1-\frac{\pi(x+z)}{\pi(x)}+\frac{\pi^{1-s}(x+z)}{\pi^{1-s}(x)}\right)q_\gamma(z)\mu_d(dz),\end{aligned}$$

where the *acceptance region* $A(x) := \{y \in \mathcal{X} | \pi(y) \geq \pi(x)\}$, and the *potential rejection region* $R(x) := \{y \in \mathcal{X} | \pi(y) < \pi(x)\}$. From (Roberts and Rosenthal, 1998, Proposition 3), we have $P_\gamma V(x)/V(x) \leq r(s)V(x)$ where $r(s) := 1 + s(1-s)^{-1+1/s}$.

Proposition 4.3.4 (Exponential tail)**.** *Suppose that the target density π is exponentially tailed. Under Assumptions 4.3.1-4.3.4, Containment holds.*

Proof: Consider $s \in [0, 1/2)$. Under Assumption 4.3.4, let

$$h(\alpha, s) = r'(s) + \frac{1}{(1-s)^2} - \frac{\alpha}{1-s} \inf_{(u,\gamma)\in S^{d-1}\times\mathcal{Y}} \int_{C_{\delta,\Delta}(u,\epsilon)} |z| \left[e^{-\alpha s|z|} - e^{-\alpha(1-s)|z|}\right] q_\gamma(z)\mu_d(dz) \text{ and}$$

$$H(\alpha, s) = 1 + \int_0^s h(\alpha, t)dt$$

where $\epsilon, \beta, \delta, \Delta$, and $C_{\delta,\Delta}(\cdot,\cdot)$ are defined in Assumption 4.3.4. So, $H(\beta\epsilon/3, 0) = 1$ and

$$\frac{\partial H(\beta\epsilon/3, 0)}{\partial s} = h(\beta\epsilon/3, 0) \leq e^{-1} + 1 - \frac{\beta\epsilon(1-e^{-1})}{3} \inf_{(u,\gamma)\in S^{d-1}\times\mathcal{Y}} \int_{C_{\delta,\Delta}(u,\epsilon)} |z|\, q_\gamma(z)\mu_d(dz) < 0.$$

Therefore, there exists $s_0 \in (0, 1/2)$ such that $H(\beta\epsilon/3, s_0) < 1$.

Denote $C(x) := x - C_{\delta,\Delta}(n(x), \epsilon)$ and $C^\top(x) := x + C_{\delta,\Delta}(n(x), \epsilon)$. For $|x| \geq 2\Delta$ and $y \in C(x) \cup C^\top(x)$, $|y| \geq |x| - \Delta \geq \Delta$ so $|n(y) - n(x)| < \epsilon/3$.

Since the target density $\pi(\cdot)$ is exponentially tailed and Assumption 4.3.2, for sufficiently large $|x| > K_1$ with some $K_1 > 2\Delta$, $\langle n(x), \nabla \log \pi(x)\rangle \leq -\beta$ and $\langle n(x), m(x)\rangle \leq -\epsilon$. Then there exists some $K_2 > K_1$ such that for $|x| \geq K_2$, $\langle n(y), m(y)\rangle \leq -\epsilon$ for $y \in C(x) \cup C^\top(x)$. Thus, $|\nabla \log \pi(y)| = \frac{\langle n(y), \nabla \log \pi(y)\rangle}{\langle n(y), m(y)\rangle} \geq \beta$. Moreover, $y = x \pm a\xi$ for some $\delta \leq a \leq \Delta$ and $\xi \in S^{d-1}$. So,

$$\langle \xi, m(y)\rangle = \langle \xi - n(x), m(y)\rangle + \langle n(x) - n(y), m(y)\rangle + \langle n(y), m(y)\rangle < -\epsilon/3. \quad (4.19)$$

Hence, by Lemma 4.3.2, for $|x| > K_2$,

$$C(x) \cap \left\{y \in \mathbb{R}^d : \pi(y) = \pi(x)\right\} = \emptyset \text{ and } C^\top(x) \cap \left\{y \in \mathbb{R}^d : \pi(y) = \pi(x)\right\} = \emptyset.$$

For $y = x + a\xi \in C^\top(x)$,

$$\begin{aligned}
&\pi(y) - \pi(x) \\
=&\int_0^a \langle \xi, \nabla\pi(x+t\xi)\rangle \, dt \\
=&\int_0^a \langle n(x+t\xi) + \xi - n(x) + n(x) - n(x+t\xi), n(\nabla\pi(x+t\xi))\rangle \, |\nabla\pi(x+t\xi)| \, dt \\
<&(-\epsilon + \epsilon/3 + \epsilon/3) \int_0^a |\nabla\pi(x+t\xi)| \, dt \le 0
\end{aligned}$$

so that $C^\top(x) \subset R(x)$. Similarly, $C(x) \subset A(x)$.

Consider the test function $V(x) = c\pi^{-s_0}(x)$ for some $c > 0$ such that $V(x) > 1$. By Assumption 4.3.1, for any compact set $C \subset \mathbb{R}^d$, $\sup_{x\in C} V(x) < \infty$.

For any sequence $\{x_n : n \ge 0\}$ with $|x_n| \to \infty$, there exists some $N > 0$ such that $n > N$, $|x_n| > K_2$. We have

$$\begin{aligned}
P_\gamma V(x_n)/V(x_n) = &\int_{\{C(x_n)-x_n\}\cup\left\{C^\top(x_n)-x_n\right\}} I_{x_n,s_0}(z) q_\gamma(z) \mu_d(dz) + \\
&\int_{\{C(x_n)-x_n\}^c\cap\left\{C^\top(x_n)-x_n\right\}^c} I_{x_n,s_0}(z) q_\gamma(z) \mu_d(dz),
\end{aligned}$$

where

$$I_{x_n,s_0}(z) = \begin{cases} \frac{\pi^{s_0}(x_n)}{\pi^{s_0}(x_n+z)}, & z \in A(x_n) - x_n, \\ 1 - \frac{\pi(x_n+z)}{\pi(x_n)} + \frac{\pi^{1-s_0}(x_n+z)}{\pi^{1-s_0}(x_n)}, & z \in R(x_n) - x_n. \end{cases}$$

For $z = a\xi \in C^\top(x_n) - x_n$ and $t \in (0, |z|)$, by Equation (4.19)

$$\langle \xi, \nabla \log\pi(x_n + t\xi)\rangle = \langle \xi, m(x_n + t\xi)\rangle \, |\nabla \log\pi(x_n + t\xi)| < -\epsilon\beta/3.$$

So, by Assumption 4.3.4,

$$\frac{\pi(x_n+z)}{\pi(x_n)} = e^{\log\pi(x_n+z) - \log\pi(x_n)} = e^{\int_0^{|z|}\langle \xi, \nabla\log\pi(x_n+t\xi)\rangle dt} \le e^{-\beta\epsilon|z|/3} \le e^{-\beta\epsilon\delta/3} \le e^{-1}.$$

Similarly, for $z = -a\xi \in C(x_n) - x_n$,

$$\frac{\pi(x_n)}{\pi(x_n + z)} \leq e^{-\beta\epsilon|z|/3} \leq e^{-1}.$$

$t^{1-s_0} - t \leq \frac{1}{1-s_0}t^{1-s_0} - t$. Since $t \to \frac{1}{1-s_0}t^{1-s_0} - t$ is an increasing function on $[0, 1]$,

$$\begin{aligned}
&\int_{\{C(x_n)-x_n\}\cup\left\{C^\top(x_n)-x_n\right\}} I_{x_n,s_0}(z)q_\gamma(z)\mu_d(dz) \\
\leq &\int_{C(x_n)-x_n} \frac{1}{1-s_0}e^{-s_0\beta\epsilon|z|/3}q_\gamma(z)\mu_d(dz) + \\
&\int_{C^\top(x_n)-x_n} \left(1 - e^{-\beta\epsilon|z|/3} + \frac{1}{1-s_0}e^{-(1-s_0)\beta\epsilon|z|/3}\right) q_\gamma(z)\mu_d(dz).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\int_{\{C(x_n)-x_n\}^c\cap\left\{C^\top(x_n)-x_n\right\}^c} I_{x_n,s_0}(z)q_\gamma(z)\mu_d(dz) \\
\leq &r(s_0)Q_\gamma\left(\{C(x_n) - x_n\}^c \cap \left\{C^\top(x_n) - x_n\right\}^c\right).
\end{aligned}$$

Define $K_{x,\gamma}(t) := \int_{C(x)-x} e^{-t|z|}q_\gamma(z)\mu_d(dz) = \int_{C^\top(x)-x} e^{-t|z|}q_\gamma(z)\mu_d(dz)$, and

$$H_{x,\gamma}(\theta, t) := \frac{K_{x,\gamma}(t\theta)}{1-t} + K_{x,\gamma}(0) - K_{x,\gamma}(\theta) + \frac{K_{x,\gamma}((1-t)\theta)}{1-t} + r(t)(1 - 2K_{x,\gamma}(0)).$$

So,

$$P_\gamma V(x_n)/V(x_n) \leq H_{x_n,\gamma}(\beta\epsilon/3, s_0).$$

For $0 \leq t < 1/2$,

$$\begin{aligned}
&\frac{\partial H_{x,\gamma}(\theta,t)}{\partial t} \\
=&r'(t)(1-2K_{x,\gamma}(0)) + \frac{K_{x,\gamma}(\theta t) + K_{x,\gamma}(\theta(1-t))}{(1-t)^2} + \frac{\theta}{1-t}\left(K'_{x,\gamma}(\theta t) - K'_{x,\gamma}(\theta(1-t))\right) \\
\leq&r'(t) + \frac{1}{(1-t)^2} - \frac{\theta}{1-t}\int_{C(x)-x}\left(e^{-\theta t|z|} - e^{-\theta(1-t)|z|}\right)|z|\, q_\gamma(z)\mu_d(dz) \\
\leq&h(\theta,t).
\end{aligned}$$

Since $H_{x,\gamma}(\theta,0) = 1$, $H_{x,\gamma}(\theta,t) \leq H(\theta,t)$ for $0 \leq t < 1/2$. Thus, $H_{x_n,\gamma}(\beta\epsilon/3, s_0) \leq H(\beta\epsilon/3, s_0) < 1$ so $\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{P_\gamma V(x)}{V(x)} < 1$. By Corollary 4.1.1, Containment holds. □

Proof of Theorem 4.3.1: For (ii), by Proposition 4.3.4, Containment holds. Then ergodicity is implied by Containment and Diminishing Adaptation.

For (i), From Assumption 4.3.3, for any $\epsilon \in (0, \eta_1)$ and any $u \in S^{d-1}$,

$$\int_{C_{\zeta/2,\zeta}(u,\epsilon)} |z|\, q_\gamma(z)\mu_d(dz) \geq \frac{\iota\zeta \mathrm{Vol}(C_{\zeta/2,\zeta}(u,\epsilon))}{2}$$

where ι is defined in Equation (4.9), ζ is defined in Assumption 4.3.3, $C_{a,b}(\cdot,\cdot)$ is defined in Equation (4.10). The right hand side of the above equation is positive and independent of γ and u. Since target density is lighter-than-exponentially tailed, $\eta_2 = +\infty$ such that there is some sufficiently large β such that Equation (4.11) holds. So, Assumption 4.3.4 is satisfied.

For (iii), adopting the proof of Fort and Moulines (2000a, Theorem 5), we will show that the simultaneous drift condition Equation (3.9) holds. Denote $R(g,x,y) := g(y) - g(x) - \langle \nabla g(x), y - x\rangle$.

$$\sup_{|z|\leq M} |R(g,x,x+z)|\,|z|^{-2} \leq \sup_{t\in\bar{\mathrm{B}}^d(x,M)} \left|\nabla^2 g(t)\right|/2.$$

Consider the test function $V(x) := 1 + f^s(x)$ where $f(x) := -\log\pi(x)$ for $\frac{2}{m} - 1 < s < \min(\frac{2}{m}, \frac{3}{m} - 2)$. By Assumption 4.3.5 and Fort and Moulines (2000a, Lemma

B.2),

$$P_\gamma V(x) - V(x) = P_\gamma f^s(x) - f^s(x) = \sum_{j=0}^{4} I_j,$$

where

$$\begin{aligned}
I_0 &:= -\, s f^{s-1}(x) \left|\nabla f(x)\right|^2 \int_{R(x)-x\cap\{|z|\le M\}} \langle m(x), n(z)\rangle^2 \left|z\right|^2 q_\gamma(z)\mu_d(dz),\\
I_1 &:= \int_{\{|z|\le M\}} R(f^s, x, x+z) q_\gamma(z)\mu_d(dz) \le M^2 \sup_{t\in\bar{\mathrm{B}}^d(x,M)} \left|\nabla^2 f^s(t)\right|/2\\
I_2 &:= \int_{R(x)-x\cap\{|z|\le M\}} R(f^s, x, x+z) R(\pi, x, x+z) \frac{q_\gamma(z)}{\pi(x)} \mu_d(dz)\\
I_3 &:= \int_{R(x)-x\cap\{|z|\le M\}} R(f^s, x, x+z) \left\langle \nabla f(x), z\right\rangle q_\gamma(z)\mu_d(dz)\\
I_4 &:= \int_{R(x)-x\cap\{|z|\le M\}} R(\pi, x, x+z) \left\langle \nabla f^s(x), z\right\rangle \frac{q_\gamma(z)}{\pi(x)} \mu_d(dz).
\end{aligned}$$

By some algebra, $\nabla^2\pi(x) = \left(\nabla f(x)\nabla f(x)^\top - \nabla^2 f(x)\right)\pi(x)$. By Definition 4.3.3 and Assumption 4.3.1,

$$\sup_{t\in\bar{\mathrm{B}}^d(x,M)} \left|\nabla^2 f^s(t)\right| = \mathrm{O}(|x|^{ms-2}) \text{ and } \sup_{t\in\bar{\mathrm{B}}^d(x,M)} \left|\nabla^2 \pi(t)\right| \le \mathrm{O}(|x|^{2(m-1)}).$$

Hence,

$$|I_1| \le \mathrm{O}(|x|^{ms-2}), |I_2| \le \mathrm{O}(|x|^{m(s+2)-4}), |I_3| \le \mathrm{O}(|x|^{m(s+1)-3}), |I_4| \le \mathrm{O}(|x|^{m(s+2)-3}).$$

Since $\frac{2}{m} - 1 < s < \min(\frac{2}{m}, \frac{3}{m} - 2)$, $|I_1|$, $|I_2|$, $|I_3|$ and $|I_4|$ converge to zero as $|x| \to \infty$. By Assumption 4.3.2, for any $\epsilon \in (0, \eta_1)$ (η_1 is defined in Equation (4.8)), $\langle n(x), m(x)\rangle < -\epsilon$ as $|x|$ is sufficiently large. By Assumption 4.3.3, for any $z \in C_{0,\zeta}(n(x), \epsilon)$ (ζ is defined in Assumption 4.3.3, ι is defined in Equation (4.9), and $C_{\cdot,\cdot}(\cdot,\cdot)$ is defined in Equation (4.10)), $-1 \le \langle m(x), n(z)\rangle = \langle m(x), n(x)\rangle +$

$\langle m(x), n(z) - n(x)\rangle \leq -\epsilon + \epsilon/3$.

$$\begin{aligned} I_0 \leq & -\frac{4\epsilon^2 \iota s f^{s-1}(x) \left|\nabla f(x)\right|^2}{9} \int_{C_{0,\zeta}(n(x),\epsilon)} |z|^2 \mu_d(dz) \\ = & -c_1 f^{s-1}(x) \left|\nabla f(x)\right|^2 \leq c_2 f^{s-(2-m)/m}(x), \end{aligned}$$

for some $c_1 > 0$ (independent of x) where $C_{0,\zeta}(n(x), \epsilon) = C_{0,\zeta}(u, \epsilon)$ for any $u \in S^{d-1}$.

So, there exist some $K > 0$ and some $c_3 > 0$ such that $V(x) > 1.1$ and $P_\gamma V(x) - V(x) \leq -c_3 V^\alpha(x)$ for $|x| > K$, some $\alpha \in (0, 1)$. Let $\tilde{V}(x) := V(x)\mathbb{I}(|x| > K) + \mathbb{I}(|x| \leq K)$. So,

$$P_\gamma \tilde{V}(x) - \tilde{V}(x) \leq -c_3 \tilde{V}^\alpha(x) + c_3 \mathbb{I}(|x| \leq K).$$

By the part (iii) of Theorem 3.3.2, Containment holds. □

4.4 Adaptive Metropolis-within-Gibbs Algorithms

In the section, we study *adaptive random-scan Metropolis-within-Gibbs algorithms* on the state space $\mathcal{X} = \mathbb{R}^d$. Consider a family $\{P_{\mathrm{RS},\gamma} : \gamma \in \mathcal{Y}\}$ of random-scan Metropolis-within-Gibbs algorithms, i.e. each $P_{\mathrm{RS},\gamma}$ is a random-scan Metropolis-within-Gibbs sampler.

Define the *symmetric proposal density family on some direction* $e \in S^{d-1}$, $\mathfrak{C}(e) := \{q : q(x, x + ze) = q(x, x - ze) = q(z)$ for $x \in \mathbb{R}^d, z \in \mathbb{R}\}$. Suppose that $P_{i,\gamma}$ is the transition kernel generated by a symmetric random-walk-based Metropolis algorithm with the proposal $q_{i,\gamma} \in \mathfrak{C}(e_i)$. Then

$$P_{\mathrm{RS},\gamma} = \frac{1}{d} \sum_{i=1}^{d} P_{i,\gamma}. \tag{4.20}$$

For a Borel set $A = A_1 \times \cdots \times A_d$ on $\mathbb{R}^d$ and $x := (x_1, \cdots, x_d) \in \mathbb{R}^d$ and $z \in \mathbb{R}$,

$$\begin{aligned} P_{i,\gamma}(x, A) := \prod_{k \neq i} \delta_{x_k}(A_k) \int_{A_i - x_i} \alpha(x, x + ze_i) q_{i,\gamma}(z) \mu(dz) + \\ \delta_x(A) \int (1 - \alpha(x, x + ze_i))\, q_{i,\gamma}(z) \mu(dz), \end{aligned} \tag{4.21}$$

where $A_i - x = \{y \in \mathbb{R} : x_i + y \in A_i\}$. Say $A(x, i) := \{z \in \mathbb{R} : \pi(x + ze_i) \geq \pi(x)\}$ and $R(x, i) := \{z \in \mathbb{R} : \pi(x + ze_i) < \pi(x)\}$ are the acceptance region and potential rejection region in the ith direction respectively.

Assumption 4.4.1 (Target Regularity)**.** *Same as Assumption 4.3.1.*

Assumption 4.4.2 (Uniform Local Positivity)**.** *Assume that* $\{q_{i,\gamma} : \gamma \in \mathcal{Y}\} \subset \mathfrak{C}(e_i)$ *for* $i = 1, \cdots, d$*. There exists* $\zeta > 0$ *such that*

$$\inf_{i=1,\cdots,d} \inf_{\gamma \in \mathcal{Y}} \inf_{|z| \leq \zeta} q_{i,\gamma}(z) > 0. \tag{4.22}$$

Assumption 4.4.3 (Exponential tails on the coordinates $\{e_1, \cdots, e_d\}$)**.** *There exist* $\beta > 0$*,* $\delta > 0$*, and* $\Delta > 0$ *with* $1/\beta \leq \delta < \Delta \leq \infty$ *such that for any sequence* $\{x_n : n \geq 0\}$ *with* $\lim_n |x_n| = \infty$*, we may extract a subsequence* $\{\tilde{x}_n : n \geq 0\}$ *with the property that for some* $i \in \{1, \cdots, d\}$ *and* $z \in [\delta, \Delta]$*,*

$$\begin{aligned} \lim_n \frac{\pi(\tilde{x}_n)}{\pi(\tilde{x}_n - \text{sign}(< \tilde{x}_n, e_i >) z e_i)} \leq & \exp(-\beta z) \text{ and} \\ \lim_n \frac{\pi(\tilde{x}_n + \text{sign}(< \tilde{x}_n, e_i >) z e_i)}{\pi(\tilde{x}_n)} \leq & \exp(-\beta z); \end{aligned} \tag{4.23}$$

Assumption 4.4.4 (Moment Condition)**.** *Under Assumption 4.4.3,*

$$\inf_{\gamma \in \mathcal{Y}} \inf_{i \in \{1,\ldots,d\}} \int_{\delta}^{\Delta} z q_{i,\gamma}(z) \mu(dz) > \frac{d + e}{\beta(e - 1)}. \tag{4.24}$$

Consider the test function $V_s(x) = c\pi^{-s}(x)$ for some $c > 0$ and $s \in (0, 1)$ such that $V_s(x) \geq 1$. For $i = 1, \cdots, d$ and $\gamma \in \mathcal{Y}$, $P_{i,\gamma} V_s(x) = \int I(z, x, i, s) q_{i,\gamma}(z) \mu(dz) \leq$

$r(s)V_s(x)$ where $r(s) = 1 + s(1-s)^{-1+1/s}$ and

$$I(z,x,i,s) := \begin{cases} \left(\frac{\pi(x)}{\pi(x+z\,e_i)}\right)^s, & z \in A(x,i), \\ 1 - \frac{\pi(x+z\,e_i)}{\pi(x)} + \left(\frac{\pi(x+z\,e_i)}{\pi(\tilde{x})}\right)^{1-s}, & z \in R(x,i). \end{cases} \tag{4.25}$$

For adaptive Metropolis-within-Gibbs algorithms, we mainly adopt the method of Fort et al. (2003).

Theorem 4.4.1. *Under Assumptions 4.4.1-4.4.4, adaptive random-scan Metropolis-within-Gibbs algorithms with Diminishing Adaptation are ergodic.*

Proof: Under Assumption 4.4.4, for $t \in [0, 1/2)$, let

$$h(\alpha,t) = r'(t) + \frac{1}{d(1-t)^2} - \frac{\alpha}{d(1-t)} \inf_{\gamma\in\mathcal{Y}} \inf_{i=1,\cdots,d} \int_\delta^\Delta z(e^{-\alpha t z} - e^{-\alpha(1-t)z}) q_{i,\gamma}(z)dz \text{ and}$$

$$H(\alpha,t) = 1 + \int_0^t h(\alpha,u)du.$$

So, $H(\beta,0) = 1$ and

$$\frac{\partial H(\beta,0)}{\partial t} = h(\beta,0) \le e^{-1} + \frac{1}{d} - \frac{\beta(1-e^{-1})}{d} \inf_{\gamma\in\mathcal{Y}} \inf_{i\in\{1,\ldots,d\}} \int_\delta^\Delta z q_{i,\gamma}(z)\mu(dz) < 0.$$

So there exists $s_0 \in (0, 1/2)$ such that $H(\beta\epsilon/3, s_0) < 1$.

Assume that $\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} P_{\mathrm{RS},\gamma} V_{s_0}(x)/V_{s_0}(x) \ge 1$. So there exists a sequence $\{(x_n,\gamma_n) : n \ge 0\}$ with $\lim_n |x_n| = \infty$ such that $\lim_n P_{\mathrm{RS},\gamma_n} V_{s_0}(x_n)/V_{s_0}(x_n) \ge 1$.

Under Assumption 4.4.3, there exists a subsequence $\{(\tilde{x}_n,\tilde{\gamma}_n) : n \ge 0\}$ such that Equation (4.23) holds with the corresponding parameters β, δ, Δ, and e_i.

$$\begin{aligned} P_{\mathrm{RS},\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) =& \frac{1}{d} P_{i,\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) + \frac{1}{d}\sum_{j\ne i} P_{j,\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) \\ \le& \frac{1}{d} P_{i,\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) + \frac{d-1}{d} r(s_0) \end{aligned}$$

Without loss of generality, assume $sign(< \tilde{x}_n, e_i >) = 1$. Let $J(\delta,\Delta) = [-\Delta,-\delta] \cup$

$[\delta, \Delta]$. It is easy to prove that

$$\lim_n R(\tilde{x}_n, i) \cap J(\delta, \Delta) = [\delta, \Delta] \text{ and } \lim_n A(\tilde{x}_n, i) \cap J(\delta, \Delta) = [-\Delta, -\delta].$$

So,

$$\begin{aligned} P_{i,\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) &= \int_{J(\delta,\Delta)} I(z, \tilde{x}_n, i, s_0) q_{i,\tilde{\gamma}_n}(z) dz + \int_{J(\delta,\Delta)^c} I(z, \tilde{x}_n, i, s_0) q_{i,\tilde{\gamma}_n}(z) dz \\ &\leq \int_{J(\delta,\Delta)} I(z, \tilde{x}_n, i, s_0) q_{i,\tilde{\gamma}_n}(z) dz + r(s_0) \int_{J(\delta,\Delta)^c} q_{i,\tilde{\gamma}_n}(z) dz. \end{aligned}$$

$t^{1-s_0} - t \leq \frac{1}{1-s_0} t^{1-s_0} - t$. Since $t \to \frac{1}{1-s_0} t^{1-s_0} - t$ is an increasing function on $(0, 1)$,

$$\begin{aligned} P_{i,\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) \leq & \frac{K_{i,\tilde{\gamma}_n}(\beta s_0)}{1 - s_0} + K_{i,\tilde{\gamma}_n}(0) + \frac{K_{i,\tilde{\gamma}_n}(\beta(1 - s_0))}{1 - s_0} - K_{i,\tilde{\gamma}_n}(\beta) + \\ & r(s_0)(1 - 2K_{i,\tilde{\gamma}_n}(0)) \end{aligned}$$

where

$$K_{i,\gamma}(t) = \int_\delta^\Delta e^{-tz} q_{i,\gamma}(z) \mu(dz). \tag{4.26}$$

Hence,

$$P_{\mathrm{RS},\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) \leq H_{i,\tilde{\gamma}_n}(\beta, s_0)$$

where

$$H_{i,\gamma}(\beta, t) = \frac{r(t)}{d}(d - 2K_{i,\gamma}(0)) + \frac{1}{d}\left(\frac{K_{i,\gamma}(\beta t)}{1 - t} + K_{i,\gamma}(0) + \frac{K_{i,\gamma}(\beta(1 - t))}{1 - t} - K_{i,\gamma}(\beta) \right). \tag{4.27}$$

For $0 \leq t < 1/2$,

$$\begin{aligned} & \frac{\partial H_{i,\gamma}(\beta, t)}{\partial t} \\ \leq & r'(t) + \frac{1}{d(1 - t)^2} + \frac{\beta}{d(1 - t)} \left(K'_{i,\gamma}(\beta t) - K'_{i,\gamma}(\beta(1 - t)) \right) \\ \leq & h(\beta, t) \end{aligned}$$

Since $H_{i,\gamma}(\beta, 0) = 1$, $H_{i,\gamma}(\beta, t) \leq H(\beta, t)$ for $t \in [0, 1/2)$. Thus, $H_{i,\tilde{\gamma}_n}(\beta, s_0) \leq H(\beta, s_0) < 1$. Contradiction! By Corollary 4.1.1, Containment holds. □

Assumption 4.4.5 (Lighter-than-exponential tails on the coordinates $\{e_1, \cdots, e_d\}$)**.** *There exist* $0 \leq \delta < \Delta \leq \infty$ *such that for any sequence* $\{x_n : n \geq 0\}$ *with* $\lim_n |x_n| = \infty$*, we may extract a subsequence* $\{\tilde{x}_n : n \geq 0\}$ *with the property that*

$$\lim_{n\to\infty} \frac{\pi(\tilde{x}_n)}{\pi(\tilde{x}_n - \text{sign}(<\tilde{x}_n, e_i>)ze_i)} = 0 \text{ and } \lim_{n\to\infty} \frac{\pi(\tilde{x}_n + \text{sign}(<\tilde{x}_n, e_i>)ze_i)}{\pi(\tilde{x}_n)} = 0. \tag{4.28}$$

Theorem 4.4.2. *Under Assumptions 4.4.1, 4.4.2, and 4.4.5, adaptive random-scan Metropolis-within-Gibbs algorithms with Diminishing Adaptation are ergodic.*

To prove it, adopt the technique in the proof of the part (i) of Theorem 4.3.1.

Example 4.4.1. *Consider the mixed distribution on* $\mathbb{R}^2$

$$\pi(x) = \beta \exp(-(x_1^2 + x_2^2)) + (1-\beta)exp(-(x_1^2 + x_1^2x_2^2 + x_2^2))$$

where $\beta \in [0, 1]$*. The family* $\{P_{\text{RS},\gamma}, \gamma \in \mathcal{Y}\}$ *consists of transition kernels generated by random-scan random-walk-based Metropolis-within-Gibbs algorithms with a set of proposal density families* $\{q_{i,\gamma}(\cdot) : \gamma \in \mathcal{Y}\}$ *for* $i = 1, 2$*, see Equation (4.20) and Equation (4.21). Assume that the proposal density family satisfies Assumption 4.4.2.*

Proposition 4.4.1. *For the target distribution and the sampler family in Example 4.4.1, any adaptive MCMC algorithm with Diminishing Adaptation is ergodic.*

Proof: We have that

$$\begin{aligned}
\frac{\nabla_1 \log \pi(x)}{-2x_1} =& \frac{\beta \exp(-(x_1^2 + x_2^2)) + (1 + x_2^2)(1-\beta)\exp(-(x_1^2 + x_1^2x_2^2 + x_2^2))}{\pi(x)} \\
\in& \left[1, 1 + x_2^2\right], \\
\frac{\nabla_2 \log \pi(x)}{-2x_2} =& \frac{\beta \exp(-(x_1^2 + x_2^2)) + (1 + x_1^2)(1-\beta)\exp(-(x_1^2 + x_1^2x_2^2 + x_2^2))}{\pi(x)} \\
\in& \left[1, 1 + x_1^2\right].
\end{aligned}$$

Clearly, $\nabla_i \log \pi(x)/(-2x_i)$ is positive bounded. So, Assumption 4.4.5 holds. Thus, by Theorem 4.4.2, the algorithm is ergodic. □

We consider the target density of Fort et al. (2003, Example 8), a mixture of two exponential distributions.

Example 4.4.2. *For $a > 1$, the target density on $\mathbb{R}^2$ is*

$$\pi(x) \propto 0.5e^{-|x_1|-a|x_2|} + 0.5e^{-a|x_1|-|x_2|}, \qquad x = (x_1, x_2).$$

The family $\{P_{\mathrm{RS},\gamma} : \gamma := (\gamma_1, \gamma_2) \in \mathbb{R}^{+2}\}$ consists of transition kernels generated by random-scan random-walk-based Metropolis-within-Gibbs algorithms with a set of proposal density families $\{q_{i,\gamma}(z) := (1-\theta)N(0,\gamma_i) + \theta\mathrm{Unif}(-b,b) : \gamma \in \mathbb{R}^2\}$ for $i = 1, 2$ (see Equation (4.20) and Equation (4.21)) where $\theta \in (0,1)$ and $b > 0$ are predetermined parameters.

Proposition 4.4.2. *For the target distribution and the sampler family in Example 4.4.2, there exists a sufficiently large $b > 0$ such that any adaptive random-scan Metropolis-within-Gibbs algorithm with Diminishing Adaptation is ergodic.*

Proof: Assumption 4.4.3 holds and $\beta = 1$, see details in Fort et al. (2003, Example 8).

$$\inf_{\gamma\in\mathcal{Y}} \inf_{i\in\{1,2\}} \int_1^b z q_{i,\gamma}(z)\mu(dz) \geq \theta \int_1^b z \frac{1}{2b} dz = \frac{\theta(b^2-1)}{4b}.$$

Obviously, there exists a sufficiently large b such that $\frac{\theta(b^2-1)}{4b} \geq \frac{2+e}{e-1}$. So, Assumption 4.4.4 holds. Since proposal densities have the same fixed part (Uniform distribution), Assumption 4.4.2 holds. By Theorem 4.4.1, the result holds. □

Chapter 5

An Adaptive Directional Metropolis-within-Gibbs algorithm

Classical Metropolis-within-Gibbs algorithms only propose values in the coordinates directions, and then accept or reject the values. When target distributions have strong correlations in some directions, the MCMC algorithm may not work very well especially on a high dimensional space, because many waste jumps are proposed. In this chapter we propose a simple adaptive Metropolis-within-Gibbs algorithm (ADMG) attempting to study directions from historical data and jump in these directions. The effective directions are extracted from the empirical covariance matrix through singular value decomposition. Some sufficient conditions for ergodicity are given. We also apply the adaptive algorithm on a Gaussian Needle example and a real-life Case-Cohort study example with competing risks. For the Cohort study, an extensive version of Competing Risks Regression model is proposed, and then the algorithm is used to estimate coefficients based on the posterior distribution.

A toy example will be presented in Section 5.1 for explanations. In Section 5.2 we propose ADMG. The idea is similar to that of the Hit-and-Run algorithm. The framework of Hit-and-Run is to uniformly draw a random direction in the unit hypersphere, and then sample a scalar from some proposal distribution in the chosen direction, see the literature Bélisle et al. (1993); Chen and Schmeiser (1993); Gilks

et al. (1994); Roberts and Gilks (1994); Chen and Schmeiser (1996); Kaufman and Smith (1998); Lovász (1999); Lovász and Vempala (2003, 2006); Bèdard and Fraser (2008). Metropolis with single particle moves, Gibbs sampler, Swendsen-Wang, data augmentation, and slice sampling have the same basic structure, see Andersen and Diaconis (2007). The ADMG algorithm tries to find directions and corresponding jumping scalars through studying certain estimates of the empirical covariance matrix of the sample chain. Then Metropolis-within-Gibbs sampler is run in the obtained directions with the jumping scalars as variances. The method can suppress the proportion of wasting moves by proposals from full dimensional Metropolis algorithm. We also compare it with Metropolis-within-Gibbs sampler and adaptive Metropolis algorithm through analysing the toy example on 10-dimensional Euclidean space. Then we show its ergodicity.

In Section 5.3 we discuss a real-life Case-Cohort study for the application, where the dataset was from the Princess Margaret Hospital, a leading cancer centre in North America. Cohort study is commonly based on the survival model. In practice, the likelihood function turns to be more and more complicated as the number of observations increases. The trade-off alternative, partial likelihood function is more interesting. Given a prior distribution, we consider the posterior distribution, and implement our algorithm to find the estimates of the coefficients of the interesting covariates in the study.

5.1 A Toy Example

Let the target density

$$t(x) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-x^\top \left(\begin{bmatrix} \cos\eta & -\sin\eta \\ \sin\eta & \cos\eta \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \cos\eta & \sin\eta \\ -\sin\eta & \cos\eta \end{bmatrix} \right)^{-1} x/2 \right), \tag{5.1}$$

where $\eta = 45^o$, $\sigma_1 = \sqrt{20}$ and $\sigma_2 = 0.01$. The target distribution has extremely small variance 0.0001 and large variance 20 respectively in the two directions $(-\sqrt{2}/2, \sqrt{2}/2)$ and $(\sqrt{2}/2, \sqrt{2}/2)$. So, the target is mainly supported on a very narrow region along the 45° degree direction between the x_1-axis and the x_2-axis. The

length of the needle region is roughly $2 * 4 * \sqrt{20} = 35.78$ (see the true sample data in Figure 5.1) because $P(|Z| < 4) \approx 1$ where Z is standard normal.

We run random-scan Metropolis-within-Gibbs sampler (MwG) defined in Section 1.3 and Adaptive Metropolis algorithm (AM) defined in Example 4.3.1 to generate sample data. For Adaptive Metropolis, the weights of mixture proposal are chosen as $\theta = 0.3$. Here the state space is $\mathbb{R}^d$ with $d = 2$. We set their burn-in time to be zero, and the initial points to be $X_0 \sim N(\overrightarrow{0}, I_2)$.

Given the sampled data $\{X_0, X_1, \cdots\}$ and the proposal values $\{Y_1, Y_2, \cdots\}$, the *k-step average of acceptance rates* is defined as

$$\alpha_i^{(k)} := \frac{1}{k} \sum_{t=ki}^{k(i+1)-1} \alpha(X_t, Y_{t+1}), \tag{5.2}$$

where $i = 0, 1, \cdots$.

We perform the random-scan MwG sampler $300,000$ iterations using the normal distribution with variance 0.1 as the proposal distribution, see the top two plots of Figure 5.1. From the sample plot, the sample data has a needle shape with the length around 4.95 ($\ll 35.78$) roughly between the two points, $(-2.0, -2.0)$ and $(1.5, 1.5)$. The 100-step average $\{\alpha_n^{(100)}\}$ of acceptance rates is roughly between 0.10 to 0.3. We also tried normal proposals with another variance 0.0001 (same as the target's) which also gives worse results. For random-scan MwG sampler, at each step, the jumping direction of the sample chain can be just in either the axis x_1 or the axis x_2 so the jumping scale is strongly limited. Moreover, the 100-step average of acceptance rates is very sensitive to the proposal variance. When the proposal variance is large, the proposal values are easily rejected. When the proposal variance is small, the proposal values are easily accepted but the chain is easily stuck.

We also perform $300,000$ iterations of AM. The algorithm attempts to find a better transition kernel by learning the empirical covariance matrix Σ of the sample chain. The sample points also span roughly the narrow stripe with the length around $4.95 \ll 35.78$ between the two points, $(-1.5, -1.5)$ and $(2, 2)$, see the third plot in Figure 5.1. At the same time, the 100-step average of acceptance rates is quite small, see the center right plot in Figure 5.1. So the sampling method for this example also does not work well.

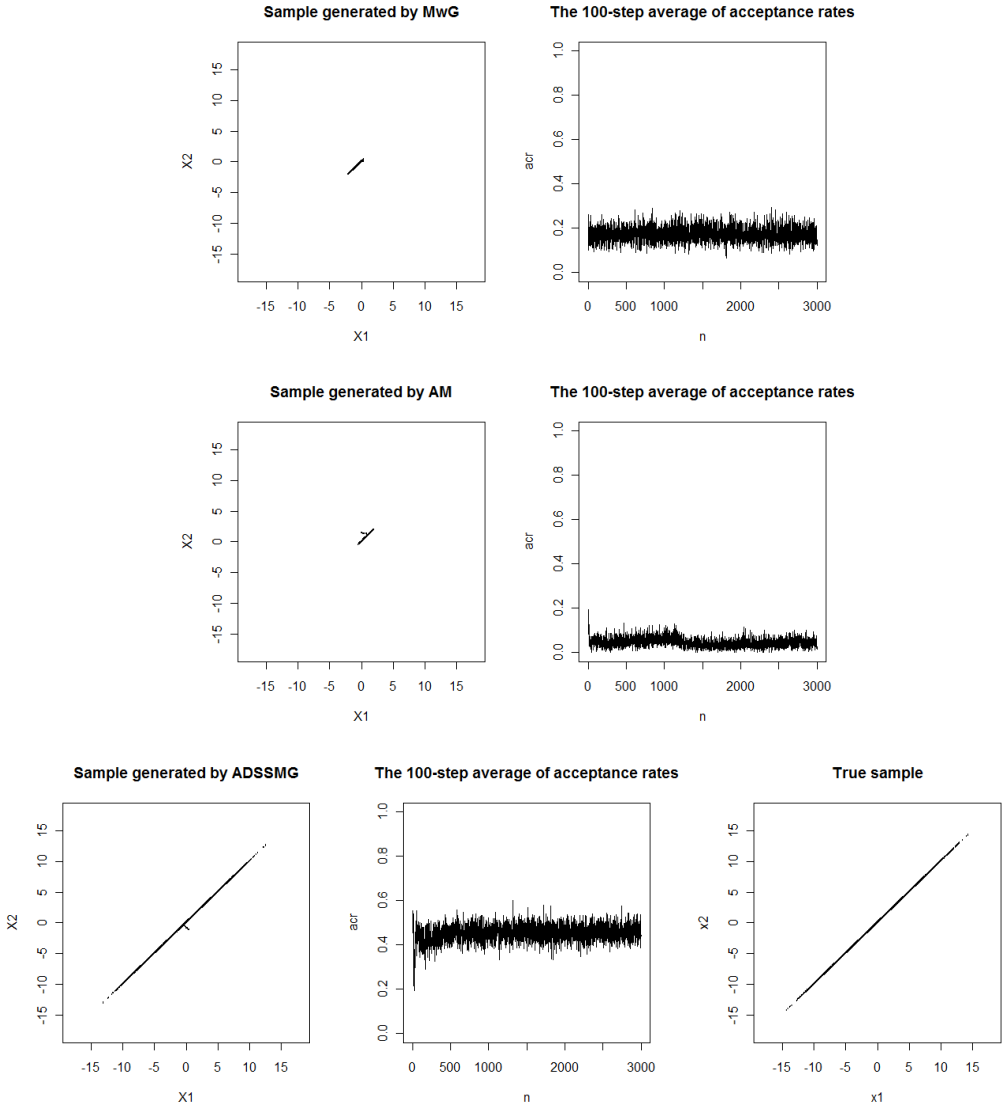

Figure 5.1: The first top plot is the sample plot by running random-scan MwG sampler. The right top plot is the 100-step average of acceptance rates by random-scan MwG sampler. The left center plot is the sample plot by running AM. The right center plot is the 100-step average of acceptance rates by AM algorithm. The left bottom plot is the sample plot by running ADSSMG. The middle bottom plot is the 100-step average of acceptance rates. The right bottom plot is the sample data directly simulated from the target distribution.

Let us observe the estimate of empirical covariance matrix Σ_n for $n = 300,000$,

$$\Sigma_n = \begin{bmatrix} 1.449585 & 1.448932 \\ 1.448932 & 1.449020 \end{bmatrix}.$$

By singular value decomposition, we have $\Sigma_n = UDV$ where

$$\begin{aligned} U &= \begin{bmatrix} -0.7071758 & -0.7070378 \\ -0.7070378 & 0.7071758 \end{bmatrix}, \\ D &= \begin{bmatrix} 2.8982345593 & 0 \\ 0 & 0.0003700081 \end{bmatrix}, \\ V &= U^\top. \end{aligned} \tag{5.3}$$

It is not difficult to find that the matrix U is approximately equal to the U matrix by singular value decomposition on the true covariance matrix of the Gaussian density $t(\cdot)$. The first diagonal element d_1 of D underestimates the variance 20 on the direction $(\sqrt{2}/2, \sqrt{2}/2)$, and the second element d_2 overestimates the variance 0.0001 on the direction $(-\sqrt{2}/2, \sqrt{2}/2)$, see Equation (5.3).

The above fact discloses that AM hardly touches the pinpoint of the needle, actually taking too much time to wander around the middle region of the needle. From the bottom right plot in Figure 5.1, the point can be also observed. The 100-step average $\{\alpha_n^{(100)}\}$ of acceptance rates is very low, approximately below 0.10 that the adaptation wastes too many proposals in "wrong" directions. Hence the inefficiency of AM is mainly due to the jumping directions.

5.2 The Algorithm and Ergodicity

Drawing a proposal value in the high dimensional space involves the direction choice and the jump scale in the direction. The direction choice can be viewed as taking a unit vector on the unit sphere. The jump scale can be viewed as the variance of the proposal marginal distribution in the chosen direction. The aim in ADMG is to find the random directions in which the efficient movement can be ensured. As illustrated in Section 5.1, the random direction can be drawn from the estimate

of empirical covariance matrix. After singular value decomposition, the orthogonal transformation can be obtained. Moreover, the diagonal matrix also approximately estimates the target extents in those directions after the rotation. Based on the orthogonal transformation and the extents in the new coordinates, the Metropolis-within-Gibbs sampler can be run flexibly.

5.2.1 ADMG

In the section, we study ergodicity of *adaptive directional random-scan Metropolis-within-Gibbs algorithms* (ADRSMwG) and *adaptive directional deterministic-scan Metropolis-within-Gibbs algorithms* (ADDSMwG). The adaptation are defined in the previous section.

The adaptive parameter set $\mathcal{Y}$ is the set of positive definite matrixes in $\mathbb{R}^{d\times d}$. So for $\gamma \in \mathcal{Y}$, there exist an unitary matrix Q and a diagonal matrix $D := \mathrm{diag}(k_1, \cdots, k_d)$ such that $\gamma = Q^\top DQ$.

For $\gamma \in \mathcal{Y}$, the collection $\{q_{i,\gamma} : i = 1, \cdots, d\}$ of the proposal densities is used to sample data on the rotated directions $(\tilde{e}_1, \cdots, \tilde{e}_d) =: \tilde{e} = Q^\top e$ where $e = (e_1, \cdots, e_d)$. On the direction $\tilde{e}_i$, the proposal distribution

$$Q_{i,\gamma}(x, \cdot) = x + (\theta N(0, k_i) + (1-\theta)N(0, \epsilon))\tilde{e}_i \tag{5.4}$$

where $\theta \in (0, 1)$ and $\epsilon > 0$ are predetermined. The sample $P_{i,\gamma}$ is the transition kernel of the symmetric random walk Metropolis-Hastings algorithm with the proposal distribution $Q_{i,\gamma}(x, \cdot)$. Denote $P_{\mathrm{DRS},\gamma} := \frac{1}{d}\sum_{i=1}^{d} P_{i,\gamma}$ and $P_{\mathrm{DDS},\gamma} := P_{1,\gamma} \cdots P_{d,\gamma}$.

The adaptation is defined as that at each iteration n, the empirical covariance matrix is

$$\Sigma_n = \begin{cases} \lambda I_d, & n \le d, \\ s_d \left(\frac{1}{n} \left(\sum_{i=0}^{n} X_i X_i^\top - (n+1)\overline{X}_n \overline{X}_n^\top \right) + \lambda I_d \right), & n > d, \end{cases} \tag{5.5}$$

where s_d is some predetermined parameter, e.g. $(2.38)^2/d$ is used in Example 4.3.1, $\lambda > 0$ is also a predetermined parameter. The matrix Σ_n is positive definite.

Step 1. Given $X_0, \cdots, X_n$, we can compute the empirical covariance matrix Σ_n defined in Equation (5.5).

Step 2. Do singular value decomposition: $\Sigma_n = U^{(n)} D^{(n)} V^{(n)}$ where $D^{(n)} := \text{diag}(d_1^{(n)}, \cdots, d_d^{(n)})$, and $U^{(n)}$ and $V^{(n)} = \left(U^{(n)}\right)^\top$ are orthonormal;

Step 3. Compute the random direction $\tilde{e}_i^{(n)} := U^{(n)} e_i$ where $e_i = \underbrace{(0, \cdots, 0, 1, 0, \cdots, 0)}_{i^{th}}$;

Step 4. According to the framework of ransom-scan Metropolis-within-Gibbs sampler, perform Metropolis algorithm on the new coordinates $\left(\tilde{e}_1^{(n)}, \cdots, \tilde{e}_d^{(n)}\right)$, i.e. in the direction $\tilde{e}_i^{(n)}$, the proposal distribution is $Q_{i,\Sigma_n}(X_n, \cdot)$ where the variance of the adaptive part distribution in Equation (5.4) is $d_i^{(n)}$.

Step 5. $n := n + 1$ and go to Step 1.

Remark 5.2.1. *In step 2, it may takes much time to do singular value decomposition when the state space is high dimensional. However, it is unnecessary to run the computation for each step. The alternative is to do singular value decomposition each m steps. Another method is to only count the accepted sample point to compute the estimate of empirical covariance matrix.*

Remark 5.2.2. *In step 4, we give one scheme to scale the variance of proposal distribution. The idea is that if the k-step average of acceptance rates is too large which implies that the jump scalar is too small, the proposal variance is required to be larger for the efficiency; if $\alpha_{[n/k]}^{(k)}$ is too small which implies that the jump scalar is too large, the proposal variance is required to be smaller for the efficiency. Here, we can increase the proposal variance if $\alpha_{[n/k]}^{(k)} > 0.3$, and decrease it if $\alpha_{[n/k]}^{(k)} < 0.3$. Actually, the pair parameter $(0.3, 0.3)$ can be tuned. E.g. define $\lambda_n = \mathbb{I}(\alpha_{[n/k]}^{(k)} > 0.5) \exp(2d(\alpha_{[n/k]}^{(k)} - 0.5)) + \mathbb{I}(\alpha_{[n/k]}^{(k)} < 0.2) \exp(2d(\alpha_{[n/k]}^{(k)} - 0.2)) + \mathbb{I}(0.2 \leq \alpha_{[n/k]}^{(k)} \leq 0.5)$.*

Considering again the example in Section 5.1, we run ADMG $300,000$ iterations, and still set their burn-in time to be zero and the initial points to be $X_0 \sim N(\overrightarrow{0}, I_2)$, see the left bottom and middle bottom plots in Figure 5.1. The simulated data span roughly from $(-15, -15)$ to $(15, 15)$ which shows that ADMG detects the target faster than MwG and AM. The 100-step average $\{\alpha_n^{(100)}\}$ of acceptance rates is between 0.35 to 0.52. The right bottom plot in Figure 5.1 is a true sample data from $t(\cdot)$.

Comparing the bottom left plot and the bottom right plot, ADMG exactly discovered the target region.

Remark 5.2.3. *From the discussion of the toy example, it is not difficult to find that when a target distribution with high correlations is mainly supported on a long narrow region, ADMG is much more efficient than the MwG sampler. In the high dimensional space, the phenomenon is more explicit.*

5.2.2 High dimensional Gaussian Needle

Here, we simulate a 10-dimensional Gaussian distribution on a long needle. Consider a 10-dimensional i.i.d. multivariate normal distribution $t^{'}(x) \propto \exp\left(-x^{\top}D^{-1}x/2\right)$ where $D = diag(20, 0.0001, \cdots, 0.0001)$ and $x \in \mathbb{R}^{10}$. We sequentially rotate the marginal plans $x_1 \perp x_2, x_2 \perp x_3, \cdots$, and $x_9 \perp x_{10}$ 45° degrees. The corresponding transformations are $Q_{1,2}(45°), \cdots, Q_{9,10}(45°)$ where

$$Q_{i,j}(\eta) = I_{10} + \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cos\eta - 1 & 0 & \cdots & 0 & -\sin\eta & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sin\eta & 0 & \cdots & 0 & \cos\eta - 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

(columns i and j)

Thus, the interesting target density is

$$t(x) \propto \exp\left(-x^{\top}\left(QDQ^{\top}\right)^{-1}x/2\right), \tag{5.6}$$

where $Q = Q_{9,10}(45°)\cdots Q_{2,3}(45°)Q_{1,2}(45°)$.

We perform MwG and AM algorithms $1,000,000$ iterations where the initial point

$X_0 \sim N(\overrightarrow{0}, diag(1, \cdots, 1))$. Figure 5.2 displays the sample data on the plane $x_1 \perp x_2$, and 300-step average acceptances of both algorithms. Both do stick in the quite short stripe. One is between $(-1.5, -1)$ to $(1.8, 1.3)$ with the length around 4.02, another is between $(1.0, 0.8)$ to $(3.0, 2.2)$ with the length around 5. Their lengths of the needle are far less than 35.78. Their estimates of autocorrelation functions (ACF) also show that the sample data have strong correlations.

We performed ADSSMG $1,000,000$ iterations where the initial point has the same distribution as that of MwG and AM. Figure 5.2 shows the sample data on the plane $x_1 \perp x_2$, the 300-step average of acceptance rates and the ACFs of ADSSMG variables x_1 and x_2 generated from ADSSMG. From these graphs, ADSSMG broadly detects the target with the narrow stripe roughly between $(-12, -10)$ to $(14, 10)$ with the length around 32.8. The average acceptance rate is roughly between 0.27 and 0.42. The ACFs of x_1 and x_2 almost tends to zero.

5.2.3 Ergodicity

Assumption 5.2.1 (Target Regularity)**.** *Same as Assumption 4.3.1.*

Assumption 5.2.2 (Exponential tails for ADRSMwG)**.** *There exist $\beta > 0$, $0 \leq \delta < \Delta$ with $1/\beta \leq \delta$ such that for any sequence $\{(x_n, \gamma_n) : n \geq 0\}$ with $\lim_n |x_n| = \infty$ and $\gamma_n \in \mathcal{Y}$ ($\gamma_n = Q_{\gamma_n}^\top D_{\gamma_n} Q_{\gamma_n}$), we may extract a subsequence $\{(\tilde{x}_n, \tilde{\gamma}_n) : n \geq 0\}$ with the property that for $n \geq 0$, there exists $i_n := i(\tilde{\gamma}_n) \in \{1, \cdots, d\}$, and denote $\tilde{e}_{i_n} := Q_{\tilde{\gamma}_n}^\top e_{i_n}$. For $z \in [\delta, \Delta]$,*

$$\begin{aligned} \lim_n \frac{\pi(\tilde{x}_n)}{\pi(\tilde{x}_n - \text{sign}(< \tilde{x}_n, \tilde{e}_{i_n} >) z \tilde{e}_{i_n})} &\leq \exp(-\beta z) \text{ and} \\ \lim_n \frac{\pi(\tilde{x}_n + \text{sign}(< \tilde{x}_n, \tilde{e}_{i_n} >) z \tilde{e}_{i_n})}{\pi(\tilde{x}_n)} &\leq \exp(-\beta z); \end{aligned} \tag{5.7}$$

Assumption 5.2.3. *Under Assumption 5.2.2,*

$$l(\epsilon, \theta, \delta, \Delta) := \frac{\epsilon\theta}{\sqrt{2\pi}} \left(e^{-\delta^2/(2\epsilon)} - e^{-\Delta^2/(2\epsilon)} \right) > \frac{d+e}{\beta(e-1)}, \tag{5.8}$$

where ϵ and θ are predetermined in Equation (5.4).

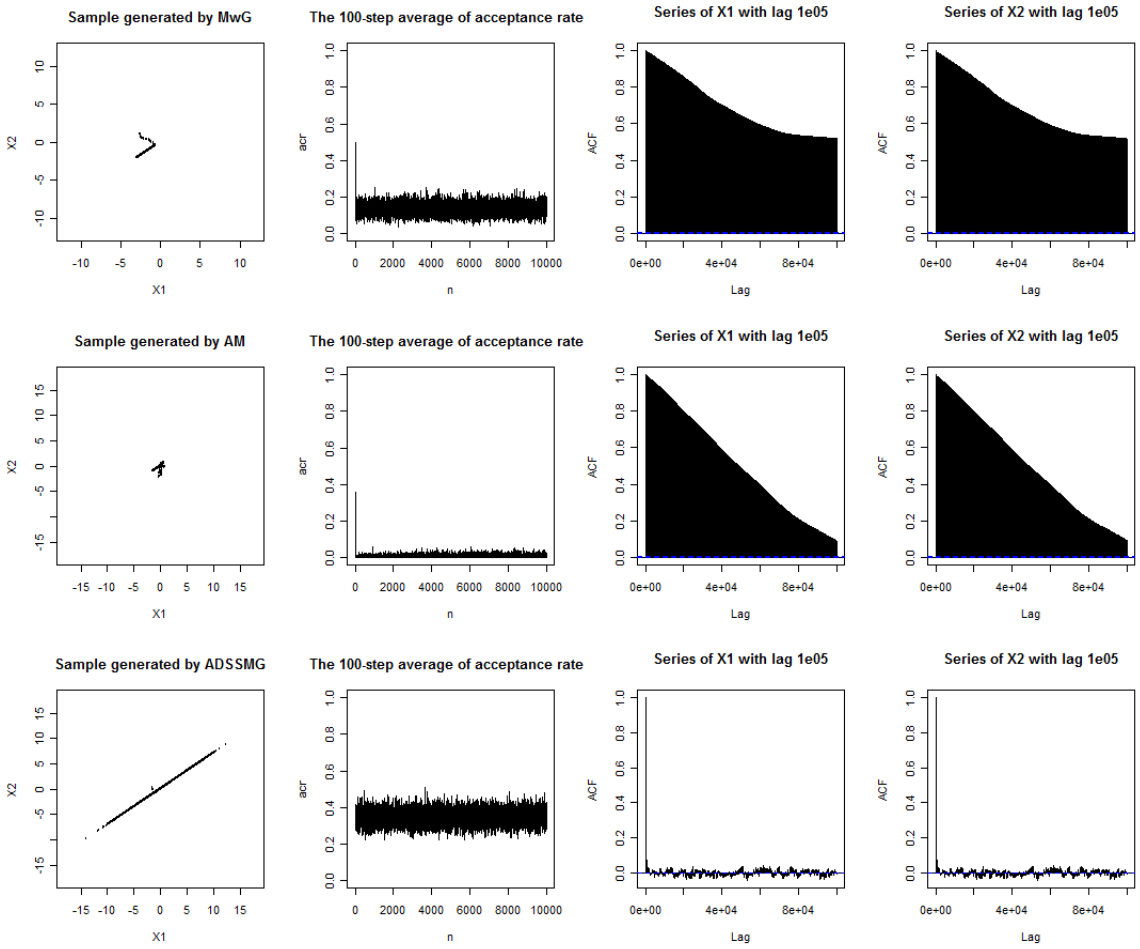

Figure 5.2: The top first plot is the sample plot by running MwG on the example of Section 5.1. The top second plot is its 300-step average of acceptance rates. The top last two plots are the ACFs of the MwG variables x_1 and x_2 with lag up to $100,000$. The center first plot is the sample plot by running AM. The center second plot is the 300-step average of acceptance rates. The center last two plots are the ACFs of the AM variables x_1 and x_2 with lag up to $100,000$. The bottom first plot is the sample plot by running ADSSMG. The bottom second plot is the 300-step average of acceptance rates. The bottom right two plots are the ACFs of the ADSSMG variables x_1 and x_2 with lag up to $100,000$.

Theorem 5.2.1. *Under Assumptions 5.2.1-5.2.3, ADRSMwG with Diminishing Adaptation is ergodic.*

Proof: Under Assumption 5.2.3, for $t \in (0, 1/2)$, let

$$h(\alpha, t) = r'(t) + \frac{1}{d(1-t)^2} - \frac{\alpha}{d(1-t)} \inf_{\gamma \in \mathcal{Y}} \inf_{i=1,\cdots,d} \int_{\delta}^{\Delta} (e^{-\alpha t z} - e^{-\alpha(1-t)z}) z q_{i,\gamma}(z) dz \text{ and}$$

$$H(\alpha, t) = 1 + \int_0^t h(\alpha, u) du.$$

So, $H(\beta, 0) = 1$ and

$$\frac{\partial H(\beta, 0)}{\partial t} = h(\beta, 0) \leq e^{-1} + \frac{1}{d} - \frac{\beta(1 - e^{-1})}{d} l(\epsilon, \theta, \delta, \Delta) < 0.$$

So there exists $s_0 \in (0, 1/2)$ such that $H(\beta, s_0) < 1$.

Assume that $\limsup_{|x| \to \infty} \sup_{\gamma \in \mathcal{Y}} P_{\text{DRS},\gamma} V_{s_0}(x)/V_{s_0}(x) \geq 1$. So there exists a sequence $\{(x_n, \gamma_n) : n \geq 0\}$ such that $\lim_n P_{\text{DRS},\gamma_n} V_{s_0}(x_n)/V_{s_0}(x_n) \geq 1$.

Under Assumption 5.2.2, there exist a subsequence $\{(\tilde{x}_n, \tilde{\gamma}_n) : n \geq 0\}$ and $\{i_n \in \{1, 2, \cdots, d\} : n \geq 0\}$, and $\{\tilde{e}_{i_n} : n \geq 0\}$ such that Equation (5.7) holds with the corresponding parameters β, δ, Δ.

Adapting the method in the proof of Theorem 4.4.1, we have

$$\begin{aligned} P_{i_n,\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) \leq & \frac{K_{i_n,\tilde{\gamma}_n}(\beta s_0)}{1 - s_0} + K_{i_n,\tilde{\gamma}_n}(0) + \frac{K_{i_n,\tilde{\gamma}_n}(\beta(1 - s_0))}{1 - s_0} - K_{i_n,\tilde{\gamma}_n}(\beta) + \\ & r(s_0)(1 - 2K_{i_n,\tilde{\gamma}_n}(0)) \end{aligned} \tag{5.9}$$

where $K_{j,\gamma}$ is defined in Equation (4.26).

Hence,

$$P_{\text{DRS},\tilde{\gamma}_n} V_{s_0}(\tilde{x}_n)/V_{s_0}(\tilde{x}_n) \leq H_{i_n,\tilde{\gamma}_n}(\beta, s_0)$$

where $H_{j,\gamma}$ is defined in Equation (4.27).

By similar arguments in the proof of Theorem 4.4.1, $H_{i_n,\tilde{\gamma}_n}(\beta, s_0) \leq H(\beta, s_0) < 1$. Contradiction! So ADRSMwG is ergodic. □

5.3 A Real-life Cohort Study with the competing risks

The Cox (1972) proportional hazards model is routinely used for failure time data. Cox (1975) studied the partial likelihood methods, also see textbook Kalbfleisch and Prentice (2002). Accordingly, Prentice (1986) proposed the Case-Cohort design to efficiently analyze Cohort data when most observations are censored, i.e. the interesting events occur with low frequency. For epidemiologic studies, the cohort may be very large under the previous assumption. Self and Prentice (1988) proved the asymptotic normal properties of the estimate $\hat{\beta}$ under certain regularity conditions by using a pseudo-likelihood. Wacholder et al. (1989) proposed a bootstrap estimate of the variance of $\hat{\beta}$. Similar estimates for the variance were derived by Lin and Ying (1993) and Barlow (1994). Pintilie et al. (2009) used a modified partial likelihood to accommodate the modeling of the hazard of subdistribution for a Case-Cohort study. They used the Jackknife method to find the estimate's covariance matrix.

These frequentist methods mainly try to find the optimal coefficient estimates of covariates such that the pseudo-likelihood reaches the maximum. Here we utilize the Bayesian method through simulating the posterior distribution of the coefficients of covariates, and compare three algorithms: MwG, AM and ADSSMG, see Table 5.2.

The following specification of Cox Regression Model is from Kalbfleisch and Prentice (2002). The extensive Cox Regression Model is from Pintilie et al. (2009).

5.3.1 The Model Description

Cox Regression Model (The relative risk model) is a semiparametric model. It has a nonparametric aspect in the sense that it involves an unspecified function in the form of an arbitrary baseline hazard function. It also incorporates a parametric modeling of the relationship between the failure rate and specified covariates.

The incorporate covariates are prefixed and independent of time, or are defined functions of time.

Let $x = (x_1, x_2, \cdots, x_K)'$ be a vector of fixed covariates that are measured at or before time 0 on individuals under study. *The relative risk models* or *Cox models* are

specified by the hazard relationship

$$\begin{aligned}\lambda(t;x) = &\lim_{h\to 0^+} \mathbf{P}(t \le T < t+h \mid T \ge t, x)/h \\ =&\lambda_0(t) r(t,x), \qquad t > 0\end{aligned} \tag{5.10}$$

where T is a corresponding absolutely continuous failure time variate, $\lambda_0(t)$ is an unspecified baseline hazard function, and the relative risk function $r(t,x)$ specifies the relationship between the covariates x and the failure rate or hazard function. We consider the usual exponential form for the relative risk function $r(t,x) = \exp(Z(t)'\beta)$, which yields the model

$$\lambda(t,x) = \lambda_0(t) \exp(Z^{(x)}(t)'\beta). \tag{5.11}$$

where $Z^{(x)}(t) = \left(Z_1^{(x)}(t), \cdots, Z_p^{(x)}(t)\right)'$ is a vector of derived, possibly time-dependent covariates obtained as functions of t and the fixed covariates x. The baseline hazard function $\lambda_0(t)$ corresponds to $Z^{(x)}(t) = (0, \cdots, 0)'$ for all t, and $\beta = (\beta_1, \cdots, \beta_p)'$ is a vector of (unknown) regression parameters.

If the failure time T has the hazard function Equation (5.11), the corresponding survivor function is

$$F(t;x) = \mathbf{P}(T > t \mid x) = \exp\left(-\int_0^t \lambda_0(u) \exp(Z^{(x)}(u)'\beta) du\right) \tag{5.12}$$

and the density function is

$$f(t;x) = \lambda(t;x) F(t;x). \tag{5.13}$$

Estimation of β

Suppose that the data consist of observations on a random vector Y with the density function $f(y;\theta,\beta)$ where β is the parameter of interest and θ is the nuisance parameter of high or infinite dimension. Suppose that Y can be transformed into $A_1, B_1, \cdots, A_m, B_m$ and $A^{(j)} = (A_1, \cdots, A_j)$ and $B^{(j)} = (B_1, \cdots, B_j)$. Assume that

the joint density function of $A^{(m)}$ and $B^{(m)}$ can be written as

$$\prod_{j=1}^{n} f(b_j \mid b^{(j-1)}, a^{(j-1)}, \theta, \beta) \prod_{j=1}^{n} f(a_j \mid b^{(j)}, a^{(j-1)}, \beta). \tag{5.14}$$

The second term is called the *partial likelihood* of β based on $\{A_j\}$ in the sequence $\{A_j, B_j\}$. One may argue that any information on β in the first term is inextricably tied up with information on the nuisance parameters θ.

suppose that the sample consists of k uncensored failure times[1] $t_1 < \cdots < t_k$ and ignore for the moment the case of ties. The remaining $n - k$ individuals are right censored[2]. Let j denote the individual failing at t_j. Let B_j specify the censoring information in $[t_{j-1}, t_j)$ plus the information that one individual fails in the interval $[t_j, t_j + dt_j)$. Let A_j specify that the item j fails in $[t_j, t_j + dt_j)$. The jth item in the partial likelihood in Equation (5.14) is

$$L_j(\beta) = f(a_j \mid b^{(j)} a^{(j-1)}, \beta). \tag{5.15}$$

Note that the conditioning event $b^{(j)}, a^{(j-1)}$ specifies all the censoring and failure information in the trial up to time t_j^- and also provides the information that a failure occurs in $[t_j, t_j + dt_j)$. Under independent censoring, it follows that

$$L_{\,j}(\beta) = \frac{\lambda(t_j, x_j) dt_j}{\sum_{l=1}^{n} Y_l(t_j) \lambda(t_j, x_l) dt_j}, \tag{5.16}$$

where $Y_l(t)$ indicates that item l is in the set $R(t)$ of items at risk of failure at time t^-, just prior to time t. Under the relative risk model Equation (5.11), Equation (5.15) simplified since the baseline hazard term cancels in the numerator and denominator. The product over j then provides the partial likelihood for β,

$$L(\beta) = \prod_{j=1}^{k} \frac{\lambda(t_j, x_j) dt_j}{\sum_{l=1}^{n} Y_l(t_j) \lambda(t_j, x_l) dt_j} \tag{5.17}$$

[1]Time censoring: the censored survival times were observed only if failure had not occurred prior to a predetermined time at which the study was to terminated. Order statistics censoring: the study terminates as soon as certain order statistics are observed.

[2]The data on these individuals who do not fail during their observation period, is called *right censored*.

Censoring variables

Suppose that there is a set of ordered pair times $(t_1^*, t_1), \cdots, (t_n^*, t_n)$ where t_j^*s are the entry times ($< t_j$) and t_js ($t_1 < \cdots < t_n$) are the event observing times. The corresponding censor variables are defined as

$$C_j = \begin{cases} 1 & \text{when the event of interest was observed} \\ 2 & \text{when the competing risk event was observed} \\ 0 & \text{when no event was observed} \end{cases}. \tag{5.18}$$

The set

$$R(t) = \{i : t_i^* \leq t \leq t_i; C_i = 0 \text{ or } 1\} \cup \{i : t_i^* \leq t; C_i = 2\}, \tag{5.19}$$

is the set of items at risk of failure at time t^-, just prior to time t.

The event $C_j = 1$ is the event of interest (failure happens). The event $C_j = 2$ is the uncensored event with competing risks. The event $C_j = 0$ is the right censored event.

By Equation (5.11) and Equation (5.17), the *modified partial likelihood function* at the time of occurrence and the competing risks events with a specific weight for the Case-Cohort study is

$$L^*(\beta; x) = \prod_{j=1}^{n} \frac{\mathbb{I}(C_j = 1) \exp(\beta^\top x_j)}{\sum_{r \in R(t_j)} w_{rj} \exp(\beta^\top x_r)}, \tag{5.20}$$

where the weights $w_{rj} = \frac{\hat{G}(t_j)}{\hat{G}(t_j \wedge t_r)}$, and $\hat{G}(t_j)$ is the Kaplan-Meier estimator for the probability of censoring, see Kaplan and Meier (1958). The set $R(t)$ represents the case and time-matched controls at the Cohort follow-up time t. The covariates x_i can be time-dependent on t_i.

Here, we choose a prior $\mu(\cdot)$ for the coefficient β. The target distribution (the posterior distribution) that we want to simulate is

$$t(\beta) \propto \mu(\beta) L^*(\beta; x). \tag{5.21}$$

Table 5.1: Hypoxia study: 10 records are extracted from dataset

age	hgb	tumsize	IFP	HP_5	pelvicln	resp	pelrec	disrec	survtime	stat	dftime
78	119	7	8	32.1428571	N	CR	N	N	6.152	0	6.152
69	131	2	8.2	2.173913	N	CR	N	N	8.008	0	8.008
55	126	10	8.6	52.3255814	N	NR	Y	N	0.621	1	0.003
55	141	8	3.3	3.2608696	N	CR	Y	Y	1.12	1	1.073
50	95	8	18.5	85.4304636	Y	NR	Y	N	1.292	1	0.003
57	132	8	20	19.3548387	N	CR	N	N	7.929	0	7.929
53	127	4	21.8	44.5783133	E	CR	N	N	8.454	0	8.454
62	142	5	31.6	59.6774194	N	CR	Y	Y	7.116	0	7.107
23	145	5	16.5	29.1666667	N	CR	N	N	8.378	0	8.378
57	142	3	31.5	85.7142857	N	CR	N	N	8.178	0	8.178
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

hgb	Haemoglobin (g/l)
pelvicln	Pelvic node involvment: N=Negative, E=Equivoval,Y=Positive
pelrec	Pelvic disease observed: Y=Yes, N=No
disrec	Distant disease observed: Y=Yes, N=No
stat	Status at last follow-up: 0=Alive, 1=Dead

5.3.2 The analysis of Hypoxia Study

In the study, 109 patients with cervical cancer were treated at a cancer center between the year 1994 to 2000. Meanwhile two cancer marker were done in the time of diagnosis: a hypoxia marker (HP_5: percentage of measurements less that 5 mmHg) and the interstitial fluid pressure (IFP). IFP are measured at a number of locations in the tumor and a mean value per patients was calculated. There are totally six diagnosis variables (age, hgb, tumorsize, IFP, HP_5, pelvicln) and five outcome variables (resp, pelrec, disrec, survtime, stat), see Table 5.1. The outcome variables include the information of the treatment, relapse and death. The response to treatment has two cases: complete response (CR) when the tumor has completely disappeared after treatment, and no response (NR) when either the disease has progressed to other sites or the tumor has not disappeared. Under the situation that resp is NR, if disease progressed to other sites then disrec=Y; if the tumor still is present then pelrec=Y, see other analysis about this case in textbook Pintilie (2006).

Consider the modified partial likelihood Equation (5.20). Here the number n of observations is 109. We use all the diagnosis variables as the covariates so the β is defined on $\mathbb{R}^6$ where the components are sequentially age, hgb, tumsize, IFP, HP_5 and pelvicln. All the entry times t_i^*s are zero, and the failure times t_js are from the variable dftime. We use the outcome variables to define the censor variables C_j for competing risks,

$$C_j = \mathbb{I}(pelrec_j = Y) + 2\mathbb{I}(pelrec_j = N, stat_j = 1), \tag{5.22}$$

Table 5.2: The settings of MwG, AM, and ADSSMG

	Initial point	Burn-in time	# iterations	proposal	other
MwG	$N(0, I_6)$	0	$1,000,000$	$N(0, 0.1)$	
AM	$N(0, I_6)$	0	$1,000,000$	see Example 4.3.1	$\theta = 0.3$
ADSSMG	$N(0, I_6)$	0	$1,000,000$	normal distribution	

Table 5.3: The coefficient estimates by CRR, MwG, AM and ADSSMG

	β_{age}	β_{hgb}	β_{tumsize}	β_{ifp}	β_{hp5}	$\beta_{\text{pelv.}}$
CRR	-0.025950	-0.013330	0.258900	0.031370	0.001198	0.497400
AM	-0.026309	-0.014401	0.245710	0.031485	0.001299	0.513099
MwG	-0.026543	-0.013669	0.257617	0.031522	0.001398	0.506934
ADSSMG	-0.026521	-0.013658	0.256224	0.031679	0.001285	0.510447

which means that the competing risk here is defined as that patients are dead and the tumors has disappeared.

Here we apply the MwG, AM and ADSSMG to sample the data for the posterior distribution $t(\cdot)$ in Equation (5.21). We compare the estimates generated by three algorithms with the R package cmprsk - CRR. Table 5.3 shows the coefficients estimate generated by CRR, AM, MwG, and ADSSMG. The three algorithms present very well. From Table 5.4, the standard errors of the coefficients generated by CRR and ADSSMG which show that the two groups of data are roughly same.

From Figure 5.3, the 100-step average of acceptance rates by AM is smallest, that by MwG stays in the middle, and that by ADSSMG is highest roughly staying in 0.4.

Figure 5.4 presents the histograms of the sample marginal densities of HP_5 and IPF where the densities by these three algorithms. The R function "hist" is called with the parameters: breaks= 172 for IPF and breaks= 400 for HP_5; color is yellow and the border color is red. Here we only show the truncated histograms (HP_5 is in $[-0.1, 0.1]$, IPF is in $[-0.04, 0.04]$) because there are few sample data on the rest of

Table 5.4: The estimates of standard errors by CRR and ADSSMG

	β_{age}	β_{hgb}	β_{tumsize}	β_{ifp}	β_{hp5}	$\beta_{\text{pelv.}}$
CRR	0.01564	0.01201	0.10690	0.01705	0.00633	0.33520
ADSSMG	0.01522	0.01298	0.10591	0.01982	0.00704	0.30021

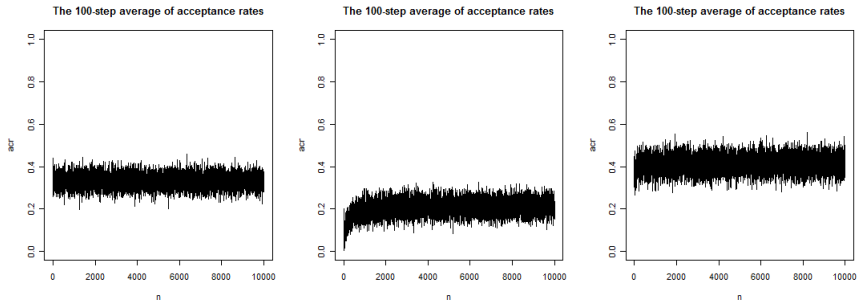

Figure 5.3: The left plot is the 100-step average of acceptance rates generated by MwG; the center plot is the 100-step average of acceptance rates generated by AM; the right plot is the 100-step average of acceptance rates generated by ADSSMG.

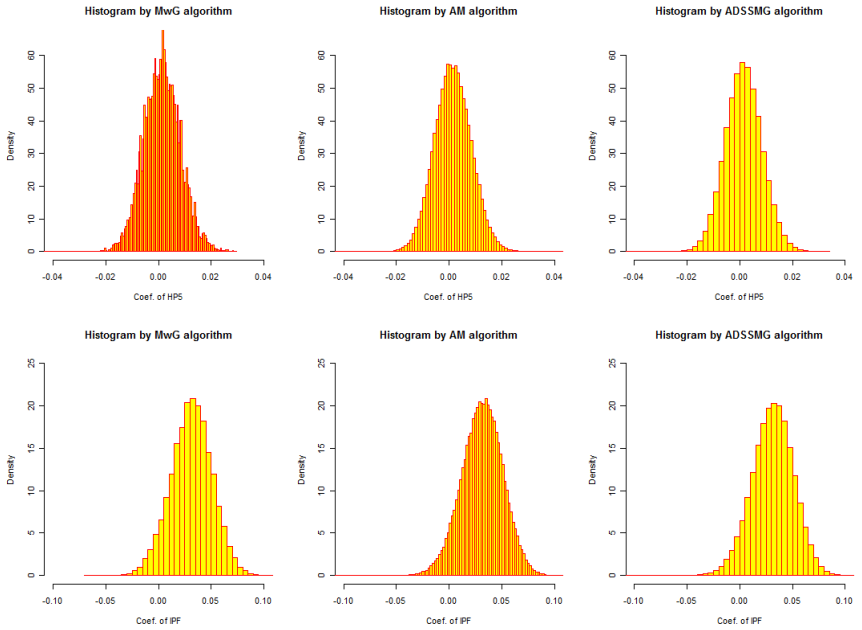

Figure 5.4: The top left is the histogram of HP_5 by MwG; the top center is the histogram of HP_5 by AM; the middle right is the histogram of HP_5 by ADSSMG; the bottom left is the histogram of IPF by MwG; the bottom center is the histogram of IPF by AM; the bottom right is the histogram of IPF by ADSSMG.

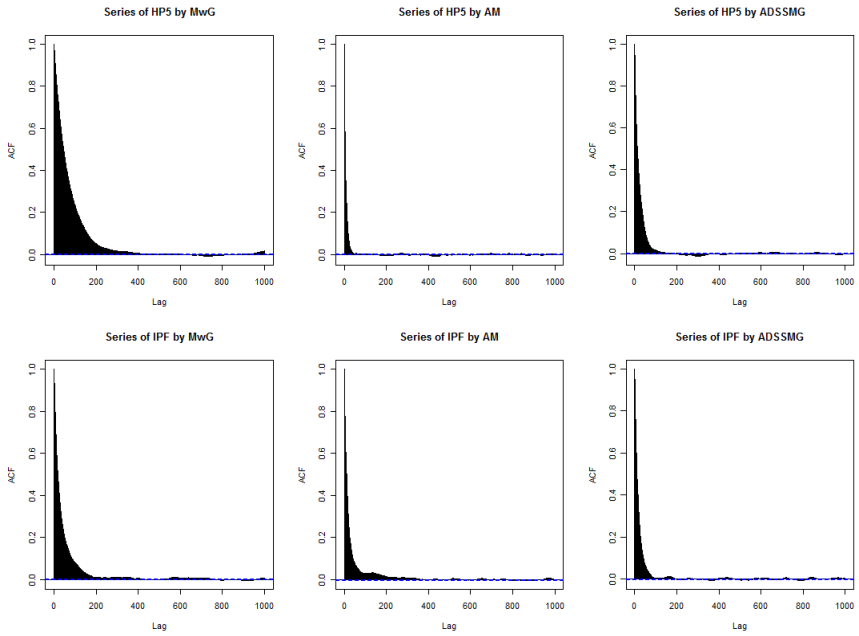

Figure 5.5: The top left is the ACF of HP_5 by MwG; the top center is the ACF of HP_5 by AM; the middle right is the ACF of HP_5 by ADSSMG; the bottom left is the ACF of IPF by MwG; the bottom center is the ACF of IPF by AM; the bottom right is the histogram of IPF by ADSSMG.

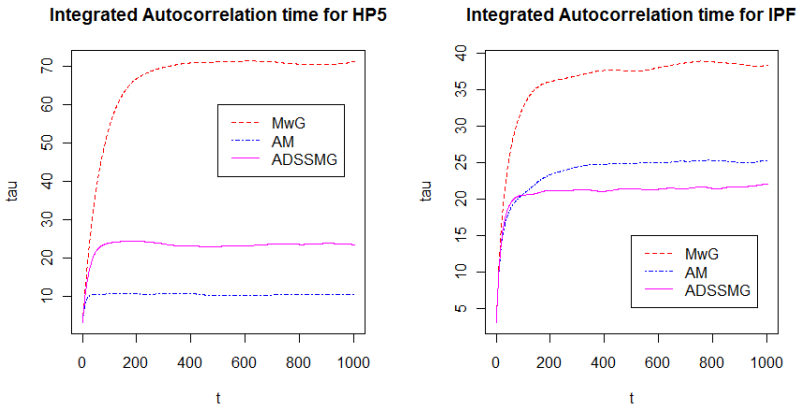

Figure 5.6: The left is the integrated autocorrelation time of HP_5 by MwG, AM and ADSSMG; the right is the integrated autocorrelation time of IPF by MwG, AM and ADSSMG.

the range. If the histogram shows more red borders, it means that more data are concentrated in the interval (it has relatively small estimate of covariance). For HP_5, ADSSMG gets relatively larger estimate of covariance. For IPF, AM gets relatively smaller estimate of covariance.

From Figure 5.5, the sample autocorrelations generated by AM and ADSSMG get close to zero earlier than MwG. The point also can be found from integrated autocorrelation time (IAT) in Figure 5.6. For HP_5, IAT from ADSSMG is smallest and roughly equal to 10. For IPF, IAT from AM is smallest and roughly equal to 20.

Although MwG, AM and ADSSMG perform well for this Cohort study, the integrated autocorrelation times of AM and ADSSMG are much smaller than MwG.

Chapter 6

Conclusions

In the thesis, first we study some relationships among Containment and Diminishing Adaptation and ergodicity of adaptive MCMC through some examples and some theoretical results. 1. Containment and Diminishing Adaptation imply ergodicity see (Roberts and Rosenthal, 2007, Theorem 13). 2. Diminishing Adaptation alone may not guarantee ergodicity, see Examples 2.1.1 and 2.2.1. Example 2.2.1 is more interesting, because there are only two transition kernels in the collection of transition kernels. 3. Containment alone may not guarantee ergodicity, see Roberts and Rosenthal (2007, Example 4). 4. Neither Diminishing Adaptation nor Containment is necessary for ergodicity, see Proposition 2.1.2. 5. Under certain additional condition, Containment is necessary, see Theorem 3.2.1 and Corollary 3.2.1. 6. For adaptive Metropolis algorithms, using some standard statistics as the adaptation of adaptive MCMC, Diminishing Adaptation can be implied by some simple conditions, see Proposition 4.3.2.

Second we study simultaneous polynomial ergodicity. For most cases of S.P.E., Containment is implied, because the boundedness of the process $\{V(X_n) : n \geq 0\}$ for some test function V can be shown. Simultaneous geometric ergodicity which is a special case of S.P.E., is also studied through considering the quantitative bound given by Rosenthal (1995).

Third we give some simple and easy-to-check conditions for adaptive Metropolis and adaptive Metropolis-within-Gibbs algorithms. The proposals are not necessarily

restricted in the family of Gaussian distributions. For targets with lighter-than-exponential tails, uniform local positivity just is required for proposals. For targets with exponential tails, the condition that the uniform first moment of proposals has some specific lower bound is required besides uniform local positivity. For targets with hyperbolic tails, the uniform local compact support condition for proposals is required besides the above two conditions.

Finally, we propose an adaptive directional Metropolis-within-Gibbs algorithm, and compare it with Metropolis-within-Gibbs sampler. We conclude that For targets with high correlations, adaptive directional Metropolis-within-Gibbs algorithms perform better than Metropolis-within-Gibbs sampler from simulation results.

Appendix A

Appendix A Markov Chain

The following notations and fundamental results are mainly drawn from textbook Meyn and Tweedie (1993) and Roberts and Rosenthal (2004).

A.1 Definition

Consider the state space $\mathcal{X}$ and the σ-field $\mathcal{B}(\mathcal{X})$. The function $P(\cdot,\cdot) : \mathcal{X}\times\mathcal{B}(\mathcal{X}) \to \mathbb{R}$ is called to be a *transition kernel* if

(i) For each $x \in \mathcal{X}$, $A \to P(x, A)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$;

(ii) For each $A \in \mathcal{B}(\mathcal{X})$, $x \to P(x, A)$ is a measurable function.

The process $\mathbf{X} = \{X_n : n \in \mathbb{Z}_+\}$ ($\mathbb{Z}_+ := \{0, 1, 2, \cdots\}$) is called *a discrete time homogeneous Markov Chain* with respect to a filtration $\mathcal{F}_n := \sigma(X_0, \cdots, X_n)$ if the property

$$P(X_n \in A \mid \mathcal{F}_{n-1}) = P(X_n \in A \mid X_{n-1}) := P(X_{n-1}, A), \quad A \in \mathcal{B}(\mathcal{X}) \tag{A.1}$$

is satisfied. Denote the n-step transition kernel by $P^n(x, A) := P(X_n \in A \mid X_0 = x)$.

Theorem A.1.1 (Chapman-Kolmogorov equation)**.** *For any m with $0 \le m \le n$,*

$$P^n(x, A) = \int_{\mathcal{X}} P^m(x, dy)P^{n-m}(y, A), \qquad x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X}). \tag{A.2}$$

We regard values of the whole chain $\mathbf{X}$ as lying in the sample path space $\Omega = \mathcal{X}^\infty = \prod_{i=0}^\infty \mathcal{X}_i$, where $\mathcal{X}_i$ is a copy of $\mathcal{X}$. The $\mathbf{X}$ is a random variable on Ω equipped with a σ-field $\mathcal{F}$. Denote the probability measure of the chain $\mathbf{X}$ starting with the initial distribution μ by $\mathbf{P}_\mu$ ($\mathbf{P}_x := \mathbf{P}_{\delta_x}$). The triple $(\Omega, \mathcal{F}, \mathbf{P}_\mu)$ defines a Markovian Chain with the initial distribution μ.

The state space $\mathcal{X}$ is called *discrete* if $\mathcal{X}$ has a finite or countable number of elements. The $\mathcal{X}$ is called *general* if it is equipped with a countably generated σ-field $\mathcal{B}(\mathcal{X})$. The state space $\mathcal{X}$ is called *topological* if it is equipped with a locally compact separable metrizable topology with $\mathcal{B}(\mathcal{X})$ as the Borel σ-field. In the thesis we only consider the discrete or general state space.

We denote *the first entry time* and *the hitting time* by $\tau_A := \min\{n > 0 : X_n \in A\}$ and $\sigma_A := \min\{n \geq 0 : X_n \in A\}$ respectively.

Let $\{a(n)\}$ be a distribution on $\mathbb{Z}_+$, consider the *sample chain* $\mathbf{X}_a$ with transition kernel $K_a(x, A) := \sum_{n=0}^\infty P^n(x, A)a(n)$, for $x \in \mathcal{X}$, $A \in \mathcal{B}(\mathcal{X})$, see the properties of sample chain in textbook Meyn and Tweedie (1993).

A.2 Irreducibility and Aperiodicity

Definition A.2.1 (Irreducible)**.** *If the state space is discrete, irreducibility means that for all $x, y \in \mathcal{X}$, y is accessible from x, i.e. there exists $n \in \mathbb{N}$ such that $P^n(x, \{y\}) > 0$. If the state space $\mathcal{X}$ is a general state space, we call the chain $\mathbf{X}$ φ-irreducible if there exists a measure φ on $\mathcal{B}(\mathcal{X})$ such that, whenever $\varphi(A) > 0$, we have $P_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$.*

If $\mathbf{X}$ is φ-irreducible for some measure φ, then there exists a probability measure ψ (*maximal irreducibility measure*) on $\mathcal{B}(\mathcal{X})$ such that

(i) $\mathbf{X}$ is ψ-irreducible;

(ii) for any other measure φ', $\mathbf{X}$ is φ'-irreducible if and only if φ' is absolutely continuous with respect to ψ;

(iii) if $\psi(A) = 0$ then $\psi\{y : P_x(\tau_A < \infty) > 0\} = 0$;

(iv) ψ is equivalent to $\int_{\mathcal{X}} \varphi'(dy)K_{a_{1/2}}(y, \cdot)$ for any finite irreducibility measure φ' where $K_{a_{1/2}}(x, A) := \sum_{n=0}^\infty P^n(x, A)2^{-(n+1)}$.

Let ψ-irreducibility represent that ψ is a maximal irreducibility measure. We write $\mathcal{B}^+(\mathcal{X}) := \{A \in \mathcal{B}(\mathcal{X}) : \psi(A) > 0\}$ for the sets of positive ψ-measure. A set A is called *full* if $\psi(A^c) = 0$ and *absorbing* if $P(x, A) = 1$ for $x \in A$.

Another property of maximal irreducibility is that for any ψ-null set, its complement set can be decomposed into two sets N and A_0 where N is ψ-null and $P(x, A_0) \equiv 1$ for any $x \in A_0$.

Definition A.2.2 (Petite set and Small set)**.** *We call a set $C \in \mathcal{B}(\mathcal{X})$ ν_a-petite if the sample chain satisfies*

$$K_a(x, B) \geq \nu_a(B), \tag{A.3}$$

for all $x \in C$, $B \in \mathcal{B}(\mathcal{X})$, where ν_a is a non-trivial measure on $\mathcal{B}(\mathcal{X})$. If the sample distribution $a(\cdot)$ is a point mass distribution on some $m \in \mathbb{Z}_+$, the ν_a-petite set is called ν_m-small set.

Definition A.2.3 (Aperiodic)**.** *Suppose that $\mathbf{X}$ is a φ-irreducible Markov Chain. The largest d for which a d-circle[1] occurs for $\mathbf{X}$ is called the period of $\mathbf{X}$. When $d = 1$, the chain $\mathbf{X}$ is called aperiodic. When there exists a ν_1-small set A with $\nu_1(A) > 0$, the chain $\mathbf{X}$ is called strongly aperiodic.*

Now we introduce the theorem which presents the relationship between petite set and small set.

Theorem A.2.1. *If $\mathbf{X}$ is irreducible and aperiodic then every petite set is small.*

A.3 Recurrence and Transience

In terms of the *occupation time* $\eta_A := \sum_{i=1}^{\infty} \mathbb{I}(X_i \in A)$, we study the transience and recurrence of the subset $A \subset \mathcal{X}$. A is called *uniformly transient* if there exists $M > 0$ such that for all $x \in A$, $E_x[\eta_A] < M$. A is called *recurrent* if $E_x[\eta_A] = \infty$ for all $x \in A$. A is called *transient* if it can be covered with a countable number of uniformly transient sets.

Definition A.3.1 (Recurrence and Transience)**.** *The chain $\mathbf{X}$ is called recurrent if it is ψ-irreducible and $E_x[\eta_A] = \infty$ for every $x \in \mathcal{X}$ and every $A \in \mathcal{B}^+(\mathcal{X})$. The chain*

[1] there exist disjoint sets $D_1, \cdots, D_d \in \mathcal{B}(\mathcal{X})$ such that (i) for $x \in D_i, P(x, D_{i+1}) = 1, i = 0, \cdots, d-1 \pmod d$; (ii) the set $N = [\cup_{i=1}^{d} D_i]^c$ is ψ null.

X *is called transient if it is ψ-irreducible and $\mathcal{X}$ is transient.*

The set A is called *Harris recurrent* if the probability of the occupation time equal to infinity is 1 starting from any point in A, i.e. $P_{\delta_x}(\eta_A = \infty) = 1$ for $x \in A$. The chain **X** is called *Harris recurrent* if it is ψ-irreducible and every set in $\mathcal{B}^+(\mathcal{X})$ is Harris recurrent.

Obviously, every Harris recurrent set is recurrent. Furthermore, if the chain **X** is recurrent then the state space $\mathcal{X}$ can be decomposed into two parts: a set N and a non-empty set H, where N is ψ-null transient and any subset of H in $\mathcal{B}^+(\mathcal{X})$ is Harris recurrent and $P(x, H) = 1$ for $x \in H$.

Theorem A.3.1. *Suppose that* **X** *is ψ-irreducible. Then either*
(i) *every set in $\mathcal{B}^+(\mathcal{X})$ is recurrent, in which case we call the chain* **X** *is recurrent; or*
(ii) *the chain* **X** *is transient.*

The *drift operator* Δ is defined for any nonnegative measurable function V by $\Delta V(x) = PV(x) - V(x)$ for $x \in \mathcal{X}$. The following theorem presents the relationship between the drift operator, and recurrence and transience.

Theorem A.3.2. *Suppose that* **X** *is ψ-irreducible.*
(i) *The chain* **X** *is transient if and only if there is a bounded non-negative function V and a set $C \in \mathcal{B}^+(\mathcal{X})$ such that $\Delta V(x) \geq 0$ for $x \in C^c$ and $\{x : V(x) > \sup_{y \in C} V(y)\} \in \mathcal{B}^+(\mathcal{X})$;*
(ii) *The chain* **X** *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 $\Delta V(x) \leq 0$ for $x \in C^c$.*

A.4 Coupling Method and Aperiodic Ergodic Theorem

For any two initial measures μ_1 and μ_2, define two process $X_n \sim P^n_{\mu_1}(\cdot)$ and $Y_n \sim P^n_{\mu_2}(\cdot)$ with the same transition kernel $P(x, \cdot)$. Let the *coupling time* T be the minimal random time of $X_n = Y_n$, i.e. $T = \min\{n \geq 0 : X_n = Y_n\}$. Then the chain **Z** where $Z_n = X_n \mathbb{I}(n < T) + Y_n \mathbb{I}(n \geq T)$ has the same distribution as **X**. The *coupling*

inequality

$$\|P(X_n \in \cdot) - P(Y_n \in \cdot)\| = \|P(Z_n \in \cdot) - P(Y_n \in \cdot)\| \leq P(T > n). \tag{A.4}$$

is satisfied where the *total variation norm*

$$\|\nu(\cdot)\| := \sup_{A \in \mathcal{B}(\mathcal{X})} |\nu(A)|\,. \tag{A.5}$$

A σ-finite measure π on $\mathcal{B}(\mathcal{X})$ with the property $\pi(A) = \int_{\mathcal{X}} \pi(dx)P(x, A)$ for $A \in \mathcal{B}(\mathcal{X})$ is called *invariant*. π is called *stationary* if it is an invariant probability measure.

Using the coupling method and the splitting technique (see textbook Meyn and Tweedie (1993)), the following theorem can be shown.

Theorem A.4.1 (Aperiodic Ergodic Theorem)**.** *Suppose the chain* $\mathbf{X}$ *is aperiodic and Harris recurrent with invariant measure* π*. The following are equivalent:*
(i) *the unique invariant measure* π *is finite;*
(ii) *there is a petite set* $C \in \mathcal{B}(\mathcal{X})$ *such that* $\sup_{x \in C} E_x[\tau_C] < \infty$*;*
(iii) *there exists some petite set* C*, some* $b < \infty$ *and a non-negative function* V *finite at some* $x_0 \in \mathcal{X}$*, satisfying*

$$\Delta V(x) \leq -1 + b\mathbb{I}_C(x), \qquad x \in \mathcal{X}. \tag{A.6}$$

Any of these condition is equivalent to the existence of a unique invariant probability measure π *such that for every initial condition* $x \in \mathcal{X}$*,* $\|P^n(x, \cdot) - \pi(\cdot)\| \to 0$*.*

When the state space $\mathcal{X}$ is discrete, the irreducibility of the chain $\mathbf{X}$ implies positive recurrent. So, if the chain $\mathbf{X}$ on the discrete space $\mathcal{X}$ is irreducible and aperiodic then $\mathbf{X}$ is ergodic.

A.5 Geometric Ergodicity and Polynomial Ergodicity

A Markov Chain satisfies a *geometric drift condition* if there are constant $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \longrightarrow [1, \infty]$, such that,

$$PV \leq \lambda V + b\mathbb{I}_C, \tag{A.7}$$

for some $C \in \mathcal{B}^+(\mathcal{X})$.

Definition A.5.1 (Geometric Ergodicity)**.** *A Markov Chain with stationary distribution π is geometric ergodic if*

$$\|P^n(x, \cdot) - \pi(\cdot)\| \leq M(x)\rho^n, \qquad n = 1, 2, 3, \cdots \tag{A.8}$$

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

If the function $M(\cdot)$ in the above definition is a constant then the chain is called *uniformly ergodic*, see its properties in Roberts and Rosenthal (2004).

The following theorem shows the criterion of geometric ergodicity, see Roberts and Rosenthal (2004).

Theorem A.5.1. *Consider a ψ-irreducible aperiodic Markov Chain $\mathbf{X}$ with stationary distribution $\pi(\cdot)$. Suppose that $C \in \mathcal{B}^+(\mathcal{X})$ is a small set. Suppose further that the geometric drift condition is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \longrightarrow [1, \infty)$ with $V(x) < \infty$ for at least one $x \in \mathcal{X}$. Then the chain is geometrically ergodic.*

To study polynomial ergodicity, Fort and Moulines (2003) developed the following result, see also Jarner and Roberts (2002).

Let $f : \mathcal{X} \to [1, \infty)$ be a Borel function, q be a positive integer and a non-empty set $C \in \mathcal{B}(\mathcal{X})$.

P1: There exist some measurable functions on $\mathcal{X}$, $1 \leq f =: V_0 \leq \cdots \leq V_q$, and some finite constants b_k, $k \in \{0, \ldots, q-1\}$, such that $\sup_{x \in C} V_q(x) < \infty$ and for all $k \in \{0, \ldots, q-1\}$,

$$PV_{k+1} - V_{k+1} \leq -V_k + b_k\mathbb{I}_C. \tag{A.9}$$

Theorem A.5.2. *Let q be a positive integer, f be a Borel function and $C \in \mathcal{B}(\mathcal{X})$ be a non-empty petite set. Suppose that P is ψ-irreducible and aperiodic and that* **P1** *holds. Then P possesses an unique invariant probability measure π such that $\pi(f) < \infty$ and for all x in the full and absorbing set $\{V_q < \infty\}$,*

$$\lim_n (n+1)^{q-1} \left\| P^n(x, \cdot) - \pi(\cdot) \right\|_f = 0, \tag{A.10}$$

where $\|\mu\|_f = \sup_{|g| \leq f} |\mu(g)|$.

Bibliography

H.C. Andersen and P. Diaconis. Hit and Run as a Unifying Device. *Journal de la Société Française de Statistique*, 148(5):5–28, 2007.

C Andrieu and E Moulines. On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms. . *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

C. Andrieu and C.P. Robert. Controlled MCMC for optimal sampling. . *Preprint*, 2001.

Y.F. Atchadé and G. Fort. Limit Theorems for some adaptive MCMC algorithms with subgeometric kernels. *Preprint*, 2008.

Y.F. Atchadé and J.S. Rosenthal. On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli*, 11(5):815–828, 2005.

Y. Bai. Simultaneous drift conditions on adaptive Markov Chain Monte Carlo algorithms. *Technical Report in Department of Statistics at the University of Toronto*, 2009a.

Y. Bai. An Adaptive Directional Metropolis-within-Gibbs algorithm. *Technical Report in Department of Statistics at the University of Toronto*, 2009b.

Y. Bai, G.O. Roberts, and J.S. Rosenthal. On the Containment condition of adaptive Markov Chain Monte Carlo algorithms. *Technical Report in Department of Statistics at the University of Toronto*, 2008.

W.E. Barlow. Robust variance estimation for the case-cohort desing. *Biometrics*, 50: 1064–1072, 1994.

M. Bèdard and D.A.S. Fraser. On a Directionally Adjusted Metropolis-Hastings Algorithm. *Preprint*, 2008.

C.J.P. Bélisle, H.E. Romeijn, and R.L. Smith. Hit-and-Run Algorithms for generating Multivariate distributions. *Math. of Operation. Research*, 18(2), 1993.

A.E. Brockwell and J.B. Kadane. Indentification of regeneration times in MCMC simulation, with application to adaptive schemes. *J. Comp. Graph. Stat*, 14:436–458, 2005.

M.H. Chen and B.W. Schmeiser. Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers. *J. Comp. and Graph. Stats.*, 2(3):251–272, 1993.

M.H. Chen and B.W. Schmeiser. General Hit-and-Run Monte Carlo smapling for evaluating multidimensional integrals. *Operation Research Letter*, 19:161–169, 1996.

D.R. Cox. Regression models and life tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972.

D.R. Cox. Partial likelihood. *Biometrika*, 62:269–272, 1975.

R.V. Craiu, J.S. Rosenthal, and C. Yang. Learning from Thy Neighbor: Parallel-Chain Adaptive MCMC. *Preprint*, 2008.

G. Fort and E. Moulines. V-Subgeometric ergodicity for a Hastings-Metropolis algorithm. *Statist. Prob. Lett.*, 49:401–410, 2000a.

G. Fort and E. Moulines. Computable Bounds for Subgeometrical and geometrical Ergodicity. 2000b.

G. Fort and E. Moulines. Polynomial ergodicity of Markov transition kernels. *Stoch. Process. Appl.*, 103:57–99, 2003.

G. Fort, E. Moulines, G.O. Roberts, and J.S. Rosenthal. On the geometric ergodicity of hybrid samplers. *J. Appl. Prob.*, 40:123–146, 2003.

A. Gelfand and A. Smith. Sampling based approaches to calculating marginal densities. *J. Amer. Stats. Assoc.*, 85:398–409, 1990.

A. Gelman, G.O. Roberts, and W.R. Gilks. Efficient Metropolis jumping rules. *Bayesian Statistics, 5(Alicate, 1994), Oxford Sci. Publ.,*, pages 599–607, 1996.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of image. *IEEE Trans. Pattn. Anal. Mach. Intel.*, 6:721–741, 1984.

W.R. Gilks, G.O. Roberts, and E.I. George. Adaptive direction sampling. *The statistician*, 43:179–189, 1994.

W.R. Gilks, G.O. Roberts, and S.K. Sahu. Adaptive Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 93:1045–1054, 1998.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

S.F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stoch. Process. Appl.*, 85:341–361, 2000.

S.F. Jarner and G.O. Roberts. Polynomial convergence rates of Markov Chains. *Ann. Appl. Probab.*, 12(1):224–247, 2002.

J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series, 2002.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958.

D.E. Kaufman and R.L. Smith. Direction Choice for Accelerated convergence in Hit-and-Run Sampling. *Operations Research*, 46(1), 1998.

D.Y. Lin and Z. Ying. Cox regression with incomplete covariate measurements. *J. Amer. Statist. Assoc.*, 88:1341–1349, 1993.

L. Lovász. Hit-and-Run mixes fast. *Math. Program.*, 86(Ser. A):443–461, 1999.

L. Lovász and S. Vempala. Hit-and-Run is fast and fun. *preprint, Microsoft Research*, 2003.

L. Lovász and S. Vempala. Hit-and-Run from a corner. *SIAM J. Comput.*, 35:985–1005, 2006.

K.L. Mengersen and R.L. Tweedie. Rate of convergences of the Hasting and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 1996.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phsy.*, 21:1087–1091, 1953.

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

M. Pintilie. *Competing Risks: A Practice Perspective*. Wiley Series, 2006.

M. Pintilie, Y. Bai, L.S. Yun, and D. Hodgson. The analysis of case cohort design in the presence of competing risks with application to the analysis of the effect of treatment for Hodgkin Lymphoma on cardiac events. *In preparation*, 2009.

R.L. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11, 1986.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

G.O. Roberts and W.R. Gilks. Convergence of Adaptive Direction Sampling. *J. Multiv. Analys.*, 49:287–298, 1994.

G.O. Roberts and J.S. Rosenthal. Two convergence properties of hybrid samplers. *Ann. Appl. Prob.*, 8:397–407, 1998.

G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.*, 16:351–367, 2001.

G.O. Roberts and J.S. Rosenthal. General state space Markov chains and MCMC algorithms. *Prob. Surv.*, 1:20–71, 2004.

G.O. Roberts and J.S. Rosenthal. Harris recurrence of Metropolis-within-Gibbs and Trans-dimensional Markov Chain. *Ann. Appl. Prob.*, 16(4):2123–2139, 2006.

G.O. Roberts and J.S. Rosenthal. Coupling and Ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Prob.*, 44:458–475, 2007.

G.O. Roberts and J.S. Rosenthal. Examples of Adaptive MCMC. *J. Comp. Graph. Stat.*, 2009.

G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110, 1996.

G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.*, 7:110–120, 1997.

G.O. Roberts, J.S. Rosenthal, and P.O. Schwartz. Convergence propperties of perturbed Markov chains. *J. Appl. Prob.*, 35:1–11, 1998.

J.S. Rosenthal. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *J. Amer. Stats. Assoc.*, 90:558–566, 1995.

E. Saksman and M. Vihola. On the Ergodicity of the Adaptive Metropolis Algorithms on Unbounded Domains. *Preprint*, 2008.

S.G. Self and R.L. Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.*, 16:64–81, 1988.

M. Tanner and W. Wong. The calculation of posterior distributions by data argumentation. *J. Amer. Stats. Assc.*, 82:528–550, 1987.

S. Wacholder, M.H. Gail, D. Pee, and R. Brookmeyer. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*, 76:117–123, 1989.

C. Yang. On the weak law of large number for unbounded functionals for adaptive MCMC. *Preprint*, 2008a.

C. Yang. Recurrent and Ergodic Properties of Adaptive MCMC. *Preprint*, 2008b.