MARKOV CHAIN MONTE CARLO ALGORITHMS AND RELATED CONVERGENCE PROPERTIES

Student: Yufan Li
Supervised by: Jeffery Rosenthal

April 2018

Abstract

This document records my work on the Engineering Science Undergraduate thesis study. I will review and discuss topics I have studied so far as well as presenting some of my original work. My thesis study so far consists of two major parts: (I) expository studies on discrete-time Markov process on general state space and convergence of generic Markov Chain Monte Carlo (MCMC) algorithms, as well as some results concerning weak convergence of a sequence of discrete-time chains to a continuous process; (II) studies on adaptive MCMC algorithm. First part of my thesis study provides necessary theoretical foundations to understand results and methods in part (II) whereas the second part provides context and examples of application to my studies in part (I).

A rough breakdown of my studies in these two parts is as follows: Part (I): (1) expository studies on various results of Markov chains on general state space, (2) expository studies on ergodicity property of Markov Chain Monte Carlo (MCMC) algorithms, along with quantitative bounds on the rate of convergence to stationarity; (3) weak convergence of discrete chains to a continuous process and related complexity bounds results; Part (II): (1) studies on generic adaptive MCMC and their ergodicity properties, (2) studies on various adaptive Gibbs Samplers, (3) original studies on the asymptotic behavior of the "Stairway to Heaven" example proposed in (Łatuszyński et al., 2013), Section 3.

Acknowledgement

I offer my sincerest gratitude to Professor Jeffery Rosenthal for supervising this thesis study. He introduced me to this immensely intriguing topic and have since guided me to study it from different perspectives. Professor Rosenthal allows me to explore topics and problems that I found interesting while giving me advice on specific directions to take. By the end of the study, I am able to understand the topic in a systematic and intuitive fashion thanks to Professor Rosenthal's guidance. I could not have asked for a better thesis experience.

I would also like to thank my parents Yinghong Wong and Ming Li for their unconditional love and support. Their encouragement is vital to my growth and development.

Contents

1 Introduction

In this section, I will briefly discuss the context of the thesis study and main areas of focus. A significant portion of my thesis work consists of expository studies on existing results. For these parts, I will identify major interests and objectives and outline important methodologies associated. Besides expository studies, there are also some original studies involved (mostly on the "Stairway to Heaven" example), I will provide context to it as well and outline my methods.

The first part of my thesis study focuses on the Markov Chain Monte Carlo (MCMC) algorithms in general and their ergodicity properties. Since analysis of MCMC algorithms often involves concepts and methods developed in the general framework of discrete-time, general state space Markov chains, my study includes a significant portion of fundamental theories and methods from this area, e.g. irreducibility, recurrence, small sets, "coupling" etc. These concepts are essential to understand current research on MCMC algorithms: for example, study on the quantitative rate of convergence relies on the Minorization condition and the coupling argument. I also studied weak convergence of Random Walk Metropolis algorithm to the Langevin process and relevant general theories. The second portion of my thesis study involves expository and original studies on adaptive MCMC algorithms and adaptive Gibbs samplers.

1.1 Motivation

The Markov Chain Monte Carlo (MCMC) is a popular sampling algorithm in statistics; it is widely used to approximately sample complicated probability distribution in high dimensions. The need for MCMC arises from the difficulty to draw samples directly from such distributions: it is not always possible to compute explicit form of the associated, complicated integrals.

For example, MCMC is instrumental in Bayesian statistics which would often requires computing large hierarchical models with a large number of parameters. To estimate expectation of function $f : \mathcal{X} \to \mathbb{R}$ under *posterior distribution*, which may be written as $\frac{\pi_u}{\int_{\mathcal{X}} \pi_u(x)dx}$ (the integral on the denominator is the *normalizing constant*):

$$E(f(X)) = \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx},$$

we need to compute closed form of $\int_{\mathcal{X}} \pi_u(x)dx$ in order to sample from it. However, this expression often takes an irregular form and contains hundreds and thousands of variables. MCMC provides an alternative to sample from such distribution without the need to compute the integral explicitly: for example, a Gibbs sampler would only require the knowledge of conditional distribution of the target distribution and Metropolis-Hastings algorithm only requires the *ratio* of the target density (Roberts and Rosenthal, 2004) (Gelfand and Smith, 1990) (Smith and Roberts, 1993) (Tierney, 1994).

MCMC algorithms should be thought of as an approximation of the direct sampling from the target distribution: the validity of MCMC sampling resides on the *convergence* of the Markov chain to the target distribution, meaning that after running the chain for sufficient long period of time, the probability distribution of the samples drawn from MCMC become fairly close to the target distribution. The "closeness" of two probability measures can be measured rigorously with *total variation distance*:

$$||v_1(\cdot) - v_2(\cdot)|| = \sup_A |v_1(A) - v_2(A)|$$

Therefore, when we say that the Markov chain converges to the target distribution, we are referring to

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

It is of great interest for statisticians to understand when and how a Markov process will converge to stationarity: in practice, it is necessary to make sure that the MCMC in question indeed converges to the target distribution so that the samples obtained actually approximate target distribution (after a sufficient large number of runs). It is also important to have a sense of how quickly the convergence occurs, that is, how many runs would be sufficient to attain satisfactory approximation.

1.2 Research Areas and Methods

To answer these questions, a theoretical framework was developed concerning general state space Markov process (Meyn and Tweedie, 2012) (Roberts and Rosenthal, 2004). The purpose has been to impose certain conditions on the Markov chains, e.g. irreducibility, so that under these conditions the Markov chain would have desirable stability and ergodicity properties. This theoretical framework can be developed by first establishing results on countable state space. The conditions can then be modified carefully so that an analogous theory can be developed for the general state space chain. One example would be the extension of recurrence/transience dichotomy from countable state space chains to general state space chains as in (Meyn and Tweedie, 2012), Chapter 8: The result is first established for irreducible countable state space chains; and then it is extended to *general state space chain with "atom set"*, a construction specifically designed to mimic countable state space chains. Then the assumption of the existence of atom set can be relaxed via Nummelin splitting, the viability of which is guaranteed by $\psi-$irreducibility (through Minorization condition).

A significant portion of my thesis study involves studying the development of this theoretic framework following (Meyn and Tweedie, 2012) and (Roberts and

Rosenthal, 2004). Specifically, I have focused on examining fundamental conditions and accompanied auxiliary constructions based on which the analysis of general state space Markov chains was made possible. Important topics include: irreducibility, small/petite sets, splitting chains, periodicity, recurrence and stationarity.

In close association with these topics, I have also studied (Rosenthal, 1995) and (Rosenthal, 1996) on *quantitative rates of convergence*. Utilizing the Minorization condition and coupling, (Rosenthal, 1995) provided a general method to analyze convergence speed of discrete-time, general state-space Markov chain. Their methods provide a rigorous, a priori bounds on how long MCMC should run to attain satisfactory results. Application of this methodology involves finding a proper auxiliary function and a number of associated parameters to establish both the Minorization condition and the Drift condition. (Rosenthal, 1996) supplies an example of how this method can be applied to a realistic hierarchical Bayesian model (James-Stein estimator). The convergence bounds developed in (Rosenthal, 1995) are often used in adaptive MCMC literature to simultaneously bound convergence speed of adaptive kernels to establish the Containment Condition. In addition, I have also studied results concerning weak convergence of Random Walk Metropolis (RWM) algorithms to the Langevin process in continuous time, along with relevant theories. Specifically, (Roberts et al., 1997) discovered that under certain technical assumptions, when proposal variance is appropriately scaled according to n, the sequence of stochastic processes formed by the first component of each Markov chain converges to appropriate limiting Langevin diffusion process. The proof is based on showing convergence of the infinitesimal generators; I have reorganized the proof so that motivation for certain technical lemmas becomes more obvious (the proof remains the same). The significance of the result is such that the limiting diffusion approximation sheds lights on the efficiency optimization problem: the asymptotically optimal acceptance rate is shown to be 0.234.

The second part of my study focuses on a special class of MCMC algorithms, i.e. *the adaptive MCMC*. This type of sampler is needed mainly due to the difficulty to "tune" parameters such as scaling manually. The intention of adaptation is to enable the algorithm to "learn" the best parameter values automatically while they run. One notable example of adaptive MCMC is the adaptive Metropolis algorithm proposed in (Haario et al., 2001). Their design allows the algorithm to optimize the proposal distribution of a Metropolis algorithm. It is shown in (Roberts et al., 1997) that the proposal $N(x, (2.38)^2\Sigma/d)$ is optimal in a large dimension settings where d is the number of dimensions and Σ is the covariance matrix of the target distribution. Therefore, the adaptive routine involves estimating Σ from empirical distribution of the chain output as it runs and adapts proposal distribution accordingly. There are a variety of designs of adaptive MCMC algorithms with different features: directional sampling, for example, rotates coordinate axis in order to improve mixing efficiency, thus suitable to sample from slanted, elongated target distribution (Bai, 2009). See (Roberts and Rosenthal, 2009) for some other designs of adaptive MCMC and empirical evaluation of their performances.

Since adaptive MCMC may no longer be Markovian, their convergence properties require further analysis. With a coupling construction, (Roberts and Rosenthal, 2007) provided conditions under which ergodicity and stationarity of the specified target distribution is ensured. Since their work employs minimal assumptions of the adaptive MCMC in question, the result provides a general guideline to analyze ergodicity properties of adaptive MCMC algorithms. Though not necessary conditions in a theoretical sense, (Łatuszyński and Rosenthal, 2014) showed that they are generally not redundant in practice.

My studies of adaptive MCMC then focused on a special subclass, i.e. the adaptive Gibbs samplers, following (Łatuszyński et al., 2013). This study applied the generic framework provided by (Roberts and Rosenthal, 2007) and focused on ergodicity properties of a series of adaptive Gibbs/Metropolis-within-Gibbs samplers that grow in sophistication. They presented various positive results guaranteeing convergence of those adaptive algorithms. To support their results, the authors presented a cautionary example of a simple-seeming adaptive Gibbs sampler that eventually fails to converge. An elaborate proof was given to show the process tends to infinity with probability larger than 0. Specifically, it is used as a counter example to refute the following proposition, which was presented erroneously in (Levine and Casella, 2006):

An adaptive random scan Gibbs sampler is ergodic if its adaptive selection probability α_n converges to α and a random scan Gibbs sampler with fixed selection probability α induces a ergodic Markov chain with stationary distribution π.

In hope to simplify the proof and strengthen their result, I worked to produce two proofs using different methodologies. My proofs strengthen the original results that such process tends to infinity with probability larger than 0 with proper choice of “adaptation speed”: I was able to show that there such probability can be larger than any $\sigma \in [0, 1)$ under certain choices of “adaptation speed”. Proof one uses an auxiliary process that converges in probability; proof two involves construction of countably infinitely many phases.

2 Part I: Markov Process on General State Space and General MCMC Algorithms

2.1 Preliminary Definitions and Examples

We will include some basic definitions and fundamental results in this section.

We first define *Transition Probability Kernels* as the following:

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

Theorem 2.1. For any initial measure μ on $\mathcal{B}(X)$, and any transition probability kernel P, there exists a stochastic process $\Phi = \{\Phi_0, \Phi_1, \cdots\}$ on $\Omega = \Pi_{i=0}^{\infty} X_i$, measurable with respect to $\mathcal{F} = \vee_{i=0}^{\infty} \mathcal{B}(X_i)$, and a probability measure P_μ on $\mathcal{F}$ such that $P_\mu(B)$ is the probability of the event $\{\Phi \in B\}$ for $B \in \mathcal{F}$; and for measurable $A_i \subseteq X_i, i = 0, ..., n$ and any n

$$P_\mu(\Phi_0 \in A_0, \Phi_1 \in A_1, ..., \Phi_n \in A_n)$$
$$= \int_{y_0 \in A_0} \cdots \int_{y_{n-1} \in A_{n-1}} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, A_n).$$

Remark. The relevant sources from which proof of this theorem can be found is outlined in (Meyn and Tweedie, 2012). This theorem says that given a transition kernel, a discrete process exists such that it “proceeds by that kernel” in a way we typically understand a time-homogeneous Markov process. We will explain this further following the next definition.

Definition 2.2. The stochastic process Φ defined on $(\Omega, \mathcal{F})$ is called a time-homogeneous Markov chain with transition probability kernel $P(x, A)$ and initial distribution μ if the finite dimensional distributions of Φ satisfy the following for every n:

$$P_\mu(\Phi_0 \in A_0, \Phi_1 \in A_1, ..., \Phi_n \in A_n)$$
$$= \int_{y_0 \in A_0} \cdots \int_{y_{n-1} \in A_{n-1}} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, A_n).$$

A time-homogeneous Markov chain is memoryless (transition probability only depends on current step) and its transition probability to “a destination” depends only on its starting position, regardless of time.

Example: Random Walk on Half Line

Given a sequence of i.d.d random variables $\{W_i\}$ taking values in $\mathbb{Z}$. We define Markov process $\{\Phi_n\}$ as $\Phi_n = [\Phi_{n-1} + W_n]^+$.

Example: Markov Process on Finite Groups

Let $(G, *)$ be a finite group. Let $n = |G|$. Given a probability distribution μ on G, the transition probabilities $P(g, h * g) := \mu(h)$ define a Markov chain. In words, the chain moves via left multiplication by a random element of G selected according to μ.

Example: Renewal process

Let $\{Y_n\}$ be a sequence of i.d.d random variables with distribution function Γ concentrated on R^+. Let Y_0 be a further independent random variable, with distribution of Y_0 being Γ_0 concentrated on R^+. The random variables $Z_n := \sum_{i=0}^n Y_i$ are called a delayed renewal process while if $\Gamma_0 = \Gamma$ then the sequence is referred to as a renewal process.

Notably, write $\Gamma_0 * \Gamma$ for the convolution of Γ_0 and Γ given by

$$\Gamma_0 * \Gamma(dt) = \int_0^t \Gamma(dt - s)\Gamma_0(ds) \tag{1}$$

By decomposing successively over the values of the first n variables $Z_0, ..., Z_{n-1}$, we have that

$$P(Z_n \in dt) = \Gamma_0 * \Gamma^{n*}(dt) \tag{2}$$

Proposition 2.1. Chapman-Kolmogorov equation: For any m with $0 \leq m \leq n$,

$$P^n(x, A) = \int_{\mathcal{X}} P^m(x, dy)P^{n-m}(y, A), x \in \mathcal{X}, A \in B(X) \tag{3}$$

Remark. Chapman-Kolmogorov equation is instrumental in Markov theories. We will see it appearing at numerous occasions in the following Chapters. Essentially, it can be interpreted as saying that the process move from starting point into A by first taking m steps to some intermediate position $y \in X$; and moves succeeding $(n - m)$ steps with law appropriate to starting afresh at y (it forgets the past before time m).

Definition 2.3. Skeletons and Resolvents: The chain Φ^m is called the $m-$skeleton of the chain Φ if it satisfies the following transition law:

$$P_x(\Phi_n^m \in A) = P^{mn}(x, A) \tag{4}$$

The resolvent K_{a_ϵ} is defined for $0 < \epsilon < 1$ by

$$K_{a_\epsilon} := (1 - \epsilon) \sum_{i=0}^{\infty} \epsilon^i P^i(x, A), x \in X, A \in B(x) \tag{5}$$

Remark. An $m-$skeleton can be understood as the sped-up version of the original Markov chain (we observe the value of the chain every two steps); A resolvent is essentially a *sampled chain*: it is still technically a sped up version of the original chain but the times by which it is sped up is determined probabilistically, where the number of steps the chain will progress per time unit is first sampled from a geometric distribution. Resolvent is associated with the concept of accessibility . As we will see later, a destination set A is accessible from x if and only if the probability of resolvent reaching A form x in one step is larger than 0.

2.1.1 Occupation Times, Return Times and Hitting Times etc.

Definition 2.4. We define occupation times, first return and first hitting times on A respectively as:

$$\gamma_A := \sum_{n=1}^{\infty} \mathbb{1}(\Phi_n \in A) \tag{6}$$

$$\tau_A := \min\{n \geq 1 : \Phi_n \in A\} \tag{7}$$

$$\sigma_A := \min\{n \geq 0 : \Phi_n \in A\} \tag{8}$$

We also define two kernels U, L for later use:

$$U(x,A) := \sum_{n=1}^{\infty} P^n(x,A) = \sum_{n=1}^{\infty} E(\mathbb{1}(\Phi_n \in A)) = E\bigg(\sum_{n=1}^{\infty}(\mathbb{1}(\Phi_n \in A))\bigg) = E_x[\gamma_A] \tag{9}$$

$$L(x,A) := P_x(\tau_A < \infty) \tag{10}$$

Remark. The occupation times is the total number of times the chain stayed in set A; τ_A is the first time the chain *return* to A (not including the starting point); σ_A is the first time the chain hits set A–it includes the starting point. $U(x,A)$ is the expected number of times the chain stayed in set A; $L(x,A)$ is the probability of x return to A in finite steps. As we will see, these concepts are very useful in later sections.

2.2 Irreducibility

Irreducibility is a fundamental concept in Markov chain theory, which basically states that all parts of the state space can be reached by a Markov chain regardless of the starting point. In order to facilitate the analysis of general state space chain, this concept is modified to $\varphi-$irreducibility with respect to measure ψ. An important result is the existence of a maximal irreducibility measure ψ, which describes the range of the chain much more completely. A closely associated concept is accessibility: the idea is that once we know which sets can be reached with positive probability from a particular starting point, then we will have some idea of how the chain will behave in the long term.

2.2.1 $\varphi-$Irreducibility

Definition 2.5. $\varphi-$Irreducibility for general space chains: We call Φ $\varphi-$irreducible if there exists a measure φ on $B(X)$ such that, whenever $\varphi(A) > 0$, we have $L(x, A) > 0$ for all $x \in X$.

Proposition 2.2. Equivalent formulation of $\varphi-$irreducibility:

1. for all $x \in X$, whenever $\varphi(A) > 0$, $U(x, A) > 0$;
2. for all $x \in X$, whenever $\varphi(A) > 0$, there exists some $n > 0$, possibly depending on both A and x, such that $P^n(x, A) > 0$;
3. for all $x \in X$, whenever $\varphi(A) > 0$ then $K_{1/2}(x, A) > 0$.

2.2.2 $\psi-$Irreducibility

Proposition 2.3. If a Markov chain is $\varphi-$irreducible for some φ, then probability measure ψ exists such that it is a maximal irreducibility measure, satisfying the following conditions:

1. Φ is $\psi-$irreducible;
2. for any other measure φ', the chain X is $\varphi'-$irreducible if and only if $\psi \succ \varphi'$;
3. if $\psi(A) = 0$, then $\psi(\{y : L(y, A) > 0\}) = 0$;
4. the probability measure ψ is equivalent to $\psi'(A) := \int_{\mathcal{X}} \varphi'(dy) K_{1/2}(y, A)$, for any finite irreducibility measure φ'

Remarks: Among different choices of irreducibility measures, ψ is the "maximal" in the sense that for any set A, $\varphi'(A) > 0$ implies $\psi(A) > 0$ where φ' is any irreducibility measure. Note that the $\succ$ sign denotes absolute continuity of measures and two measures which are mutually absolutely continuous are called equivalent. (2) says that if a set is "negligible" by the maximal irreducibility measure, which would imply it is negligible by all other irreducible measures, then the set of starting points that will reach A is also negligible: $\psi(\{y : L(y, A) > 0\}) = 0$.

Definition 2.6. We call a set $A \in B(X)$ full if $\psi(A^c) = 0$; we call a set $A \in B(X)$ absorbing if $P(x, A) = 1$ for $x \in A$.

One important usage of $\psi-$irreducibility, as opposed to $\varphi-$irreducibility, is to guarantee effective analysis of restrictions of chains to full set utilizing the following two propositions. An example we will present later is the theorem concerning periodicity of $\psi-$irreducible chains where the cyclic classes cover only a full set on the state space instead of the whole state space as for countable state space chains.

Proposition 2.4. Suppose that Φ is $\psi-$irreducible. Then every absorbing set is full; every full set contains a non-empty, absorbing set. Here an absorbing set is essentially equivalent to the absorbing communicating class in countable space.

Proposition 2.5. Suppose that A is an absorbing set. Let P_A denote kernel P restricted to the states in A. Then there exists a Markov chain Φ_A whose state space is A and whose transition matrix is given by P_A. Moreover, if Φ is $\psi-$irreducible then Φ_A is $\psi-$irreducible.

2.2.3 Accessibility

Definition 2.7. We say that a set $B \in B(X)$ is accessible from another set A if $L(x, B) > 0$ for every $x \in A$;

We say that a set $B \in B(X)$ is uniformly accessible from another set A if there exists a $\sigma > 0$ such that

$$\inf_{x \in A} L(x, B) \geq \sigma \tag{11}$$

Note that the relation $A \rightsquigarrow B$, the uniformly accessibility of A to B, is nonreflexive in general, but it is transitive.

Proposition 2.6. Let $\bar{A} := \{x \in X : L(x, A) > 0\}$ and $\bar{A}(m) := \{x \in X : \sum_{n=1}^{m} P^n(x, A) \geq m^{-1}\}$.

$$\bar{A} = \cup_m \bar{A}(m)$$

and for each m we have $\bar{A}(m) \rightsquigarrow A$.

Remark. The first statement is obvious: consider $m \to \infty$. The second statement follows from the following:

$$L(x, A) \geq P_x(\tau_A \leq m) \geq m^{-1}$$

The purpose of this Proposition has been to show that state space X can be covered by sets from which any given $A \in B^+(X)$ is uniformly accessible.

2.3 Pseudo-atoms

The purpose of an atom set has been to endow general state space chain with analogous properties of countable state space chain. The rationale of this construction can be understood as an attempt to "simplify" the Chapman Kolmogorov equation. A very important bridge between general state space chain with atom and any general state space chain is the existence of a "split" chain. A split chain is defined in a way such that it always admits an atom set and the transition probability is directly linked to the original chain. Provided that a *Minorizaion Condition* is satisfied, a chain can always be split and thus analyzed effectively through its split counterpart. The Minorization condition, or existence of small sets, can in turn be guaranteed via aperiodicity and irreducibility. This series of deduction allows us to infer important stability and ergodicity properties of general state space chains through easily verifiable conditions, i.e. irreducibility and aperiodicity.

Definition 2.8. A set $\alpha \in B(X)$ is called an atom for Φ if there exists a measure v on $B(X)$ such that

$$P(x, A) = v(A), x \in \alpha \tag{12}$$

If Φ is $\psi-$irreducible and $\psi(\alpha) > 0$ then α is called an accessible atom

Remark: The existence of "artificial atom" for φ-irreducible chains is one single result that makes general state space Markov chain theory as powerful as countable space theory.

2.3.1 Splitting φ-Irreducible Chains

Proposition 2.7. Suppose there is an atom α in X such that $\sum_n P^n(x, \alpha) > 0$ for all $x \in X$. Then α is an accessible atom and Φ is v-irreducible with $v = P(\alpha, \cdot)$

Proposition 2.8. If $L(x, A) > 0$ for some state $x \in \alpha$, where α is an atom, then $\alpha \to A$.

Definition 2.9. Minorization Condition: For some $\sigma > 0$, some $C \in B(X)$ and some probability measure v with $v(C^C) = 0$ and $v(C) = 1$

$$P(x, A) \geq \sigma \mathbb{1}_C(x) v(A), A \in B(X), x \in X. \tag{13}$$

Definition 2.10. Construction of Split Chain: Define a new Markov chain based on the original chain by the following steps

1. Split state space X into $X' = X \times \{0,1\}$ where $X_0 := X \times \{0\}$ and $X_1 := X \times \{1\}$ are thought of as copies of X equipped with copies of $B(X_0), B(x_1)$ of the $\sigma-$field $B(X)$. Let $B(X')$ be the σ-field generated by $B(X_0), B(x_1)$;
2. For any measure λ on $B(X)$, split it into two measures on each of X_0, X_1 by defining the measure λ' on $B(X')$ through: $\lambda'(A_0) = \lambda(A \cap C)[1 - \sigma] + \lambda(A \cap C^C)$; $\lambda'(A_1) = \lambda(A \cap C)\sigma$ where σ, C are the constant and the corresponding small set, and $A_0 \in X_0, A_1 \in X_1$;
3. Define the split kernel $P'(x_i, A)$ for $x_i \in X'$ and $A \in B(X')$ by

$$P'(x_0, \cdot) = P(x, \cdot)^*, x_0 \in X_0 - C_0 \tag{14}$$

$$P'(x_0, \cdot) = [1-\sigma]^{-1}[P(x, \cdot)^* - \sigma v'(\cdot)], x_0 \in C_0 \tag{15}$$

$$P'(x_1, \cdot) = v'(\cdot), x_0 \in X_1 \tag{16}$$

Remark: The exact way a split chain operates is in fact not of our interests as far as application is concerned. The purpose of this construction has been nothing but to find a chain such that it has the same transition probability as the original chain and equipped with an atom set (stated formally in Theorem 2.2). Of course, the existence will be inferred from the Minorization condition, guaranteed through strong aperiodicity and irreducibility. In the above definition, we assume small set C exist; it is later shown that for $\varphi-$irreducible chains small sets for which the minorization condition holds exist, at least for the $m-$skeleton.

A few notes on the construction: original measure λ is the marginal measure induced by $\lambda' : \lambda'(A_0 \cup A_1) = \lambda(A)$. Outside C the chain Φ' behaves just like Φ, moving on the top half X_0 of the split space. Each time it arrives in C, it is split; with probability $1 - \sigma$ it remains in X_0 with probability σ it drops to C_1. The bottom level X_1 is an atom, with $\varphi'(X_1) = \sigma\varphi(C) > 0$ whenever the chain Φ is φ-irreducible. From the definition above, $C_1 \subseteq X_1$ is the only part of the bottom level which is reached with positive probability.

Theorem 2.2. The original chain Φ is the marginal chain of Φ': that is, for any initial distribution λ on $B(X)$ and any $A \in B(X)$,

$$\int_X \lambda(dx)P^k(x, A) = \int_{X'} \lambda'(dy_i)P^{'k}(y_i, A_0 \cup A_1). \tag{17}$$

The chain Φ is $\varphi-$irreducible if Φ' is $\varphi'-$irreducible; and if Φ is $\varphi-$irreducible with $\varphi(C) > 0$ then Φ' is $v^{'}-$irreducible, and C_1 is an accessible atom for the split chain.

Remarks. This theorem equates transition probability of the original chain to its split counterpart. If some results can be proved for general state space chain endowed with atom set, we may immediately transfer such result to the original chain, provided that the splitting is viable.

2.3.2 Small Sets

Definition 2.11. Small Sets: A set $C \in B(X)$ is called a small set if there exists an $m > 0$, and a non-trivial measure v_m on $B(X)$, such that for all $x \in C$, $B \in B(X)$,

$$P^m(x, B) \geq v_m(B). \tag{18}$$

We say C is v_m-small if the above holds.

Remark. Intuitively, a set is small if the transition probability from each of its element to any set B can be bounded uniformly by a "shrunk" probability measure (the non-trivial measure v_m). However, small set should be understood as a way to "translate" irreducibility to more concrete information regarding behavior of the chain. For example, it may be used with Chapman-Komogorov equation to derive certain inequalities; it may be combined with strong aperiodicity condition to produce Minorization condition, which guarantees splitting of the chain. We shall see many of those applications in the following sections.

Theorem 2.3. If Φ is $\psi-$irreducible, then for every $A \in B^+(X)$, there exists $m \geq 1$ and a v_m-small set $C \subseteq A$ such that $C \in B^+(X)$ and $v_m\{C\} > 0$. Here $B^+(X)$ denote $\{B|\psi(B) > 0\}$.

This shows that if Φ is $\psi-$irreducible, every set $A \in B^+(X)$ contains a small set in $B^+(X)$. As a consequence, every ψ-irreducible chain admits some $m-$skeleton which can be split and for which the atomic structure of the split chain can be exploited. In other words, some $m-$skeleton always obeys the Minorization condition when $\psi-$irreducibility holds.

Proposition 2.9.

1. If $C \in B(X)$ is v_n-small, and for any $x \in D$ we have $P^m(x, C) \geq \sigma$, then D is $v_{n+m}-$small, where v_{n+m} is a multiple of v_n
2. Suppose Φ is $\phi-$irreducible. Then there exists a countable collection C_i of small sets in $B(X)$ such that

$$X = \cup_{i=1}^{\infty} C_i \tag{19}$$

3. Suppose Φ is $\psi-$irreducible. If $C \in B^+(X)$ is v_n-small, then we may find $M \in Z_+, M > n$ and a measure v_M such that C is v_M-small, and $v_M(C) > 0$.

Remark. (1) can be proved by the Chapman-Kolmogorov equation. For (2), the collection can be found by identifying the v_m-small set $C \in B^+(X)$. Since the chain is $\psi-$irreducible (so $\exists n \in \mathbb{N}^+, P^n(x, A) > 0, \forall x \in X, \psi(A) > 0$), the following set covers :

$$C(n, k) := \{y : P^n(y, C) \geq k^{-1}\}$$

Notice that each $C(n, k)$ is small because of (1) – they all admit positive probability of reaching small set C in some n steps. To see (3), since $C \in B^+(X)$, from equivalence definition of $\varphi-$irreducible, we have $K_{a1/2}(x, C) > 0$ for all $x \in X$. Therefore, it can be deduced that the probability of starting from $v(\cdot)$ and reaching C after some m steps is positive. Let $v_M(C) := vP^m(C) > 0$. For all $x \in C$,

$$P^{n+m}(x, B) = \int_X P^n(x, dy) P^m(y, B) \geq vP^m(B) = v_M(B),$$

where $M = n + m$. The purpose of (3) has been to show that with a small set C, we may assume with out loss of generality that $v_M(C) > 0$.

Example: Small Set on Random walk on a half line: Provided that $\Gamma(-\infty, 0) > 0$, i.e. there is a positive probability of negative increment: there exists $\epsilon > 0, \sigma > 0$ such that the increment $P(W < -\epsilon) > \sigma$. Then $\{0\}$ is small by definition and for any compact set $D := [a, b]$ we have $P^{b/\epsilon}(x, \{0\}) > \sigma^{b/\epsilon}, \forall x \in D$. So any compact set is small by Proposition above.

In addition to the above Proposition, we document two results useful for proving existence of small sets:

Proposition 2.10. Given a positive integer k_0 and a subset $R \subseteq X$, there exists a probability measure $Q(\cdot)$ so that

$$P^{k_0}(x, \cdot) \geq \epsilon Q(\cdot) \forall x \in R, \tag{20}$$

where

$$\epsilon = \int_X (\inf_{x \in R} P^{k_0}(x, dy)) \tag{21}$$

Remarks. This Proposition follows trivially from the following inequality:

$$P^k(x, A) \geq \int_A (\inf_{x \in R} P^k(x, dy))$$

This is proven to be a very handy result to be used to show that certain set is small. Most typical application is when $P^k(x, \cdot)$ is a unimodal distribution with x being the mode whilst $R := [a, b]$ Therefore, $\inf_{x \in R} P^{k_0}(x, dy) = \min(P^{k_0}(a, dy), P^{k_0}(b, dy))$.

Proposition 2.11. Consider a sequentially-updated Gibbs sampler with n dimension. Suppose that for some d, conditional on values for $X_1^k, ..., X_d^k$, the random variables $X_{d+1}^k, ..., X_n^k$ are independent for all $X_i^{k'}, k' < k$. Then if we establish Minorization condition for $X_1^k, ..., X_d^k$, i.e.

$$\mathcal{L}(X_1^{k_0}, ..., X_d^{k_0} | (X_1^0, ..., X_n^0) = x) \geq \epsilon' Q'(\cdot), \forall x \in R \tag{22}$$

Then there is a probability measure $Q(\cdot)$ on X such that

$$P^{k_0}(x, \cdot) \geq \epsilon' Q(\cdot) \tag{23}$$

Proof. Suppose that for any measurable set $A \subseteq \mathcal{X}_1 \times ... \times \mathcal{X}_d, \forall x \in R$,

$$P^{k_0}(x, A) \geq \epsilon' Q'(A).$$

Define $Q(\cdot)$ on $\mathcal{X}$ such that its marginal distribution on the first d coordinates agrees with $Q'(\cdot)$ and the conditional distribution on the first d coordinates is defined as

$$Q(X_{d+1}, ..., X_n | X_1, ..., X_d) = \mathcal{L}(X_{d+1}, ..., X_n | X_1, ..., X_d)$$

Notice that a distribution is fully described with marginal distribution of a set of coordinates and conditional distribution on the same set of coordinates, i.e. let $X \sim Q(\cdot)$ and X_1, B_1 represent first d coordinates and X_2, B_2 the remaining coordinates,

$$\begin{aligned} Q(B) &= P(\{X_1 \in B_1\} \cap \{X_2 \in B_2\}) \\ &= P(\{X_2 \in B_2\} | \{X_1 \in B_1\}) Q'(B_1) \\ &= E_{X_1 \in B_1}(E(\mathbb{1}\{X_2 \in B_2\} | X_1 = x, \{X_1 \in B_1\})) Q'(B_1) \\ &= E_{X_1 \in B_1}(P(\{X_2 \in B_2\} | X_1 = x)) Q'(B_1) \end{aligned}$$

where the expectation is known since marginal distribution of X_1 is known. The result follows. □

2.4 Cyclic Behavior

The existence of small sets allows us to study cyclic behavior of Markov chain on general state space. It should serve yet another example of application of small sets. First I will review some results on countable space.

Definition 2.12. In the countable space, we define period of a single state α as the following:

$$d(\alpha) = \gcd n \geq 1 : P^n(\alpha, \alpha) > 0 \tag{24}$$

Proposition 2.12. For any $y \in C(\alpha) := \{y : \alpha \leftrightarrow y\}$, $d(\alpha) = d(y)$

Proposition 2.13. Let Φ be a irreducible Markov chain on countable space, and d the common period of the states. Then there exists disjoint sets $D_1, ..., D_d$ such that they cover X and $P(x, D_{k+1}) = 1, x \in D_k, k = 0, ..., d-1$

Definition 2.13. An irreducible chain on a countable space X is called

1. aperiodic, if $d(x) \equiv 1, x \in X$;
2. strongly aperiodic, if $P(x, x) > 0$ for some $x \in X$.

Proposition 2.14. Let Φ be a irreducible Markov chain on countable space, and d the common period of the states, and cyclic classes are $D_1, ..., D_d$. Then for a Markov chain with transition matrix P^d, each D_i is an irreducible absorbing set of aperiodic states.

One major theme of the Markov chain theory on general state space is to "transfer" results on countable state space to general state space chain through existence of small set. Now I will document one instance of using this method– using small sets to prove existence of "disjoint" $d-$cycles for a $\psi-$irreducible Markov chain:

Theorem 2.4. Suppose that Φ is a $\psi-$irreducible Markov chain on X. Let $C \in B^+(X)$ be a v_M-small set and let d be the greatest common divisor of the following set:

$$E_C = \{n \geq 1 : \text{the set } C \text{ is } v_n\text{−small, with } v_n = \delta_n v \text{ for some } \delta_n > 0\}$$

Then there exists disjoint sets $D_1, ..., D_d \in B(X)$ such that for $x \in D_i$, $P(x, D_{i+1}) = 1, i = 0, ..., d-1 \mod d$. And the set not covered by the $d-$cycle, i.e. $N = [\cup_{i=1}^d D_i]^c$, is $\psi-$null.

The $d-$cycle $\{D_i\}$ is maximal in the sense that for any other collection $\{d', D'_k, k = 1, ..., d'\}$ satisfying the properties above, we have $d'|d$; whilst if $d = d'$, then, by reordering the indices if necessary, $D'_d = D_i,\ a.e.\psi$

Proof. The idea of the proof is to first construct d overlapping sets as the following:

$$D_i^* = \left\{ y : \sum_{n=1}^{\infty} P^{nd-i}(y, C) > 0 \right\}, i = 0, .., d-1$$

which covers X due to irreducibility – $\psi(C) > 0$. Then we proceed to prove that $\psi(D_i^* \cap D_k^*) = 0$, i.e. the overlapping is $\psi-$null. By contradiction, assume that $\exists A \subseteq D_i^* \cap D_k^*$ with $\psi(A) > 0$. Therefore, for some $\omega \in A$, there exists some n_i, n_k such that

$$P^{n_i d-i}(\omega, C) \geq \sigma_i > 0$$

$$P^{n_k d-k}(\omega, C) \geq \sigma_k > 0$$

Utilizing the property of maximal irreducibility, since $\psi(A) > 0$, $\exists r$ such that

$$\int_C v(dy) P^r(y, A) = \sigma_c > 0.$$

Notice that we cannot simply say that since $\psi(A) > 0$, there exists some r such that $P(x, A) \geq \sigma_c > 0$. It is necessary to eliminate "variable" x via integration: this is the reason why we take a detour to C first.

For $x \in C, B \subseteq C$

$$P^{2M+md-i+r}(x, B) \geq$$

$$\int_C P^M(x, dy) \int_A P^r(y, dw) \int_C P^{md-i}(w, dz) P^M(z, B) \geq \sigma_c \sigma_m v(B)$$

Notice that here the path the chain follows is: $x \to C \to A \to C \to B$. As explained we take a detour to C before go to A to uniformly bound the integral. This shows that $2M+md+r-i \in E_C$. Similarly, we have $2M+nd+r-k \in E_C$. This contradicts the definition of d and we have thus shown that $\psi(D_i^* \cap D_k^*) = 0$.

Let N be the union of overlapping between $\{D_i^*\}$. Since $\psi(N) = 0$, the sets $\{D_i^* \setminus N\}$ form a disjoint class of sets whose union is full. We know that there exists an absorbing set $D \in \cup_i D_i^* \setminus N$. $D_i = D \cap D_i^* \setminus N$ thus are disjoint and if $x \in D$ is such that $P(x, D_j) > 0$, then we have $x \in D_{j-1}$ because that implies $P^{nd-j+1}(x, C) = P^{nd-(j-1)}(x, C) > 0$ for some n.

To show the maximality, it suffices to show that for each $n \in E_C$, $d'|n$. In order to show this, we shall show that for some cyclic class D_i', $P^n(x, D_i') > 0, x \in D_i'$ for all $n \in E_C$. However, we know that for some cyclic class D_i', $v_1(D_i' \cap C) > 0$ because otherwise $v_1(C) = 0$; this implies that $P^n(x, D_i' \cap C) > v_1(D_i' \cap C) > 0, \forall x \in C$. The result follows. We can further show that $C \cap D_j' \equiv \emptyset, \forall j \neq k$; because if not, for $x \in C \cap D_j'$, $P^M(x, C \cap D_k') = 0$ (the chain is bound to return to D_j' after multiple of d' steps).

□

This result motivates the following definition regarding periodicity of Markov chains on general state space.

Definition 2.14. Suppose that the chain is $\psi-$irreducible. The largest d for which a $d-$cycle occurs is called period of the chain. If $d = 1$, the chain is called aperiodic. When there exists v_1-small set A with $v_1(A) > 0$, then the chain is called strongly aperiodic.

Based on these definitions and the main theorem, we have the following useful propositions,

Proposition 2.15. Suppose that Φ is a $\psi-$irreducible Markov chain.
(i) If Φ is strongly aperiodic, then the Minorization Condition (single step) holds;
(ii) The resolvent, or $K_{a_\epsilon}-$chain, is strongly aperiodic for all $0 < \epsilon < 1$.
(iii) If the chain is aperiodic then every skeleton is $\psi-$irreducible and aperiodic, and some $m-$skeleton is strongly aperiodic

Remarks. Our results regarding existence of small sets suggest that Minorization is always satisfied for some $m-$skeleton of the chain. (i) tells us that for a strongly aperiodic chain, $m = 1$, i.e. Minorization condition is satisfied for the original chain and we can split it immediately. So it is always preferable to work with strongly aperiodic chain. The general method is to prove results for strongly aperiodic chains and then extend them to general state space chains through the $m-$skeleton or $K_{a_\epsilon}-$chain.

Notice that for a periodic chain, such v_1-small set does not exists such that it has positive $\psi-$measure: to prove by contradiction, assume such small set does exist; $A = (A \cap N) \cup (\cup_{i=0}^{d-1}(A \cap D_i))$: $\psi(A \cap N) = 0 \implies v_1(A \cap N) = 0$ because otherwise $P(x, A \cap N) > v_1(A \cap N) > 0, \forall x \in A \implies \psi(A) = 0$; for any i such that $A \cap D_i \neq \emptyset$, we have $x \in A \cap D_i$, $P(x, D_1 \cap A) = 0 \implies v_1(D_1 \cap A) = 0$. Therefore, $v_1(A) = 0$ and the result follows.

2.5 Petite Sets and Sampled Chains

Let $a = \{a(n)\}$ be a distribution, or probability measure, on $\mathbb{Z}_+$ and consider the Markov chain Φ_a with probability transition kernel

$$K_a(x, A) := \sum_{n=0}^{\infty} P^n(x, A)a(n)$$

Probabilistically, Φ_a has the interpretation of being the chain Φ "sampled" at time-points drawn successively according to the distribution a. For example, if

$$a_\epsilon(n) = [1 - \epsilon]\epsilon^n, n \in \mathbb{Z}_+,$$

we sample from the geometric distribution (thus obtaining some integer n) and progress the original chain by n steps. Notice that in this case, the chain is the resolvent K_ϵ as defined before. One use of this definition is that it allows for development of useful conditions under which uniform accessibility can be inferred, i.e. if a set $B \in B(X)$ is *uniformly accessible using* a from another set $A \in B(X)$, that is, if there exists a $\sigma > 0$ such that

$$\inf_{x \in A} K_a(x, B) > \sigma,$$

then $A \rightsquigarrow B$. This is because $L(x, B) > K_a(x, B)$ for any a.

In addition to the above, we will document some other useful Propositions concerning this definition:

Proposition 2.16. (i) Let $a * b$ denote the convolution of a and b,

$$K_{a*b}(x, A) = \int K_a(x, dy)K_b(y, A)$$

;
(ii) If $A \rightsquigarrow_a B$ and $A \rightsquigarrow_b B$, then $A \rightsquigarrow_{a*b} B$;
(iii) If a is a distribution on $\mathbb{Z}_+$ then

$$U(x, A) \geq \int U(x, dy)K_a(y, A)$$

Remarks: The probabilistic interpretation of (i) is that if the chain is sampled at a random time $\eta = \eta_1 + \eta_2$, where η_1 has distribution a and η_2 has independent distribution b, then since η has distribution $a * b$ (definition of convolution), it follows that (i) is just a Chapman-Kolmogorov decomposition at the intermediate random time.

Small sets always exist in the $\psi-$irreducible case, and provide most of the properties needed. Petite sets, on the other hand, have more tractable properties:

Definition 2.15. We call a set $C \in B(x)$ v_a-petite if the sampled chain satisfy the bound

$$K_a(x, B) \geq v_a(B), \forall x \in C, B \in B(X),$$

where v_a is a non-trivial measure on $B(X)$

Obviously, a small set is a petite set with $a := \delta_m(m)$. Petite set, on the other hand, is small set with respect to the sampled chain. Other important properties are the following:

Proposition 2.17. (i) If $A \in B(X)$ is v_a-petite, and $D \rightsquigarrow_b A$ then D is $v_{b*a}-$petite;
(ii) If Φ is $\psi-$irreducible and if $A \in B^+(X)$ is v_a-petite, then v_a is an irreducibility measure for Φ.

Remarks. Apparently, (i) is useful to deduce that some set D is petite set upon knowing another petite set and their relation–similar result exists for small sets as we have seen before. The proof for (i) is just writing out the definition; to see (ii), we need to show that given $v_a(B) > 0$, $P^n(x, B) > 0$ for all $x \in X$ and some n. It is sufficient to show that $P^n K_a(x, B) > 0$ since this essentially implies that it is probable to reach B from x in finite steps. We also use the small set cover of X that we have proved existence, i.e. $A(n, k) = \{y : P^n(y, A) \geq k^{-1}\}$:

$$P^n K_a(x, B) \geq \int_A P^n(x, dy) K_a(y, B) \geq m^{-1} v_a(B) > 0.$$

In summary, this proposition shows when a petite/small set associated measure is an irreducibility measure, which is generally very useful in later proofs.

Here is another set of useful propositions concerning petite sets. These properties generally do not apply to small sets. (ii), (iii) are apparently consequences of (i); as we will see later, the existence of an increasing sequence of petites that covers the state space is especially useful in certain proofs.

Proposition 2.18. Suppose Φ is $\psi-$irreducible.

(i) If A is v_a-petite, then there exists a sampling distribution b such that A is also ψ_b-petite where ψ_b is a maximal irreducibility measure;
(ii) The union of two petite sets is petite;
(iii) There exists a sampling distribution c, an everywhere strictly positive, measurable functions: $s : X \to R$, and a maximal irreducibility measure ψ_c such that

$$K_c(x, B) \geq s(x)\psi_c(B), x \in X, B \in B(X)$$

Thus there is an increasing sequence $\{C_i\}$ of ψ_c-petite sets, all with the same sampling distribution c and Minorizing measure equivalent to ψ, with $\cup C_i = X$

Remarks. The proof of (i) is to first show that v_a is an irreducibility measure: we use the fact that there exists small set $C \in B^+(X)$ which is also v_b-petite for some b where v_b is an irreducibility measure by previous results; then we can show that $A \rightsquigarrow_{a*a_\epsilon} C$ and thus derive that A is $v_{a*a_\epsilon*b}-$petite where $v_{a*a_\epsilon*b}$ is a irreducibility measure since it is multiple of v_b. Then we can show that with $0 < \epsilon < 1$,

$$K_{a*a_\epsilon}(x, B) = K_a K_{a_\epsilon}(x, B) \geq v_a K_{a_\epsilon}(B), x \in A, B \in B(X)$$

Since v_a is an irreducibility measure, the measure $v_a K_{a_\epsilon}$ here is a maximal irreducibility by previous results.

Proposition 2.19. Suppose that Φ is $\psi-$irreducible and that C is v_a-petite.

(i) Without loss of generality we can take a to be either a uniform sampling distribution $a_m(i) = 1/m, 1 \leq i \leq m$, or a to be the geometric sampling distribution a_ϵ. In either case, there is a finite mean sampling time

$$m_a = \sum_i ia(i).$$

(ii) If Φ is strongly aperiodic then the set $C_0 \cup C_1 \subseteq X'$ corresponding to C is v_a^*-petite for the split chain Φ'

Remark. Too see (i), we note that for any v_n-small set $A \in B^+(X)$, $\sum_{k=1}^N P^k(x, A) \geq 1/2\psi_b(A), x \in C$, using Proposition 2.18. Then we can show that

$$\sum_{k=1}^{N+n} P^k(x, B) \geq \sum_{k=1}^{N} P^{k+n}(x, B) \geq \frac{1}{2}\psi_b(A)v_n(B)$$

2.6 Transience and Recurrence

Stable chains are conceived as those which do not vanish from their starting points in some ways. The focus of this portion of the study will be on the behavior of the occupation time:

Definition 2.16 (Uniform Transience and Recurrence)**.** A set A is called uniformly transient if there exists $M < \infty$ such that

$$E_x[\eta_A] \leq M, \forall x \in A$$

The set A is called recurrent if

$$E_x[\eta_A] = \infty, \forall x \in A$$

2.6.1 Chains with an atom

In this section, I will review how to classify a chain that admits an atom to either recurrent or transient, through the splitting technique we reviewed previously.

Theorem 2.5. *Suppose that a chain is $\psi-$irreducible and admits an atom $\alpha \in B^+(X)$. Then*
(i) if α is recurrent, then every set in B^+ is recurrent. (ii) if α is transient, then there is countable covering of X by uniformly transients sets.

Remarks. To see (i), just apply Chapman-Kolmogorov equation (x goes to α; α goes to α; α to any target set A). For (ii), it is important to pay attention here to the use of the "atom" construct: it is essentially conceived to mimic behavior of a single state on the countable state space.

An important theme of this section is to explore the relation between return/hitting time "τ_A" and occupation time "η_A" in the context of recurrence. The first entrance and last exist decomposition provides a link between the two. For general state space Markov chain, the decomposition equations assumes the following form (see (Meyn and Tweedie, 2012) p.184 for detail):

$$U^{(z)}(x, B) = U_A^{(z)}(x, B) + \int_A U_A^{(z)}(x, dw)U^{(z)}(w, B),$$

$$U^{(z)}(x, B) = U_A^{(z)}(x, B) + \int_A U^{(z)}(x, dw)U_A^{(z)}(w, B),$$

With the decomposition equations and the existence of atomic set, we may classify the general chains as the following:

Theorem 2.6. Suppose that Φ is $\psi-$irreducible and admits an atom $\alpha \in B^+(X)$. Then
(i) if α is recurrent, then every set in $B^+(X)$ is recurrent.
(ii) if α is transient, then there is a countable covering of X by uniformly transient sets.

Proof. For (i), using the property of the atomic set α,

$$\sum_n P^{r+s+n}(x,A) \geq \sum_n \int_\alpha P^r(x,dw) \int_\alpha P^n(w,dz)P^m(z,A)$$

$$= P^r(x,\alpha)[\sum_n P^n(\alpha,\alpha)]P^m(\alpha,A) = \infty$$

We can see that the purpose of atom here has been to transform the Chapman-Kolmogorov equation on general state space to its countable state space form. The advantage □

Due to the property of the atom set α, the above can be simplified to the form that is identical to the countable space case. For example, the last exist decomposition is:

$$U^{(z)}(x,\alpha) = U_\alpha^{(z)}(x,\alpha) + U^{(z)}(x,\alpha)U_\alpha^{(z)}(\alpha,\alpha),$$

So results on the countable state space are directly applicable. For example, we can solve for expected hitting time (with z) in terms of the *taboo* probability here:

$$U^{(z)}(x,\alpha) = \frac{U_\alpha^{(z)}(x,\alpha)}{1 - U_\alpha^{(z)}(\alpha,\alpha)} \leq \frac{1}{1 - L(\alpha,\alpha)}.$$

Notice that

$$L(x,A) = \sum_{n=1}^{\infty} {}_A P^n(x,A) = \lim_{z\uparrow 1} U_A^{(z)}(x,A)$$

The condition that α is transient ensures that $U(x,\alpha)$ is bounded for all x. Consider the countable covering of X given by,

$$\bar{\alpha}(j) = \{y : \sum_{n=1}^{j} P^n(y,\alpha) > j^{-1}\}$$

We can use Chapman-Kolmogorov to show that for each j, $\bar{\alpha}(j)$ is indeed transient:

$$U(x,\alpha) \geq j^{-1}U(x,\bar{\alpha}(j)) \inf_{y\in\bar{\alpha}(j)} \sum_{n=1}^{j} P^n(y,\alpha) \geq j^{-2}U(x,\bar{\alpha})(j)$$

This leads to the following definition:

Definition 2.17 (Transient Sets)**.** If $A \in B(X)$ can be covered with a countable number of uniformly transient sets, then we call A transient.

2.6.2 The General Recurrence/transience Dichotomy

This section will extend results in the last section, which assumes existence of atom. This is one example of using the splitting techniques introduced previously to extend results from countable state space to general state space. Regarding stability classification of $\psi-$irreducible chains, we have the following definition:

Definition 2.18. (i) The chain is called recurrent if it is $\psi-$irreducible and $U(x, A) \equiv \infty, \forall x \in X, \forall A \in B^+(X)$.
(ii) The chain is transient if it is $\psi-$irreducible and X is transient.

Starting from the Proposition below, we can see how our construction of split chain becomes useful in extending results from countable state space to general state space.

Proposition 2.20. Suppose that Φ is $\psi-$irreducible and strongly aperiodic. Then either both Φ and $\check{\Phi}$ are recurrent, or both Φ and $\check{\Phi}$ are transient.

Remarks. Strong aperiodicity ensure that Minorization Condition holds and thus Nummelin Splitting is viable for Φ. With the "split" Dirac measure δ_x^*, we have the following equality,

$$\sum_{n=1}^{\infty} \int \delta_x^*(dy_i) \check{P}^n(y_i, B) = \sum_{n=1}^{\infty} P^n(x, B).$$

Then the result follows.

The following Proposition provides a link between the recurrence of the chain and its resolvent. In the proof, we can see that it uses some results we had for sampled chains.

Proposition 2.21. For any $0 < \epsilon < 1$,

$$\sum_{n=1}^{\infty} K_{a_\epsilon}^n = \frac{1-\epsilon}{\epsilon} \sum_{n=0}^{\infty} P^n$$

Remarks. The proof presented in (Meyn and Tweedie, 2012) (not stated explicitly) uses the following property of the convolution, which in turn follows from Fubini's theorem :

$$\int_{R^d} (f * g)(x) dx = (\int_{R^d} f(x) dx)(\int_{R^d} g(x) dx).$$

As a direct consequence of the Proposition above,

Proposition 2.22. Suppose that Φ is $\psi-$irreducible.
(i) The chain is transient if and only if each resolvent chain K_{a_ϵ} is transient;
(ii) The chain is recurrent if and only each K_{a_ϵ} chain is recurrent

Since the chain is $\psi-$irreducible, we know that the resolvent is strongly aperiodic. The dichotomy (either transient or recurrent) is already established for the split chain of the resolvent and by Proposition 2.20 we know that the dichotomy extends to the resolvent chain. The Proposition 2.22 further extends the dichotomy to the original chain, which gives us the following theorem.

Theorem 2.7. If the chain is $\psi-$irreducible, then it is either recurrent or transient.

Remarks. Here we can see how the dichotomous results from countable state space chain is extended to general state space: we first prove similar results with the assumption that atom set exists and "remove" that assumption by utilizing Nummelin Splitting technique.

We also have the following concerning the skeleton chain.

Theorem 2.8. Suppose that Φ is $\psi-$irreducible and aperiodic.
(i) The chain is transient if and only if one, and then every, $m-$skeleton Φ^m is transient.
(ii) The chain is recurrent if and only if one, and then every, $m-$skeleton Φ^m is recurrent.

Remarks. We will note the following equality:

$$\sum_{i=1}^{\infty} P^j(x,A) = \sum_{r=1}^{m}\sum_{j=1}^{\infty} P^{r+jm}(x,dy) = \sum_{r=1}^{m}\sum_{j=1}^{\infty}\int P^r(x,dy)P^{jm}(y,A)$$

$$= \sum_{r=1}^{m}\int P^r(x,dy)\sum_{j=1}^{\infty} P^{jm}(y,A)$$

With this (i), and $\Leftarrow$ of (ii) should be easy to see. For (ii) $\Rightarrow$, notice that from $\psi-$irreducibility and aperiodicity of the original chain, we know any skeleton is also $\psi-$irreducible (here aperiodicity is necessary). So we know that the skeleton is dichotomous but it cannot be transient due to (i).

The recurrence considered in this section is weaker than more desirable recurrence property known as *Harris Recurrence*, which requires that $L(x,A) \equiv 1$ for all $x \in A$ and $A \in B^+(X)$. For countable chains, recurrence would imply Harris Recurrence due to the following Proposition:

Proposition 2.23. For countable chains, $U(x,x) = \infty$ if and only if $L(x,x) = 1$.

Unfortunately, we are unable to establish analogous results for chains on general state space thus far.

The following theorem provides various means to bound $U(x, A)$ so that they can be used to ensure that a set is uniformly transient.

Proposition 2.24. Suppose that Φ is a Markov chain but not necessarily irreducible.
(i) If set $A \in B(X)$ is uniformly transient with $U(x, A) \leq M, \forall x \in A$, then $U(x, A) \leq 1 + M, \forall x \in X$;
(ii) If any set $A \in B(X)$ satisfies $L(x, A) = 1, \forall x \in A$, then A is recurrent. If Φ is $\psi-$irreducible, then $A \in B^+(X)$, then $A \in B^+(X)$ and we have $U(x, A) \equiv \infty, \forall x \in X$;
(iii) if any set $A \in B(X)$ satisfies $L(x, A) \leq \epsilon < 1, \forall x \in A$, then we have $U(x, A) \leq 1/[1 - \epsilon], \forall x \in X$ so that A is uniformly transient;
(iv) Let $\tau_A(k)$ denote the $k-$th return time to A, and suppose that for some m

$$P_x(\tau_A(m) < \infty) \leq \epsilon < 1, x \in A;$$

then $U(x, A) \leq 1 + m/[1 - \epsilon], \forall x \in X$.

Remarks. (i) is a direct result of using first entrance decomposition, which is often applied to link $U(x, A)$ with $x \in A$ to $U(x, A)$ with $x \in X$. It mostly serves to extend bound on $U(x, A)$ with $x \in A$ to the entire state space X; (ii) basically states that Harris recurrence is stronger condition than recurrence, providing a link between τ_A and η_A; (iii) allows us to bound $U(x, A)$ given that $L(x, A)$ is uniformly bounded; it can be shown using last exist decomposition (yet again); (iv) instead can be used to bound $U(x, A)$ given that probability of "making more than $k - th$ return" is bounded below 1. The proof uses induction: for fixed $m \in \mathbb{Z}_+$, given

$$P_x(\eta_A \geq m) \leq \epsilon, x \in A$$

i.e. the probability of "making more than $(m-1)-th$ return" is bounded below 1. By induction, we can bound the probability of making more than or equal to $m(k + 1)$ hits,

$$P_x(\eta_A \geq m(k + 1)) = \int_A P_x(\Phi_{\tau_A(km)} \in dy) P_y(\eta_A \geq m) \leq \epsilon^{k+1}$$

With this bound over the $m-$skeleton, we may bound each step in a "stairway" fashion: for $x \in A$,

$$U(x, A) = \sum_{n=1}^{\infty} P_x(\eta_A \geq n)$$

$$\leq m[1 + \sum_{k=1}^{\infty} P_x(\eta_A \geq km)] \leq m/(1 - \epsilon)$$

.

The following proposition can be used to identify other uniform transient set given the existence of one uniformly transient set.

Proposition 2.25. If A is uniformly transient, and $B \rightsquigarrow_a A$ for some a, then B is uniformly transient. Hence if A is uniformly transient, there is a countable covering of $\bar{A}$ by uniformly transient sets.

Proof. This is another example of using accessibility to infer property of one set from another. As we have seen in previous sections, similar results exist with petite set (small set), transitive property of uniform accessibility. To prove such property usually requires Chapman-Kolmogorov equation or formula of similar form. To see the first statement, recall the following inequality,

$$U(x, A) \geq \int U(x, dy) K_a(y, A) \geq \delta U(x, B)$$

The second statement comes from the following result from previous sections:

$$\bar{A} = \cup_m \bar{A}(m)$$

where $\bar{A}(m) \rightsquigarrow A, \forall m$ (so each $\bar{A}(m)$ is uniformly transient). □

2.6.3 Harris Recurrence

In this section, we will review some stronger concepts of recurrence. This section will be focused on establishing some link between the "return time counts" and first return time probabilities. Let's first define the event that $\Phi \in A$ infinitely often (i.o), or $\eta_A = \infty$:

$$\{\Phi \in A \ i.o.\} := \cap_{N=1}^{\infty} \cup_{k=N}^{\infty} \{\Phi_k \in A\}$$

which obviously a stronger condition than $U(\cdot, A) = E(\eta_A) = \infty$. We will use the following notation:

$$Q(x, A) := P_x\{\Phi \in A \ i.o.\}.$$

We have the following equality due to the strong Markov property:

$$Q(x, A) = P(\{\tau_A < \infty\} \cap \{\Phi \in A \ i.o. \ \Phi_0 \in A\})$$

$$= E(1\{\{\tau_A < \infty\} \cap \{\Phi \in A \ i.o. \ \Phi_0 \in A\}\})$$

$$= P_{\Phi_{\tau_A}}\{\Phi \in A \ i.o.\} \cdot E_x[1\{\tau_A < \infty\}]$$

$$= E_x[P_{\Phi_{\tau_A}}\{\Phi \in A \ i.o.\} 1\{\tau_A < \infty\}]$$

$$= \int_A (\sum_{i=1}^{\infty} P^i(x, dy)) Q(y, A) = \int_A U_A(x, dy) Q(y, A)$$

Here, to calculate the expectation of the random variable $P_{\Phi_{\tau_A}}\{\Phi \in A \ i.o.\} 1\{\tau_A < \infty\}$, we partition the sample space according to the location of the first return.

The definition for Harris Recurrence is as the following:

Definition 2.19. The set A called Harris recurrent if

$$Q(x, A) = P_x(\eta = \infty) = 1, x \in A$$

A chain is called Harris recurrent if it is $\psi-$irreducible and every set in $B^+(X)$ is Harris recurrent.

We first develop conditions to ensure that a set is Harris recurrent based on first return time probabilities $L(x, A)$.

Proposition 2.26. Suppose for some set $A \in B(X)$ we have $L(x, A) \equiv 1, x \in A$. Then

$$Q(x, A) = L(x, A), \forall x \in X$$

and in particular A is Harris recurrent.

Proof. The concepts/methodology used to prove this proposition is useful for our later studies. So I will supply a detailed proof here. Firstly, we are able to acquire the following equality using strong Markov property:

$$P_x(\tau_A(2) < \infty) = \int_A U_A(x, dy)L(y, A) = 1$$

Intuitively, this can be understood as the probability of reaching to some point in A ($U_A(x, dy)$) with finite steps and then return to A in finite steps. Strong Markov property allows us to inductively extend to any finite $k-$return time: i.e. with probability 1, the chain will return to A. For any x, we have the following:

$$P_x(\eta_A \geq k) = P_x(\tau_A(k) < \infty) = 1$$

i.e. the probability of the chain "occupying" A for at least k times equals to the probability it returns to A more than k-th times. Then,

$$Q(x, A) = \lim_{k \to \infty} P_x(\eta_A \geq k) = 1, \forall x \in A$$

. Then,

$$Q(x, A) = \int_A U_A(x, dy)Q(y, A) = L(x, A)$$

□

Remarks. Intuitively, this Proposition states that if the chain returns to set A with probability 1, then it returns to set A infinitely often with probability 1. This shows that the two definitions of Harris recurrence are in fact identical, i.e. the one that uses occupation time $P_x(\eta_A = \infty) = 1, x \in A$ and the one that uses return time $L(x, A) = 1, x \in A$.

The following theorem can be very useful in proving Harris recurrence of a set given uniform accessibility (denoted with "$\rightsquigarrow$"). To prove the theorem, we need the Martingale Convergence Theorem. Let's first define martingale and supermartingale:

Definition 2.20. A sequence of integrable random variables $\{M_n : n \in Z_+\}$ is called adapted to an increasing family of $\sigma-$fields $\{\mathcal{F}_n : n \in Z_+\}$ if M_n is $\mathcal{F}-$measurable for each n. The sequence is called a martingale if $E[M_{n+1}|\mathcal{F}_n] = M_n$ for all $n \in Z_+$ and a supermartingale if $E[M_{n+1}|\mathcal{F}_n] \leq M_n$ for $n \in Z_+$

Theorem 2.9. For a supermartingale M_n for which

$$\sup_n E(|M_n|) < \infty,$$

$\{M_n\}$ converges to a finite limit with probability one.

Theorem 2.10. (i) Suppose that $D \rightsquigarrow A$ for any sets D and A in $B(X)$. Then

$$\{\Phi \in D \; i.o.\} \subseteq \{\Phi \in A \; i.o.\} \; a.s.$$

and therefore $Q(y, D) \leq Q(y, A), \forall y \in X$;
(ii) If $X \rightsquigarrow A$ then A is Harris recurrent, and $Q(x, A) \equiv 1, \forall x \in X$.

Proof. Intuitively, (i) is easy to understand: since the chain visits set D infinitely often and there is some $\epsilon > 0$ chance for the chain to go from D to A each time, the chain will end up visiting A infinitely often as well. The proof is to show that the chain visiting D infinitely often implies $\lim_n L(\Phi_n, A) = 1$, which can be shown to be equivalent to the event of the chain returning to set A infinitely often. We use the following notation:

$$E_n = \{\Phi_{n+1} \in A, n \in \mathbb{Z}_+\}$$

One important step of the proof is to show that

$$P[\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}] \to \mathbb{1}(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} E_i) \; a.s.$$

This equality establishes a link between the probability of “returning” and the occurrence of “infinitely often returning to the set”. To see this, note that for fixed $k \leq n$

$$P(\cup_{i=k}^{\infty} E_i | \mathcal{F}_n^{\Phi}) \geq P(\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}) \geq P(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} E_i | \mathcal{F}_n^{\Phi}).$$

Notice that

$$\lim_{n\to\infty} P(\cup_{i=k}^{\infty} E_i | \mathcal{F}_n^{\Phi}) = \mathbb{1}(\cup_{i=k}^{\infty} E_i)$$

$$\lim_{n\to\infty} P(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} E_i | \mathcal{F}_n^{\Phi}) = \mathbb{1}(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} E_i)$$

By the Martingale Convergence Theorem, we know that the following limit exists,

$$\limsup_n P[\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}], \; \liminf_n P[\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}]$$

Therefore, we have the following inequality,

$$\mathbb{1}(\cup_{i=k}^{\infty} E_i) \geq \limsup_n P[\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}]$$

$$\geq \liminf_n P[\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}] \geq \mathbb{1}(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} E_i)$$

As $k \to \infty$, the two extreme terms converge, which shows the equality we desired. As a result, by strong Markov property,

$$\lim_{n\to\infty} L(\Phi_n, A) = \lim_{n\to\infty} P[\cup_{i=n}^{\infty} E_i | \mathcal{F}_n^{\Phi}] = \mathbb{1}(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} E_i) \ a.s.$$

Also note that

$$\mathbb{1}(\cap_{m=1}^{\infty} \cup_{i=m}^{\infty} \{\Phi_i \in D\}) \leq \mathbb{1}(\limsup_n L(\Phi_n, A)) > 0 = \mathbb{1}(\lim_n L(\Phi_n, A) = 1)$$

The result follows. (ii) follows from (i) easily.

Another method of proof could be the following: we know that $\eta_D = \infty, a.s.$, if the chain visits D infinitely often i.e.

$$P(\{\omega \in \Omega | \eta_D(\omega) = \infty\}) = 1$$

while due to uniform accessibility

$$L(x, A) = E(E(\mathbb{1}(\{\exists n < \infty, \Phi_n \in A\} | \sigma(\eta_D))))$$

$$= E(1 - (1-\epsilon)^{\eta_D}) = \int_{\Omega} (1 - (1-\epsilon)^{\eta_D(\omega)}) P(d\omega) = 1$$

□

The result we just proved can be used to show the following:

Theorem 2.11. If Φ is Harris recurrent then $Q(x, B) = 1$ for every $x \in X$ and every $B \in B^+(X)$.

Remarks. This theorem simply means that for a Harris recurrent chain, the chain will hit every "large enough" set infinitely often regardless of its starting point. From Harris recurrent, we know that every $B \in B^+(X)$ is recurrent and that the chain is $\psi-$irreducible. We can use results we obtained from our study of the petite sets: there exists an increasing sequence of ψ_c-petite sets (countable) $\{C_i\}$ that covers X, all with the same sampling distribution c and Minorizing measure equivalent to ψ (recall that for each v_a-petite set it is also a ψ_b-petite where ψ_b is maximal irreducibility measure, which is unique to the specific chain). We can choose a large enough n such that $x \in C_n$ and $\psi(C_n) > 0$ (this ensures that C_n is Harris recurrent). By definition,

$$K_c(y, B) \geq \psi_c(B) = \sigma > 0, \forall y \in C_n$$

which implies that $C_n \rightsquigarrow_a B \Rightarrow C_n \rightsquigarrow B$. The result follows.

Example. Recurrent But Not Harris-Recurrent
We may construct a chain that is recurrent but not Harris recurrent: consider a chain Φ that is Harris recurrent on state space X. We can expand the state space to $X' := X \cup N$ where N consists of a sequence of individual points $\{x_i\}$. We can further define the transition probability:

$$P'(x_i, x_{i+1}) = \beta_i,\ P'(x_i, \alpha) = 1 - \beta_i$$

where $\alpha \in X$ and $0 < \prod_{i=0}^{\infty} \beta_i < 1$.

This construction ensures that

$$L'(x_i, A) = L'(x_i, \alpha) = 1 - \prod_{i=1}^{\infty} \beta_i < 1, A \in B^+(X)$$

Therefore, any set $B \subset X'$ with $B \cap X \in B^+(X)$ and $B \cap N$ non-empty is not Harris recurrent. However, since

$$U'(x_i, A) = \sum_{n=1}^{\infty} P^n(x_i, A) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} P^{n+m}(x_i, A) \geq \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} P^m(x_i, \alpha) P^n(\alpha, A)$$

$$= \sum_{m=1}^{\infty} P^m(x_i, \alpha) \sum_{n=1}^{\infty} P^n(\alpha, A) \geq L'(x_i, \alpha) U(\alpha, A) = \infty, A \in B(X)$$

every set in $B^+(X')$ is recurrent. We will later see that the only way in which an irreducible chain can be recurrent and not Harris recurrent is by the existence of an absorbing set which is Harris recurrent, accompanied by a single $\psi-$null set on which the Harris recurrence fails.

Definition 2.21. Maximal absorbing set: For any Harris recurrent set D, we write $D^\infty = \{y : L(y, D) = 1\}$. We know that $D \subseteq D^\infty$ and D^∞ is absorbing (the chain can never move from D^∞ to any point such that the probability to reach D is less than 1.); D is a maximal absorbing set if $D = D^\infty$.
Maximal Harris sets: We call a set maximal Harris if H is a maximal absorbing set such that Φ restricted to H is Harris recurrent.

With this definition, we have the following theorem:

Theorem 2.12. If Φ is recurrent, then

$$X = H \cup N$$

where H is non-empty maximal Harris set and N is transient.

Proof. The proof first requires construction of set H and set N. Then it proceeds to prove that H is indeed a non-empty maximal Harris set and that N is transient. First, we set

$$H = \{y : Q(y, C) = 1\}, N = H^c$$

where C is a ψ_a−petite set in $B^+(X)$. Since $H^\infty = H(H^\infty \in H$ because if the chain reaches H with probability 1 then it reaches C with probability 1), either H is empty or H is maximal absorbing. Suppose that H is empty, i.e. $Q(x, C) < 1$ for all x. It can be show that

$$C_1 := \{x \in C : L(x, C) < 1\}$$

is in $B^+(X)$. For if it is not, we can find an absorbing full set $F \subset C_1^c$ and

$$L(x, C \cap F) = 1, x \in C \cap F \implies Q(x, C \cap F) = 1, x \in C \cap F$$

which contradicts the premise $Q(x, C) < 1$ for all x. Notice that $C \cap F$ is non-empty because that would imply $C \subseteq F^c \implies \psi_a(C) = 0$.

Since $\psi(C_1) > 0$ there exists $B \subseteq C_1, B \in B^+(X)$ and $\delta > 0$ with $L(x, C_1) \leq \delta < 1, \forall x \in B$, i.e.

$$L(x, B) \leq L(x, C_1) \leq \delta, x \in B.$$

Then by previous results on transient sets, we have $U(x, B) \leq [1-\delta]^{-1}$ and this contradicts that Φ is recurrent.

Therefore H is a non-empty maximal absorbing set, and H is full. Then we know that N is ψ_a−null. Note that from equivalent definition of the irreducibility, we know that $\exists k, P^k(x, N^c) > 0, \forall x \in X$. Therefore, since

$$E(\eta_N) = \sum_n nk(1 - P^k(x, N^c))^n < \infty$$

we may conclude that set N is transient.

To show that H is also Harris, we note that $C \rightsquigarrow A, \forall A \in B^+(X)$ and by construction $Q(x, C) = 1, \forall x \in H$. Therefore, $Q(x, A) = 1, \forall x \in H, A \in B^+(X)$.

□

2.7 Quantitative Convergence Rates of MCMC

In this section, I will review some results on convergence rates of MCMC. This section mainly follows (Rosenthal, 1995) and (Rosenthal, 1996). The main results specifies a generic method to derive *quantitative convergence rates* of a general Markov process upon verifying the Minorization and Drift condition. These results are also used in adaptive MCMC literature to bound convergence speed so that it provides an alternative to check the Containment Condition.

2.7.1 Minorization and Convergence Rates

Here I state another definition of Minorization conditions for Markov chains:

Definition 2.22. (Another definition of Minorization Condition) A Markov chain with transition kernel $P(x, dy)$ on a state space X satisfies a minorization condition on a subset $R \subseteq X$ if there is a probability measure $Q(\cdot)$ on X, a positive integer k_0, and $\epsilon > 0$, such that

$$P^{k_0}(x, A) \geq \epsilon Q(A), \forall x \in R, A \in B(X) \tag{25}$$

Theorem 2.13. Suppose that a Markov chain $P(x, dy)$ on a state space X satisfies Minorization condition. Let $X^{(k)}, Y^{(k)}$ be two realizations of the Markov chain (started in any initial distribution), defined jointly as described in the proof, Let

$$t_1 = \inf\{m : (X^{(m)}, Y^{(m)}) \in R \times R\}$$

, and for $i > 1$ let

$$t_i = \inf\{m : m \geq t_{i-1} + k_0, (X^{(m)}, Y^{(m)}) \in R \times R\}$$

Set $N_k = max\{i : t_i < k\}$. Then for any $j > 0$,

$$||L(X^{(k)}) - L(Y^{(k)})||_{var} \leq (1 - \epsilon)^{[j/k_0]} + P(N_k < j) \tag{26}$$

Remarks on the Proof:

1. N_k can be conveniently interpreted as counts of both chains stepping into small set R before k if $k_0 = 1$; when $k_0 \neq 1$, then N_k is the latest time when both chains stepping into small set R and since the "coin tossing" procedure is only administrated for k_0-skeleton chain, there are at most $[j/k_0]$ times the chain may couple in the small set given $N_k \geq j$

2. The proof of this theorem provides a good example of utilizing Minorization condition to show convergence through "coupling" argument. The idea is to construct simultaneously two joint Markov chain such that they each follows the transition probability $P(x, dy)$ marginally, and one of

them is at stationarity. It is critical that these two chains coincide after a random amount of time T(coupling time). This generic "coupling" argument leads to the following inequality ("coupling" inequality):

$$||L(X^{(k)}) - L(Y^{(k)})||_{var} \leq P(X^{(k)} \neq Y^{(k)}) \leq P(k < T) \tag{27}$$

3. The idea of Minorization condition comes from the desire to "bring together" both chains: one starts from stationarity and the other from any initial distribution. The small set here is the venue at which such encounter may happen while they still each abide by the transition probability P marginally:

$$X_{n+1}/Y_{n+1} \sim \begin{cases} Q^*(\cdot), & \text{if } X_n, Y_n \in R, I = 0 \\ \dfrac{P(X_n/Y_n, \cdot) - \epsilon Q(\cdot)}{1-\epsilon}, & \text{if } X_n, Y_n \in R, I = 1 \\ P(X_n/Y_n, \cdot), & \text{otherwise} \end{cases} \tag{28}$$

$Q^*(\cdot)$ denotes that both chains proceed according to distribution $Q(\cdot)$ but with perfect correlation, while in other cases both chains proceed independently, and I is a Bernoulli random variable with $P(I = 0) = \epsilon$.

The results in Theorem 2.13 can be applied to form quantitative bound to the convergence rate by carefully bounding $P(N_k) < j$. Specifically, with the following proposition:

Proposition 2.27.

$$P(N_k < j) \leq \alpha^{-k} E(\prod_{i=1}^{j} \alpha^{r_i}), \forall \alpha > 1, r_i = t_i - t_{i-1} \tag{29}$$

Remarks: This can be proven using Markov's inequality.

Proposition 2.28. Suppose that there is $\alpha > 1$ and a function $h : X \times X \to \mathbb{R}$ such that $h \geq 1$ and

$$E(h(X^1, Y^1)|X^0 = x, Y^0 = y) \leq \alpha^{-1} h(x, y), \forall (x, y) \notin R \times R. \tag{30}$$

Then

$$E(\alpha^{r_1}) \leq E(h(X^0, Y^0)), \tag{31}$$

and for $i > 1$ and any choice of $r_1, ..., r_{i-1}$,

$$E(\alpha^{r_i} | r_1, ..., r_{i-1}) \leq \alpha^{k_0} \sup_{(x,y) \in R \times R} E(h(X^1, Y^1)|X^0 = x, Y^0 = y). \tag{32}$$

Theorem 2.14. Suppose that Markov chain $P(x, dy)$ satisfies Minorization condition and satisfies hypotheses of the above proposition. Set

$$A = \sup_{(x,y) \in R \times R} E(h(X^1, Y^1)|X^0 = x, Y^0 = y)$$

. Then, with initial distribution v,

$$||L(X^k) - \pi||_{var} \leq (1-\epsilon)^{[j/k_0]} + \alpha^{-k+(j-1)k_0} A^{j-1} E_{v\times\pi}(h(X^0, Y^0)). \tag{33}$$

Remarks: The above can be shown iteratively with the total expectation formula. The application of the result above involves finding a small set as well as associated ϵ and all of the following components: α, A, and $E_{v\times\pi}(h(X^0, Y^0))$. The propositions in last section may be used to find small set and associated ϵ. As to the rest, we may use the "drift condition" to simplify analysis (See below).

Theorem 2.15. Suppose a Markov chain $P(x, dy)$ on a state space X satisfies the drift condition

$$E(V(X^1)|X^0 = x) \leq \lambda V(x) + b, x \in X \tag{34}$$

for some $V : X \to R^{\geq 0}$, and some $\lambda < 1$ and $b < \infty$; and further satisfies a minorization condition

$$P(x, \cdot) \geq \epsilon Q(\cdot), \forall x \in X, V(x) \leq d, \tag{35}$$

for some $\epsilon > 0$, some probability measure $Q(\cdot)$ on X, and some $d > \frac{2b}{1-\lambda}$. Then for any $0 < r < 1$, beginning in the initial distribution v, we have

$$||L(X^k) - \pi|| \leq (1-\epsilon)^{rk} + (\alpha^{-(1-r)} A^r)^k (1 + \frac{b}{1-\lambda} + E_v(V(X_0))), \tag{36}$$

where

$$\alpha^{-1} = \frac{1 + 2b + \lambda d}{1 + d} < 1; \tag{37}$$

$$A = 1 + 2(\lambda d + b) \tag{38}$$

Remarks: This is an extension of 2.14 by setting $h(x, y) = 1 + V(x) + V(y), R = \{x \in X|V(x) \leq d\}$.

2.7.2 Example: Bivariate Normal Model

Given a Bivariate normal distribution that has common mean μ and covariance matrix $\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ for which the conditional distributions are given by

$$\mathcal{L}(X_1|X_2 = x) = N(x, 1),$$

$$\mathcal{L}(X_2|X_1 = x) = N(\frac{x+\mu}{2}, 1/2)$$

Suppose we use a Gibbs sampler to sample this distribution. We may use the method given in previous section to acquire a quantitative exponential bound on total variation distance. Notice that for this example, we use the original method developed in Theorem 5, (Rosenthal, 1995). It differs from the simplified version slightly.

Step 1. Auxiliary Function
Notice that due to the special structure of the sequential Gibbs sampler, we may choose auxiliary functions that simply disregard the first coordinate: since the update scheme is as the following (12)(12)(12)..., the expectation of the auxiliary function can be acquired without knowing the previous value of the first coordinate. Another important factor to consider when screening for auxiliary function is its stability with respect to the chain, that is, it "tends to" admit a smaller value. Intuitively, for points that are far away from modal points of the target distribution, it is more likely to move closer to the modal points at the next step. Therefore, given that the small set chosen does include modal points, it is preferable to elect auxiliary functions that admit a small value at those modal points. Notice that in the simplified version of Theorem 5 (Theorem 12 in (Rosenthal, 1995)), the small set is chosen to be $\{x : h(x) \geq d\}$ directly.

Taking in to consideration of these factors, we may choose the following auxiliary function for this particular example:

$$h(x, y) = 1 + (x_2 - \mu)^2 + (y_2 - \mu)^2.$$

Step 2. Small Set and Drift Condition
Choose the following set to be the perspective small set:

$$R = \{x \in X | (x_2 - \mu)^2 \geq 3\}$$

Compute ϵ using Proposition 2.10:

$$\epsilon = \int (\inf_{x \in R} N(\frac{x_2 + \mu}{2}, 3/4; y))dy$$

$$= \int_{-\infty}^{0} N(\sqrt{3}/2, 3/4; y)dy + \int_{0}^{\infty} N(-\sqrt{3}/2, 3/4; y)dy > 0$$

For this choice of auxiliary function, using the variance-expectation formula, we have

$$E(h(X^{(1)}, Y^{(1)}) | x_2, y_2) = 9/4 + (1/4)h(x, y), \forall x, y,$$

which implies

$$E(h(X^{(1)}, Y^{(1)}) | x_2, y_2) \leq (13/16)h(x, y), \forall x, y \notin R \times R.$$

The condition we will verify here is not exactly the drift condition–it is a similar condition developed by (Rosenthal, 1995).

To continue, we note that

$$A = \sup_{(x,y) \in R \times R} E(h(X^{(1)}, Y^{(1)}) | x, y) = 9/4 + (1/4)(1 + 3 + 3) = 4$$

and knowing that $Y_2 \sim N(\mu, 1)$,

$$E_{v\times\pi}(h(X^{(0),Y^{(0)}}) = 1 + E_\pi((y_2 - \mu))^2) + E_v(x_2 - \mu)^2 = 2 + E_v(x_2 - \mu)$$

2.7.3 Example: Quantitative Rates of the James-Stein estimator

We will go through a realistic example of applying Theorem 2.15 to a hierarchical Bayesian sampling problem. The procedure can be broken down to multiple steps.

Step 1: Set up the Gibbs Sampler

Firstly, we shall define the Bayesian model:

$$Y_i|\theta_i \sim N(\theta_i, V), 1 \le i \le K$$
$$\theta_i|\mu, A \sim N(\mu, A), 1 \le i \le K$$
$$\mu \sim \text{flat prior on } R$$
$$A \sim IG(a, b)$$

This model defines a procedure by which the K observations are "generated". In the context of Bayesian probability, the observations $\{Y_i\}$ are known and we are to find and sample from the "posterior" distribution of parameters μ and $\{\theta_i\}$. It is quite standard to work with such "hierarchical" models. Informally, starting with $\{Y_i\}$, μ and A (the last two are sampled from prior), we just need to write conditional posterior distribution $P(\theta_i|Y_i)P(\theta_i)$:

$$\theta_i^{(k)} \sim L(\theta_i|A = A^{(k)}, \mu = \mu^{(k-1)}, Y_i) = N(\frac{\mu^{(k)}V + Y_i A^{(k)}}{V + A^{(k)}}, \frac{A^{(k)}V}{V + A^{(k)}})$$

Notice that this formula is quite standard (prior being normal with parameters (μ, A) and likelihood being also normal with only one data point Y_i). See these papers for derivation of such formulae: (Murphy, 2007), (Jordan, 2010). Similarly, we are able to write the following conditional distributions:

$$A^{(k)} \sim L(A|\theta_i = \theta_i^{(k-1)}, Y_i) = IG((a + \frac{K-1}{2}), b + \frac{1}{2}\sum(\theta_i^{(k-1)} - \bar{\theta}^{(k-1)})^2);$$

$$\mu^{(k)} \sim L(\mu|A = A^{(k)}, \theta_i = \theta_i^{(k-1)}, Y_i) = N(\bar{\theta}^{(k-1)}, \frac{A^{(k)}}{K})$$

With these, we are able to use a Gibbs sampler to sample parameters μ and $\{\theta_i\}$ in an iterative manner. We hope that the sampling results will approximate true posterior distribution:

$$\pi(\cdot) = L(A, \mu, \theta_1, ..., \theta_K|Y_1, ..., Y_K)$$

Step 2: Find appropriate auxiliary function $f(\cdot)$

The goal here is to derive the drift condition:

$$E(f(x^{(k)})|x^{(k-1)} = x) \leq \lambda f(x) + \Lambda,$$

for some $0 < \lambda < 1, \Lambda > 0, f : \mathcal{X} \to R^+$.

Some observations:

1. When searching for a proper auxiliary function, not only is it necessary to satisfy the drift condition, we must also keep λ, Λ as small as possible so that the Minorization condition is satisfied on $f_d : \{x \in \mathcal{X}|f(x) \leq d\}$: intuitively, the larger f_d, less "likely" will Minorization condition be satisfied.
2. Another factor that affects size of ϵ and satisfaction of Minorisation condition is whether transition probability within f_d is "close": the point is that if the transition probability is very disparate, their respective "minima" points tend to scatter around on $\mathcal{X}$, which would require ϵ to be very small for $\epsilon \cdot Q(\cdot)$ to be bounded below of all the "minima" points.
3. Based on the observations above, we thus conclude that the choice of auxiliary function f should have following properties: (1) if it is very large at one iteration, it tends to get smaller at the next; (2) all values of x for which $f(x)$ is small have similar transition probabilities for the next iteration.

For this problem, a good choice is

$$f(x) = f(A, \mu, \theta_1, ..., \theta_K) = \sum_{i=1}^{K}(\theta_i - \bar{Y})^2 = K(\bar{\theta} - \bar{Y})^2 + \sum_{i=1}^{K}(\theta_i - \bar{\theta})^2$$

We know that the samples of $\{\theta_i\}$ should be close to data points $\{Y_i\}$. Therefore, if $f(x)$ is large in one iteration (the values of θ_i are afar from $\bar{Y}$), we expect $f(x)$ to attain smaller value at the next iteration.

Step 3. Show drift condition is satisfied

This step involves evaluating the following:

$$E(\sum(\theta_i^{(k)} - \bar{\theta}^{(k)})^2|x^{(k-1)}), E(K(\bar{\theta} - \bar{Y})^2|x^{(k-1)})$$

Since we only know distribution of θ_i conditioned on $A^{(k)}, \mu^{(k)}, Y_i$, we have to use the "double expectation" to peel away θ_i

$$E(\sum(\theta_i^{(k)} - \bar{\theta}^{(k)})^2|A^{(k)}, \mu^{(k)}, x^{(k-1)})$$

This can be easily attained by using the following equality (this can be derived by exchanging summation and expectation and use variance formula):

$$E(\sum_{i=1}^{n}(Z_i - \bar{Z})^2) = (\frac{n-1}{n})\sum_{i=1}^{n} Var(Z_i) + \sum_{i=1}^{n}(E(Z_i) - E(\bar{Z}))^2$$

Apply this formula and variance-expectation formula, we can derive:

$$E(f(x^{(k)})|x^{(k-1)}, \mu^{(k)}, A^{(k)}) = E(\sum(\theta_i^{(k)} - \bar{\theta}^{(k)})^2|x^{(k-1)}, \mu^{(k)}, A^{(k)})$$

$$+E(K(\bar{\theta} - \bar{Y})^2|x^{(k-1)}, \mu^{(k)}, A^{(k)})$$

$$= (K-1)(\frac{A^{(k)}V}{V + A^{(k)}}) + (\frac{A^{(k)}}{V + A^{(k)}})^2\Delta$$

$$+K[\frac{A^{(k)}V}{K(v + A^{(k)})} + (\mu^{(k)} - \bar{Y})^2(\frac{V}{V + A^{(k)}})^2],$$

where $\Delta = \sum(Y_i - \bar{Y})^2$. We cam then take expected value over $\mu^{(k)}$,

$$E(f(x^{(k)})|x^{(k-1)}, A^{(k)}) = K(\frac{A^{(k)}V}{V + A^{(k)}}) + (\frac{A^{(k)}}{V + A^{(k)}})^2\Delta$$

$$+(K(\bar{\theta}^{(k-1)} - \bar{Y})^2 + A^{(k)})(\frac{V}{V + A^{(k)}})^2$$

Then we can identify that $K(\bar{\theta}^{(k-1)} - \bar{Y})^2 \leq f(x^{(k-1)})$ and simplify the expression further through other inequalities (detail not shown here):

$$E(f(x^{(k)})|x^{(k-1)}, A^{(k)}) \leq (1 + \frac{A^{(k)}}{V})^{-2} f(x^{(k-1)}) + (K + \frac{1}{4})V + \Delta$$

and take expectation over $A^{(k)}$, we then can obtain the drift condition:

$$E(f(x^{(k)})|x^{(k-1)}) \leq \lambda f(x) + \Lambda,$$

where

$$\lambda = E(1 + \frac{W}{V})^{-2} \text{ with } W \sim IG(a + \frac{K-1}{2}, b)$$

and $\Lambda = \Delta + (K + 1/4)V$, with $\Delta = \sum(Y_i - \bar{Y})^2$.

Step 4. Verify Minorization condition

The problem here is how to find a probability measure $Q(\cdot)$ and corresponding ϵ. One strategy used here is to define $Q(\cdot)$ so that it mimics transition probabilities $P(x, \cdot)$ (as defined by conditional distribution in Step 1) but with appropriate infimums over values of $x \in f_d$, i.e.,

$$Q'(dA) = (\inf_{0 \leq r \leq d} IG(a + \frac{K-1}{2}, b + \frac{r}{2}; A))dA$$

$$Q'(d\mu|A) = \left(\inf_{K(s-\bar{Y})^2 \leq d} N(s, \frac{A}{K}; \mu) \right) d\mu;$$

$$Q'(d\theta_i|\mu, A) = N(\frac{\mu V + Y_i A}{V + A}, \frac{AV}{V + A}),$$

The point of this construction is to ensure that

$$P(x, \cdot) \geq Q'(\cdot), x \in f_d$$

Intuitively, $Q(\cdot)$ differs from $P(x, \cdot)$ in that it is defined to "start" from the point in f_d to ensure the minimum value while $P(x, \cdot)$ can start from any point $x \in f_d$. This is why the infimum operator is not applied to $Q'(d\theta_i|\mu, A)$: it does not contain any θ_i.

To illustrate, suppose our "destination" is the following set: $(A, \mu, \theta) \in \mathbb{R} \times \mathcal{Y} \times \mathbb{R} \subseteq \mathcal{X}$,

$$Q(\mu \in \mathcal{Y}) = \int_{\mu \in \mathcal{Y}} \int_{A \in \mathbb{R}} Q'(dA) Q'(d\mu|A)$$

$$\leq \int_{\mu \in \mathcal{Y}} \int_{A \in \mathbb{R}} Q'(dA) P(x, d\mu) \leq \int_{\mu \in \mathcal{Y}} P(x, d\mu) = P(x, \mathcal{Y}), \forall x \in f_d$$

Therefore,

$$P(x, \cdot) \geq \epsilon Q(\cdot), \text{ where } Q(\cdot) = \frac{Q'(\cdot)}{Q'(\mathcal{X})}, \epsilon = Q'(\mathcal{X})$$

where

$$Q'(\mathcal{X}) = \int_0^\infty Q'(dA) \int_{-\infty}^\infty Q'(d\mu|A) \prod_{i=1}^K \int_{-\infty}^\infty Q'(d\theta_i|\mu, A)$$

$$= \int_0^\infty (\inf_{0 \leq r \leq d} IG(a + \frac{K-1}{2}, b + \frac{r}{2}; A)) dA \int_{-\infty}^\infty \left(\inf_{K(s-\bar{Y})^2 \leq d} N(s, \frac{A}{K}; \mu) \right) d\mu$$

Recall the following Proposition we have reviewed in previous section:

Proposition 2.29. Given a positive integer k_0 and subset $R \subseteq \mathcal{X}$, there exists a probability measure $Q(\cdot)$ such that

$$P^{k_0}(x, \cdot) \geq \epsilon Q(\cdot), \forall x \in R$$

where

$$\epsilon = \int_{\mathcal{X}} (\inf_{x \in R} P^{k_0}(x, dy))$$

The construction of $Q(\cdot)$ here is exactly to replicate this proposition in the hierarchical settings. Observe the integral expression of $Q'(\mathcal{X})$, it is actually just $\int_{\mathcal{X}}(\inf_{x\in R} P^{k_0}(x,dy))$ with $k_0 = 1$. The expression of ϵ can often be simplified further using the uni-modality of the probability distribution function (See also the bivariate normal example in (Rosenthal, 1995)). In this example, IG and normal distribution are both unimodal. Therefore, we can write

$$\inf_{0\le r\le d} IG(a+\frac{K-1}{2}, b+r/s; A) =$$

$$\min[IG(a+\frac{K-1}{2}, b; A), IG(a+\frac{K-1}{2}, b+d/2; A)].$$

$$\int_{-\infty}^{\infty}(\inf_{K(s-\bar{Y})^2\le d} N(s, A/K; \mu))d\mu = 2\int_0^{\infty} N(-\sqrt{d/K}, A/K; \mu)d\mu$$

In summary, to establish a quantitative bounds involves choosing an auxiliary function, computing its expectation (given starting point), finding appropriate constant λ, Λ so that drift condition is satisfied, finding $Q(\cdot)$ (often motivated by Proposition 2.29) for which f_d is a small set. Application of this “program” seems to follow a relatively standard procedure. With appropriate choice of auxiliary function, it is possible to derive a quantitative convergence rates for certain algorithms.

2.8 Complexity Bounds

2.8.1 Definitions and Useful results concerning continuous stochastic process

We include a short section on definitions and useful results necessary to understand the concept of *weak convergence of stochastic process*, particularly how it is motived by weak convergence of probability measures. The definition uses fundamental thoughts of the functional analysis: intuitively, each process is treated as a functional defined on "time" with certain common characteristics. Weak convergence of stochastic processes can therefore be defined by establishing a metric space on the set of such functionals. The results and definitions here are mostly taken from Chapter 3 of (Ethier and Kurtz, 2009).

We first develop weak convergence of probability measures.

Definition 2.23. Let $C(S)$ be the space of real-valued bounded continuous functions on the metric space (S, d) with norm $||f||= \sup_{x\in S}|f(x)|$. A sequence $\{P_n\} \subset \mathcal{P}(S)$ is said to converge weakly to $P \in \mathcal{P}(S)$ if

$$\lim_{n\to\infty} \int f dP_n = \int f dP, f \in C(S)$$

.

Recall that the distribution of an $S-$valued random variable X, denoted by PX^{-1}, is the element of $\mathcal{P}(S)$ given by $PX^{(-1)}(B) = P\{X \in B\}$. We may define convergence in distribution (covered in undergraduate probability) of a sequence of random variables $\{X_n\}$ accordingly by

$$\lim_{n\to\infty} E[f(X_n)] = E[f(X)].$$

We may denote weak convergence by $P_n \Rightarrow P$ and convergence in distribution by $X_n \Rightarrow X$.

Cadlag functions and Skorokhod space. As stated in introduction, to study convergence of a sequence of stochastic processes to another stochastic process, we must establish some metric space of which stochastic processes are the elements. Most stochastic processes arising in applications have the property that they have right nd left limits at each time point for almost every sample path. The convention is such that the sample paths are assumed to be actually right continuous without altering finite-dimensional distributions. This motivates the adoption of Cadlag functions (real-valued functions on $[0, \infty)$ that are right continuous and have left-hand limits) as elements of our metric space. Note that we do not merely use continuous functions so that "jump processes" can be accommodated in our theory. If we suppose that the processes (functionals) considered are defined on (E, r) metric space, we may denote this set as $D_E[0, \infty)$.

The most obvious way to define metrics for $D_E[0,\infty)$ is the following $||f||= \sup_{t\in[0,\infty)}|f(t)|$. This gives us a Banach space but the resulting metric space is non-separable, which causes well-known problems of measurability in the theory of weak convergence of measures on the space (Paulauskas, 2011). A definition that preserves separability and completeness is given in ($EthierandKurtz$, 2009) as the following:

$$d(x,y) = \inf_{\lambda\in\Lambda}\left[\gamma(\lambda) \vee \int_0^\infty e^{-\mu} d(x,y,\lambda,\mu)d\mu\right],$$

where

$$d(x,y,\lambda,\mu) = \sup_{t\geq 0} q(x(t\wedge\mu), y(\lambda(t)\wedge\mu)), q \equiv r\wedge 1$$

$$\gamma(\lambda) = \sup_{s>t\geq 0}|\log\frac{\lambda(s)-\lambda(t)}{s-t}|< \infty$$

Therefore, we may treat each stochastic process as a random variable and keep using the weak convergence definition of probability measures given above (thus the name weak convergence of stochastic processes). The following theorem is given in Chapter 3, 7.8 (a) of (Ethier and Kurtz, 2009) (we will not restate the proof here). It is illustrative of how weak convergence of stochastic processes leads to weak convergence of the values of processes at finitely many time points to those of the limiting process. We will use this theorem in (Roberts and Rosenthal, 2014).

Theorem 2.16. Let E be separable and let $X_n, n = 1, 2, \dots,$ and X be processes with sample paths in $D_E[0,\infty)$. If $X_n \Rightarrow X$, then

$$(X_n(t_1), \dots, t_k) \Rightarrow (X(t_1), ..., X(t_k))$$

for every finite set $\{t_1, ..., t_k\} \subset \{t \geq 0 : P\{X(t) = X(t-)\} = 1\}$.

Remarks. This theorem says that if a sequence of stochastic processes converges weakly to a limiting process, then upon fixing finitely many points in time, the corresponding values (vector of random variables indexed by n) converges in distribution ($n \to \infty$) to the corresponding values on the limiting process. Notably, this is only true for points where there is no jumps in limiting process X (almost surely). However, the proposition shows that there are not there are not "too many" of points that "admit jumps":

Proposition 2.30. If X is a process with sample paths in $D_E[0,\infty)$, then the complement in $[0,\infty)$ of

$$\{t \geq 0 : P\{X(t) = X(t-)\} = 1\}$$

is at most countable.

This also implies that we can always find some $t \geq 0$ such that the process is not jumping. This proposition will be used to fix a minor problem in the proof of the main theorem of (Roberts and Rosenthal, 2014).

Convergence Theorems for Feller processes I will not give a complete proof of these. However, the general methodology should be useful for us to understand relation between weak convergence of the processes and convergence of their *generators* and *semigroup.* We must first familiarize ourselves with some crucial concepts in stochastic process literature. The following definition is taken from the *Encyclopedia of Mathematics* which I find most illustrative of the intended "continuity" idea for Feller process. That said, there are other equivalent, more common definitions for Feller process (e.g. by defining Feller semigroup first).

Definition 2.24. Feller Process A homogeneous Markov process $X(t), t \in T$, where T is an additive sub-semi-group of the real axis $\mathbb{R}$, with values in a topological space E with a topology $\mathcal{C}$ and a Borel $\sigma-$algebra $\mathcal{B}$, the transition function $P(t, x, B), t \in T, x \in E, B \in \mathcal{B}$, of which has a certain property of smoothness, namely that for a continuous bounded function f the function

$$x \mapsto P^t f(x) = \int f(y) P(t, x, dy)$$

is continuous. We refer to the following set

$$\mathcal{P} = \{P^t : t \in T\}$$

as the Feller semigroup.

Remarks. This definition is actually easy to understand once we understood its purpose: to "qualify" for being a Feller process, the Markov chain must exhibit continuous behavior with respect to the starting point. Informally, a Feller process is such that if the starting point changes by a small amount, the distribution at any fixed time t will only deform by a very small amount.

We also introduce the concept of infinitesimal generator:

Definition 2.25. Let $\{P_t\}$ be a semigroup for a Markov process. The infinitesimal generator for the semigroup (and the process) is as follows:

$$Af = \lim_{t \downarrow 0} \frac{P^t f - f}{t}$$

for all $f \in B(E)$ for which this limit exists as a limit in $B(E)$.

Definition 2.26. Let A be a closed linear operator on a Banach space. A linear subspace $D \subseteq dom(A)$ is a core of A if the closure of A restricted to D is, again A.

Remark. The core of a generator is a rather technical concept related to functional analysis. To understand it, we make the following comments: (i) A closure of a linear operator A is *a closed extension* of A, which is essentially another linear operator B such that B is *closed* and maps a superset of $dom(A)$ to the same codomain while retaining the original mapping for elements in $dom(A)$; (ii) A linear operator A is closed if $\{f, g : f \in dom(A), g = Af\}$ is a closed set (A is a *closed map*); (iii) core of operator A is just a subset of $D \subseteq dom(A)$ such that D itself is closed under linear operations (linear subspace) and A restricted to D is closed. (iv) The idea of a core is that we can get away with knowing how the operator works on a linear subspace, which is often much easier to deal with, rather than controlling how it acts on its whole domain.

The following theorem is crucial to proving weak convergence of stochastic processes. It is taken from (Kallenberg, 2006) Chapter 17. A similar version without referring to Feller process explicitly can be found in (Ethier and Kurtz, 2009) Chapter 4 Section 8. I find the former more accessible.

Theorem 2.17. Let $X, X^1, X^2, ...$ be Feller processes with semigroups $(T_t), (T_{1,t}), (T_{2,t}), ...$ and generators $A, A_1, A_2, ...$, and fix a core D for A. Then these conditions are equivalent:
(i) If $f \in D$, there exists some $f_n \in dom(A_n)$ with $f_n \to f$ and $A_n f_n \to Af$;
(ii) $T_{n,t} \to T_t$ strongly for each $t > 0$;
(iii) $T_{n,t} f \to T_t f$ for each $f \in C_0$, uniformly for bounded $t > 0$;
(iv) If $X_0^n \Rightarrow X_0$, then $X^n \Rightarrow X$.

The theorem above does not apply to discrete-time Markov chains (viewed in continuous time) since they are not time-homogeneous (think of the time points where transition occur and time in between). Therefore, this theorem is not directly applicable to prove say weak convergence of discrete random walk to Wiener process. The following theorem amends this deficiency:

Theorem 2.18. Let $Y^1, Y2, ...$ be discrete Markov chains with transition operators $U_1, U_2, ...$, and consider a Feller process X in S with semigroup (T_t) and generator A. Fix a core D for A, and assume that $0 < h_n \to 0$. Then condition (i) through (iv) of the previous theorem remain equivalent for the operators and processes:

$$A_n = h_n^{-1}(U_n - I),\ T_{n,t} = U_n^{[t/h_n]},\ X_t^n = Y_{[t/h_n]}^n.$$

Remark. This theorem essentially says that the Feller process requirement for X_n can be relaxed if X_n can be shown to be a sped-up process (infinitely fast as $n \to \infty$) of some other discrete process Y_n.

Notice that the requirement for X to be Feller has not been relaxed in the theorem above. The following theorem is very useful in practice.

Theorem 2.19. If a process satisfies a stochastic differential equation of the form:

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t,$$

we refer to it as an *Ito diffusion.*

An Ito diffusion X is a continuous Feller process

These theorems provide means to prove weak convergence of stochastic processes. In particular, with (i), we may prove weak convergence of processes by proving convergence of generators–as we will see in next section, (Roberts et al., 1997) sets $f_n := V$ and showed $A_n V \to AV$.

Example: Functional Central Limit Theorem. This well-know theorem states that

$$Y_n := \frac{1}{n^{1/2}} \sum_{i=0}^{[nt]} X_i \Rightarrow W$$

where $X_1, X_2, \ldots$ is a sequence of iid, standard-normal random variables and W is the Wiener process. From Theorem 2.19, we know that W is Feller and the discrete process Y_n certainly satisfies requirement of Theorem 2.18. The central limit theorem implies that

$$\frac{1}{n^{1/2}} \sum_{i=0}^{[nt]} X_i \Rightarrow \sqrt{(t)}Z, Z \sim \mathcal{N}(0,1).$$

Therefore, by definition of weak convergence of random variables:

$$\mathbb{E}[f(y + \frac{1}{n^{1/2}} \sum_{i=0}^{[nt]} X_i)] \to \mathbb{E}[f(y + \sqrt{t}Z)], \forall f \in C_0.$$

Therefore, (iii) in Theorem 2.17 is satisfied. Functional Central Limit Theorem follows.

2.8.2 Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms

This section summarizes (Roberts et al., 1997)'s result that under certain conditions the Markov process associated with the MH algorithm converges to a Langevin diffusion as the dimension of the target densities goes to infinity and the process is sped up proportionally. The set-up of this asymptotic problem is as the following:

The target distribution is of the following form:

$$\pi_n(x^n) = \prod_{i=1}^{n} f(x_i^n)$$

where f is assumed to be Lipschitz continuous and

$$E_f[(\frac{f'(X)}{f(X)})^8] \equiv M < \infty,$$

$$E_f[(\frac{f''(X)}{f(X)})^4] < \infty.$$

The proposal distribution is supposed to be a multivariate Gaussian distribution as the following:

$$q_n(x^n, y^n) = \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp(\frac{-1}{2\sigma_n^2}|y^n - x^n|^2)$$

The the main theorem states:

Theorem 2.20. Suppose f is a real-valued function with continuous second derivative and satisfies the conditions outlined in the set-up. Let $\sigma_n^2 = l^2/(n-1)$ and $U_t^n := X_{[nt],1}^n$ where $X_{[nt],1}^n$ denotes the first-dimension component of the sped-up (by $n-$times for setting of n dimension) MH process. Assume that $X_{0,j}^i = X_{0,j}^j \forall i \leq j$. Then, as $n \to \infty$,

$$U^n \Rightarrow U$$

where $\Rightarrow$ denotes weak convergence of processes in the Skorokhod topology and U_0 is distributed according to f and U satisfies the Langevin SDE

$$dU_t = (h(l))^{1/2} dB_t + h(l)\frac{f'(U_t)}{2f(U_t)}dt$$

and

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

with Φ being the standard normal cumulative cdf and

$$I \equiv E_f[(\frac{f'(X)}{f(X)})^2]$$

Remarks. Here I will discuss the methods of proof presented in (Roberts et al., 1997). I will focus on illustrating the thought process and intuition of the proof–rather than just replicate the algebraic derivations in the original paper. I also corrected several typos in the original paper. The aim is to prove weak convergence of a sequence of stochastic process $U^n \to U$. Refer to (Ethier and Kurtz, 2009) for relevant theoretic background on infinitesimal generators, convergence etc..

To show that $U^n \Rightarrow U$, we need to show that the (discrete-time) generator of X^n,

$$G_n V(x^n) = nE[(V(Y^n) - V(x^n))(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})],$$

converges uniformly to the generator of the limiting Langevin diffusion, for suitably large class of real-valued function V (V is function of the first component only), where

$$GV(x) := \lim_{t\to 0} \frac{E^x[V(X_t)] - f(x)}{t}$$
$$= h(l)[\frac{1}{2}V^{''}(x) + \frac{1}{2}\frac{d}{dx}(\log f)(x)V'(x)].$$

Using total expectation law,

$$G_nV(x^n) = nE_{Y_1}[(V(Y^n) - V(x^n))E[(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})|Y_1]]$$

The key to the proof is to apply Taylor expansion: observe the expressions of $GV(x)$ and $G_nV(x^n)$–the latter essentially describes expectation of "change" in derivatives form of $V, \log(f)$ (in infinitesimal time) and the former describes change "directly" as expectation of difference of two discrete steps, which vanishes as n increases (the variance of per step is scaled to $1/n$).

Therefore, we would like to express $E[(V(Y^n) - V(x^n))(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})]$ in some polynomial form of, say, $(1/n)^k, k \in N$. Note that we already know that $E[(Y_1 - x_1)^2]$ is scaled to $1/n$, $E(Y_1 - x_1) = 0$ and higher moments $E(|Y_1 - x_1|^p), \forall p > 2$ converges to 0 faster than $1/n$. Therefore, intuitively, we will be able to show the convergence results so long as we can find the expansion of $(V(Y^n) - V(x^n))E[(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})|Y_1]$, i.e.,

$$(V(Y^n) - V(x^n))E(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)}|Y_1)$$
$$= A_1(n, x_1)(Y_1 - x_1) + A_2(n, x_1)(Y_1 - x_1)^2 + A_3(n, x_1)(Y_1 - x_1)^3 + ...$$

One way to obtain this expression is to expand $(V(Y^n) - V(x^n))$ and $E[(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})|Y_1]$ separately and evaluate the product.

The inner expectation can be evaluated as the following:

$$E_{Y_1,x_i,n} = E[(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})|Y_1] = E[1 \wedge \exp\left(\epsilon(Y_1) + \sum_{i=2}^{n}(\log f(Y_i) - \log f(x_i))\right)]$$

where $\epsilon(Y_1) := \log(\frac{f(Y_1)}{f(x_1)})$.

As stated above, our strategy is to expand $E[(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})|Y_1]$, which would require evaluating explicit expression of the expectation. Let's first expand the expression inside of the expectation:

$$E_{Y_1,x_i,n} = E[(1 \wedge \frac{\pi_n(Y^n)}{\pi_n(x^n)})|Y_1]$$

$$= E[1 \wedge \exp\Bigg(\epsilon(Y_1) + \sum_{i=2}^{n}[(\log f(x_i))'(Y_i - x_i) + 1/2(\log f(x_i))''(Y_i - x_i)^2$$

$$+1/6(\log f(Z_i))'''(Y_i - x_i)^3]\Bigg)]$$

Let's write the above expression as $E(1 \wedge e^A)$ where A is the only random component. We will need to first find density function of A and evaluate the integral on $A < 0$ and $A \geq 0$. Here we know $(Y_i - x_i)$ terms are Gaussian, but A also involves higher-order terms that could complicate the calculation. The intuition here is that if we may somehow discard higher order terms, then we may immediately apply the following lemma to evaluate the expectation is the following: Suppose $A \sim N(\mu, \sigma^2)$,

$$\mathbb{E}(1 \wedge e^A) = \Phi(\mu/\sigma) + \exp(\mu + \sigma^2/2)\Phi(-\sigma - \mu/\sigma)$$

Therefore, we want to find some "auxiliary sequence" $W_{Y_1,x^n,n}$, which does not contain higher order terms, such that

$$\sup_{x^n,Y_1} |E_{Y_1,x^n,n} - W_{Y_1,x^n,n}| \to 0 \; as \; n \to \infty \tag{39}$$

With this sequence, since

$$\sup_{x^n} |G_n V - nE[(V(Y_1) - V(x_1))W_{Y_1,x^n,n}]|$$

$$\leq \sup_{x^n} n\mathbb{E}_{Y_1}[|(V(Y_1) - V(x_1)) \cdot (E_{Y_1,x^n,n} - W_{Y_1,x^n,n})|]$$

$$\leq \sup_{Y_1,x^n} |E_{Y_1,x^n,n} - W_{Y_1,x^n,n}| \cdot \sup_{x^n} |n\mathbb{E}_{Y_1}[(V(Y_1) - V(x_1))]| \to 0 \; as \; n \to \infty$$

we only need to prove that

$$nE[(V(Y_1) - V(x_1)) \cdot W_{Y_1,x^n,n}] \to GV(x) \; as \; n \to \infty$$

Here the auxiliary sequence can be chosen to be the following:

$$W_{Y_1,x^n,n} := \mathbb{E}[1 \wedge \exp\Bigg(\epsilon(Y_1) + \sum_{i=2}^{n}[(\log f(x_i))'(Y_i - x_i) - \frac{l^2}{2(n-1)}((\log f(x_i))')^2]\Bigg)|Y_1]$$

As we have noted above, this choice is mainly motivated by the need to rid of second order and third terms. Now we need to show (39). First notice that the function $g(x) = 1 \wedge e^x$ is Lipschitz with coefficient 1, i.e.

$$|g(x) - g(y)| \leq |x - y|, \forall x, y \in \mathbb{R}$$

And note that absolute moments of normal random variable X are

$$E[|X|^p] = \sigma^p \cdot \frac{2^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}$$

Since by our construction $\sigma \sim O(1/n)$, the third order terms will just vanish as $n \to \infty$.

To show that (39), we only need to show that

$$\sup_{x^n} \mathbb{E}\bigg|\sum_{i=2}^{n}[\frac{(\log f(x_i))''}{2}(Y_i - x_i)^2 + \frac{l^2}{2(n-1)}(\log f(x_i))^{'2}]\bigg| \to 0 \; as \; n \to \infty$$

Rearrange the expression, using the fact that Y_i are independent for $i = 2, ..., n$:

$$\begin{aligned}
&\bigg\{\mathbb{E}\bigg|\sum_{i=2}^{n}[\frac{(\log f(x_i))''}{2}(Y_i - x_i)^2 + \frac{l^2}{2(n-1)}(\log f(x_i))^{'2}]\bigg|\bigg\}^2 \\
&\le \mathbb{E}\bigg\{\bigg(\sum_{i=2}^{n}[\frac{(\log f(x_i))''}{2}(Y_i - x_i)^2 + \frac{l^2}{2(n-1)}(\log f(x_i))^{'2}]\bigg)^2\bigg\} \\
&= Var\bigg[\sum_{i=2}^{n}[\frac{(\log f(x_i))''}{2}(Y_i - x_i)^2 + \frac{l^2}{2(n-1)}(\log f(x_i))^{'2}]\bigg] \\
&+\bigg(\mathbb{E}\bigg[\sum_{i=2}^{n}[\frac{(\log f(x_i))''}{2}(Y_i - x_i)^2 + \frac{l^2}{2(n-1)}(\log f(x_i))^{'2}]\bigg]\bigg)^2 \\
&= \frac{Var((Y_i - x_i)^2)}{4}\sum_{i=2}^{n}\bigg((\log f(x_i))''\bigg)^2 \\
&+\bigg(\sum_{i=2}^{n}[\frac{(\log f(x_i))''}{2}Var(Y_i - x_i) + \frac{l^2}{2(n-1)}(\log f(x_i))^{'2}]\bigg)^2 \\
&= \frac{l^4}{4(n-1)^2}\sum_{i=2}^{n}\bigg((\log f(x_i))''\bigg)^2 \\
&+\frac{l^4}{4(n-1)^2}\bigg(\sum_{i=2}^{n}[(\log f(x_i))'' + (\log f(x_i))^{'2}]\bigg)^2
\end{aligned}$$

It seems that there is a typo in (Roberts et al., 1997) where they omitted l^4 but this will not affect the proof. Since $(\log f)''$ is bounded, the first term will tend to 0. We now can focus on the second term.

For the ease of notation,

$$R_n(x_2, ..., x_n) = \frac{1}{n-1}\sum_{i=2}^{n}[(\log f(x_i))']^2$$

and

$$S_n(x_2, ..., x_n) = \frac{-1}{n-1} \sum_{i=2}^{n} [(\log f(x_i))''].$$

The remaining arguments rely on the fact that as $n \to \infty$, we are able to find a “tightening” set F_n in which R_n and S_n are bounded closer and closer as n increases whilst the probability of the process falls outside of the set tends to 0.

We can define F_n as the following:

$$F_n = \{|R_n(x_2, ..., x_n) - I| < n^{-1/8}\} \cap \{|S_n(x_2, ..., x_n) - I| < n^{-1/8}\}$$

Here R_n and S_n are bounded closer and closer to I and the bound is $n^{-1/8}$. For $x^n \in F_n$, using triangular inequality,

$$\left| \sum_{i=1}^{n} \frac{(\log f(x_i))'' + ((\log f(x_i))')^2}{2(n-1)} \right|$$

$$\leq |1/2R_n - 1/2I| + |1/2I - 1/2S_n| \to 0 \; as \; n \to \infty$$

Therefore, it is necessary to show the following

$$\forall t, P[Z_s^n \in F_n, 0 \leq s \leq t] \to 1 \; as \; n \to \infty.$$

That is, the process will be “contained” within F_n as $n \to \infty$.

To show this, we utilize the stationarity of the process to π_n and the assumption of the initial distribution, i.e. $Z_0^n \sim \pi_n$ and $Z_s^n \sim \pi_n, 0 \leq s \leq t$. Another observation is that since

$$\mathbb{E}[R_n] = \mathbb{E}[(\frac{f'(x_i)}{f(x_i)})^2] = I$$

we may apply weak law of large numbers: for all $\epsilon > 0$,

$$P_{\pi_n}[|R_n(Z) - T| > \epsilon] \to 0 \; as \; n \to \infty.$$

Here we may choose $\epsilon := n^{-1/8}$, by extended Markov's inequality (for monotonically increasing functions) and the assumption of f,

$$P_{\pi_n}[Z \notin F_n] = P_{\pi_n}[|R_n(Z) - I| > n^{-1/8}] \leq \frac{\mathbb{E}_{\pi_n}[(R_n(Z) - I)^4]}{(n^{-1/8})^4}$$

$$= \mathbb{E}_{\pi_n}[(R_n(Z) - I)^4] n^{1/2} \leq \frac{3M}{(n-1)^{3/2}}$$

It follows that (using $P(A \cup B) = P(A) + P(B) - P(A \cap B)$),

$$P[Z_s^n \notin F_n, \; for \; some \; 0 \leq x \leq t] \leq tnP_{\pi_n}[Z \notin F_n] \to 0 \; as \; n \to \infty$$

. The case for S_n is analogous.

Now that we proved 39, we will proceed to show that the auxiliary process $W_{Y_1,x^n,n}$ indeed converges to infinitesimal generator GV as $n \to \infty$.

$W_{Y_1,x^n,n}$ can be evaluated to be

$$W_{Y_1,x^n,n} = \Phi(R_n^{-1/2}(l^{-1}\epsilon(Y_1) - lR_n/2))$$

$$+\exp(\epsilon(Y_1))\Phi(-\frac{lR_n^{1/2}}{2} - \epsilon(Y_1)R_n^{-1/2}l^{-1}) :\equiv M(\epsilon).$$

Therefore, we only need to prove convergence of the term

$$n\mathbb{E}[(V(Y_1) - V(x_1))M(\log \frac{f(Y_1)}{f(x_1)})].$$

Apply Taylor series expansion:

$$(V(Y_1) - V(x_1))M(\log \frac{f(Y_1)}{f(x_1)})$$

$$= (V'(x_1)(Y_1 - x_1) + 1/2V''(x_1)(Y_1 - x_1)^2 + \frac{V'''(Z_1)}{6}(Y_1 - x_1)^3)$$

$$\times[M(0) + (Y_1 - x_1)M'(0)(\log f(x_1))' + \frac{1}{2}(Y_1 - x_1)^2 T(x_1, W_1)].$$

where $T(x_1, W_1)$ denotes the Taylor remainder term and Z_1, W_1 are in between x_1, Y_1.

We notice that in the expression above, all terms with $(Y_1 - x_1)^k, k \neq 2$ vanish as $n \to \infty$ after taking expectation, i.e.

$$\mathbb{E}\left[n(V(Y_1) - V(x_1))M(\log \frac{f(Y_1)}{f(x_1)})\right]$$

$$= 2n\Phi(-\frac{R_n^{1/2}l}{2})[\frac{1}{2}V''(x_1) + \frac{1}{2}(\log f(x_1))'V'(x_1)] \cdot \mathbb{E}[(Y_1 - x_1)^2]$$

$$+\mathbb{E}[B(x_1, Y_1, n)] \to 0 \; as \; n \to \infty$$

where we used the fact $E(Y_1 - x_1) = 0$ and that $f^{(i)}(x), V^{(i)}(x), i = 1, 2, 3$ can be bounded by some K and the higher order terms converge to 0 as $n \to \infty$, i.e.

$$\mathbb{E}[|B(x_1, Y_1, n)|] \leq a_1(K)nE[|Y_1 - x_1|^3] + a_2(K)nE[|Y_1 - x_1|^4] + a_3(K)nE[|Y_1 - x_1|^5].$$

Therefore,

$$\sup_{x^n \in F_n} |G_n V(x) - GV(x)| \to 0 \; as \; n \to \infty$$

This concludes the proof. The following Corollary highlights the practical significance of the main theorem:

Corollary 2.1. Let $\alpha_n(l) = \int\int \pi_n(x^n)\alpha(x^n, y^n)q_n(x^n, y^n)dx^n dy^n$ be the average acceptance rate of random walk Metropolis algorithm in n dimensions, and let

$$\alpha(l) = 2\Phi(-\frac{l\sqrt{I}}{2})$$

.

We then have
(i) $\lim_{n\to\infty} \alpha_n(l) = \alpha(l)$
(ii) $h(l)$ is maximized (to two decimal places) by

$$l = \hat{l} = \frac{2.38}{\sqrt{I}}$$

Also

$$\alpha(\hat{l}) = 0.23$$

and

$$h(\hat{l}) = 1.3/I.$$

Remark. The average acceptance rate defined above can be understood intuitively as average of $\alpha(x^n, y^n)$, the acceptance rate from x^n to y^n, weighted over each occurrence of the chain (initial distribution being stationarity π) reaching y^n in one step. $h(l)$ is sometimes referred to as the speed measure of the limiting diffusion process. Therefore, we may interpret $\hat{l}$ as l that maximizes the limiting diffusion's convergence speed to stationarity. Under this $\hat{l}$, the asymptotic average rate of acceptance is 0.23. Therefore, an approximation of optimal avg. rate of acceptance can be taken to be 0.23 for finite RWM algorithms.

2.8.3 Complexity Bounds via Diffusion Limits

In the last section, we established the result that under certain assumptions, a random-walk Metropolis-Hasting chain (its the first coordinate to be precise) converges weakly (in the usual Skorokhod topology)to a limiting ergodic Langevin diffusion as $d \to \infty$ where d is both the dimension and the factor by which we speed up the Markov chain.

Let Lip_1^1 denote all Lipschitz functions with Lipschitz constant ≤ 1 and with $|f(x)|\leq 1, \forall x \in \mathcal{X}$:

$$Lip_1^1 = \{f : \mathcal{X} \to \mathbb{R}, |f(x) - f(y)|\leq \rho(x, y)\forall x, y \in \mathcal{X}, |f|\leq 1\}$$

We can define "KR" distance function as the following:

$$||\mathcal{L}_x(X_t) - \pi||_{KR} := \sup_{f\in Lip_1^1} \left|\mathbb{E}_x[f(X_t)] - \pi(f)\right|.$$

Compare the "KR" distance function with usual total variation distance:

$$||\mathcal{L}_x(X_t) - \pi||_{TV} := \sup_A |P^{(t)}(x, A) - \pi(A)|$$

$$\equiv \sup_{|f|\leq 1} \left| \mathbb{E}_x[f(X_t)] - \pi(f) \right|.$$

where the equivalence is due to the fact that the maximum is achieved by choosing $f = 1$ on regions where density of $P^{(t)}$ is greater than π and $f = 0$ otherwise (or vice-versa). We notice that "KR" has a more limited set of function we may choose from (e.g. f has to be Lipschitz continuous). This means that

$$||\mathcal{L}_x(X_t) - \pi||_{KR} \leq ||\mathcal{L}_x(X_t) - \pi||_{TV}$$

The purpose of redefining a distance metrics is that TV is not suitable to bound weak convergence (it may not go to 0 if the process only converges weakly to stationarity). We will see how the Lipschitz condition is necessary in the proof.

Theorem 2.21. Let $X^{(d)}$ denote a sequence of stochastic processes. Suppose that this sequence of stochastic processes converges weakly in the Skorokhod topology to another stochastic process $X^{(\infty)}$ as $d \to \infty$. Assume that $X^{(d)}$ has stationary distribution π for all d and $X^{(\infty)}$ converges to π. Then for any $\epsilon > 0$, there are $D < \infty$ and $T < \infty$ such that

$$\mathbb{E}_{X_0^{(d)} \sim \pi} ||\mathcal{L}_{X_0^{(d)}}(X_t^{(d)}) - \pi||_{KR} < \epsilon, t \geq T, d \geq D.$$

Remark. To see why this is true. We must first establish that $||\ldots||_{KR}$ is indeed a norm: (i) $||0|| = 0$; (ii) $||a\mu|| = a||\mu||$ (iii)$||-\mu|| = ||\mu||$ ($f \in Lip_1^1 \Leftrightarrow -f \in Lip_1^1$) (iv) the triangular inequality

$$||\mu + \nu|| = \sup_{f \in Lip} (\mu(f) + \nu(f)) \leq \sup_{f \in Lip} (\mu(f)) + \sup_{f \in Lip} (\nu(f)) = ||\mu|| + ||\nu||$$

The following proposition is critical to the proof of the main theorem:

Proposition 2.31. The metric $\delta := ||\mu - \nu||_{KR}$ metrises weak convergence of probability measures on $(\mathcal{X}, \mathcal{F}, \rho)$, i.e.

$$\{\mu_t\} \Rightarrow \mu_t \text{ iff } \lim_{t \to \infty} \delta(\mu_t, \mu) = 0$$

From (Gibbs and Su, 2002), we learned the following proposition:

Proposition 2.32. Given state space $\mathbb{R}$ or any metric space, the Wasserstein metric is equivalent to our definition of KR, i.e. for distribution function μ, ν

$$d_W(\mu, \nu) := \sup\{ \left| \int h d\mu - \int h d\nu \right| : ||h||_L \leq 1\}$$

where the supremum is taken over all h satisfying the Lipschitz condition $|h(x) - h(y)| \leq \rho(x, y)$.

The Wassertein metric metrises weak convergence on spaces of bounded diameter.

(Roberts and Rosenthal, 2014) is able to conclude that "KR", i.e. the Wasserstein metric, in fact metrises weak convergence on all spaces regardless of boundedness by showing the following:

Proposition 2.33. Let $\rho^* = \min(\rho \wedge 2)$. Weak convergence on $(\mathcal{X}, \rho)$ is equivalent to weak convergence on $(\mathcal{X}, \rho^*)$. And $||\ldots||_{KR}$ is the same under both metrics.

This is easy to see once we realize that $|f(x) - f(y)| \leq 2$ for $f \in Lip_1^1$–this means that our set of $1-$Lipschitz functions remain unchanged upon switching metrics to ρ^*. It is odd why (Gibbs and Su, 2002) would include the bounded condition (and a few other papers too). As a direct consequence, we have the following proposition:

Proposition 2.34. If $X^{(\infty)}$ converges to π, either weakly or in total variation distance, the for all $\epsilon > 0$ there is $T < \infty$ such that

$$||\mathcal{L}_x(X_T^{(\infty)}) - \pi||_{KR} \leq \epsilon/2, \forall t \geq T$$

This is obvious because (i) total variance converges to 0 implies that weak convergence (though the reverse is not necessarily true) (ii) "KR" metrises weak convergence. Then we have the following proposition:

Proposition 2.35. Under the assumptions of the main theorem, for any $x \in \mathcal{X}$ and $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that

$$||\mathcal{L}_x(X_T^{(d)}) - \pi||_{KR} < \epsilon, d \geq D$$

Proof of this proposition as presented in (Roberts and Rosenthal, 2014) has a minor problem that there is no guarantee that $X_n(T) \Rightarrow X(T)$ for the T we find in 2.34. This problem can be fixed easily by applying Proposition (2.30) from the first section: from this proposition, we know that there exists some $T' > T$ such that

$$X_n(T') \Rightarrow X(T').$$

For if not, $[T, \infty)$ would be a subset of the complement of $\{t \geq 0, P\{X(t) = X(t-)\} = 1\}$, which would make it uncountable (thus contradiction). The rest of the proof is identical to the original. We apply triangular inequality as the following:

$$||\mathcal{L}_x(X_t^{(d)}) - \pi||_{KR} \leq ||\mathcal{L}_x(X_t^{(d)}) - \mathcal{L}_x(X_t^{(\infty)})||_{KR} + ||\mathcal{L}_x(X_t^{(\infty)}) - \pi||_{KR}.$$

Then we can just set $t := T'$. Since, "KR" metrices weak convergence, the results follows.

Remarks. We have shown that such D exists for a single point in time (which we know exists). However, we do not know if D exists for all points in $[T', \infty)$.

Next step is to show that for each $\epsilon > 0$, there exists $D < \infty$ and $T < \infty$ such that

$$\mathbb{E}_{X_0 \sim \pi} || \mathcal{L}_{X_0}(X_T^{(d)}) - \pi ||_{KR} < \epsilon, d \geq D.$$

In addition, we want to show that $\mathcal{E}_{X_0 \sim \pi} || \mathcal{L}_{X_0}(X_t^{(d)}) - \pi ||_{KR}$ is a non-increasing function of t:

$$\begin{aligned}
&\mathbb{E}_{X_0 \sim \pi} || P^{s+t}(X_0, \cdot) - \pi || \\
&= \mathbb{E}_{X_0} || \int_{y \in \mathcal{X}} P^s(X_0, dy) P^t(y, \cdot) - \pi || \\
&\leq \mathbb{E}_{X_0 \sim \pi} \int_{y \in \mathcal{X}} || P^s(X_0, dy) P^t(y, \cdot) - \pi \cdot P^s(X_0, dy) || \\
&\leq \mathbb{E}_{X_0 \sim \pi} \int_{y \in \mathcal{X}} P^s(X_0, dy) || P^t(y, \cdot) - \pi \cdot || \\
&= \mathbb{E}_{Y_0 \sim \pi} || P^t(Y_0, \cdot) - \pi ||
\end{aligned}$$

The third inequality step uses stationarity. This proves the claim. The main theorem follows.

As we can see, this theorem can be directly applied to the Random Walk Metropolis algorithm (by combining with the result in the previous section) to derive the following result:

Theorem 2.22. Let $Z^{(d)}$ be a $d-$dimension RWM algorithm satisfying the technical assumption we discussed in previous section. Then for any $\epsilon > 0$, there is $D < \infty$ and $T < \infty$ such that

$$\mathbb{E}_{Z_0^{(d)} \sim \pi} || \mathcal{L}_{Z_0^{(d)}}(Z_{\lfloor dt \rfloor, 1}^{(d)}) - \pi ||_{KR} < \epsilon, t \geq T, d \geq D$$

3 Part II: Adaptive Markov Chain Monte Carlo

This is the second part of my thesis study, which focuses on a special class of MCMC algorithm: adaptive Markov Chain Monte Carlo. The study mostly follows (Roberts and Rosenthal, 2007) and (Łatuszyński et al., 2013), though a set of other papers were reviewed related this topic: (Roberts and Rosenthal, 1997) provided results useful for analysis in (Łatuszyński et al., 2013); (Craiu et al., 2015) investigated whether bounded modifications of stable Markov chain remains stable based on methodology provided by (Roberts and Rosenthal, 2007); (Rosenthal and Yang, 2017) extended results in the previous paper to the "combo-continuous" case; (Roberts and Rosenthal, 2009) surveyed various adaptive MCMC algorithms and their empirical properties;

3.1 Ergodicity Conditions for General Adaptive Markov Chain Monte Carlo

We mostly follow (Roberts and Rosenthal, 2007) in this section.

3.1.1 Formal Setup

Let $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain kernels on $\mathcal{X}$. Each kernel should have $\pi(\cdot)$ as their stationary distribution:

$$(\pi P_\gamma)(\cdot) = \pi(\cdot)$$

We also assume that P_γ is $\phi-$irreducible and aperiodic so that P_γ is ergodic for $\pi(\cdot)$:

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

The reasoning behind adaptive MCMC is that since some γ may lead to far less (or more) efficient algorithms, the adaptive algorithm should choose γ by a $\mathcal{Y}-$valued random variable Γ_n updated at each step by specified rules. Let X_n represent state of the algorithm at time n on $\mathcal{X}$. Let

$$\mathcal{G} = \sigma(X_0, ..., X_n, \Gamma_0, ..., \Gamma_n)$$

be the filtration generated by (X_n, Γ_n). Thus, $P[X_{n+1} \in B|X_n = x, \Gamma_n = \gamma, \mathcal{G}_{n-1}] = P_\gamma(x, B)$. We also adopt the following notation:

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

Note that $A^{(n)}$ here represents *unconditional* distribution of the algorithm at step n with only initial state and kernel known. Finally, we let

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

denote the total variation distance between the distribution of the adaptive algorithm at time n and the target distribution $\pi(\cdot)$. The adaptive algorithm is referred to as *ergodic* if

$$\lim_{n\to\infty} T(x,\gamma,n) = 0, \forall x, \gamma$$

.

3.1.2 Uniform Converging Case

To motivate the main theorem, we state the following proposition which is a special case with only finite adaptation.

Proposition 3.1. A finite adaptation MCMC algorithm, in which each individual kernel P_γ is ergodic and adaptation stops after some finite steps τ, is ergodic.

The result can be proved by simply noting that $\lim_{n\to\infty}||P^n_{\Gamma_\tau}(X_\tau,\cdot)-\pi(\cdot)||=0$. This proposition provides a viable yet "safe" means to run adaptive MCMCs, i.e., there is an initial, trial stage with finite number of runs, within which the algorithm will explore different tunings to determine good parameter values. A second stage follows where the actual samples will be obtained to preserve asymptotic convergence.

The main results in this section therefore concern *dependent, infinite* adaptation schemes: the adaptations will continue to modify Γ_n based on $\mathcal{G}_{n-1}$. We will specify under which conditions the adaptive algorithm will still converge to target distribution $\pi(\cdot)$. In this section, we will start from more straightforward conditions: we will require the convergence to $\pi(\cdot)$ of the kernels P_γ to be uniformly bounded. This condition will be extended to more general case.

Theorem 3.1. Consider an adaptive MCMC algorithm on state space $\mathcal{X}$, with adaptation index $\mathcal{Y}$, so $\pi(\cdot)$ is stationary for each kernel $P_\gamma, \gamma \in \mathcal{Y}$. The adaptive algorithm is ergodic if both of the following conditions are satisfied:

(a) [Simultaneous Uniform Ergodicity] For all $\epsilon > 0$, there is $N = N(\epsilon) \in \mathbb{N}$ such that $||P^N_\gamma(x,\cdot)-\pi(\cdot)||\leq \epsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$; and

(b) [Diminishing Adaptation] Let

$$D_n = \sup_{x\in\mathcal{X}} ||P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)||$$

denote a $\mathcal{G}_{n+1}-$measurable random variable (depending on Γ_n and Γ_{n+1}). The Diminishing Adaptation condition requires

$$\lim_{n\to\infty} D_n = 0$$

in probability.

Proof. We will supply a detailed proof for this theorem since the methods involved are very important to the study of adaptive MCMC in general. Intuitively, the flow of the proof is as the following: using condition (b), we identify a sufficiently large step K the adaptation D_n is properly bounded with minimal probability to "escape" (convergence in probability); then we construct a second, auxiliary chain where kernel is fixed. For this chain, we know will converge to $\pi(\cdot)$ by (a), which corresponds to a fixed number of steps N after step K in our $\delta - \epsilon$ proof due to simultaneous uniformity; the rest of the proof involves coupling the original chain to this auxiliary chain and thus show that it indeed converges to the target distribution.

Fix $\epsilon > 0$. Let N be some positive integer that is larger than $N(\epsilon)$ as in (a). By condition (b), we can find $n* \in \mathbb{N}$ such that

$$P(\{D_n \geq \epsilon/N^2\}) \leq \epsilon/N, \forall n \geq n*$$

The coupling construction will start from $n*$, after which the adaptation is sufficiently small in a probabilistic sense. We also denote the "target time" with $K \geq n * +N$.

Let

$$E_n = \cap_{i=n+1}^{n+N}(\{D_i < \epsilon/N^2\})$$

Due to convergence in probability as in (2), we have $P(E_n) \geq 1 - \epsilon$ for $n \geq n*$. It can be easily seen with triangular inequality that on event E,

$$||P_{\Gamma_{n*}}(x, \cdot) - P_{\Gamma_m(x,\cdot)}|| < \epsilon/N, x \in \mathcal{X}, n* \leq m \leq K$$

The second chain we will now construct is just a Markov chain with fixed transition probability (kernel) P_{n*} and starts from the state X_{n*}. Let's denote this chain with X'_n. We claim that such chain exists so that on E

$$P[X'_i \neq X_i, n* \leq i \leq m] < [m - n*]\epsilon/N, n* \leq m \leq K$$

Indeed, this is direct result of *coupling inequality* and induction. In particular, this implies that at the target step,

$$P(X'_K \neq X_K, E) < \epsilon$$

Recall that with condition (a), for the auxiliary chain (with fixed kernel Γ_{n*}),

$$||P^N_{\Gamma_{n*}}(X_{n*}, \cdot) - \pi(\cdot)|| < \epsilon$$

Suppose $Z \sim \pi(\cdot)$. With coupling inequality again, $P[X'_K \neq Z] < \epsilon$. We can construct all of X_n, X'_n and Z on a common probability space, by first constructing X_n and X'_n as above and then constructing Z conditioning on them but with $Z \sim \pi(\cdot)$. We then have

$$P(X_K \neq Z) \leq P(X_K \neq X_K, E) + P(X'_K \neq Z, E) + P(E^c) < 3\epsilon$$

. Hence, with coupling inequality again,

$$T(x_0, \gamma_0, K) = ||\mathcal{L}(X_K) - \pi(\cdot)||< 3\epsilon$$

Since we may choose any arbitrarily large $N > N(\epsilon)$, the result follows.

□

Under diminishing adaptation, the process gradually tends to a certain kernel and under uniform ergodicity, with any such kernel the process will converge in a uniformly bounded manner: since there could be infinitely many possible kernels, without uniformity we may end up with an infinite sequence of kernels that, though each converges eventually, do not all converge to targeted distribution within given ϵ with any finite N.

The uniform ergodicity condition may be "substituted" with conditions that are easier to verify. We will therefore document some useful propositions regarding this.

Proposition 3.2. Suppose an adaptive MCMC satisfies diminishing adaptation and also that each kernel is ergodic for $\pi(\cdot)$.

(a) If $\mathcal{X}, \mathcal{Y}$ are finite. Then the adaptive MCMC is ergodic;
(b) If $\mathcal{X} \times \mathcal{Y}$ is compact in some topology with respect to which the mapping $(x, \gamma) \to T(x, \gamma, n)$ is continuous for each fixed $n \in \mathbb{N}$. The the adaptive algorithm is ergodic;

Remarks. (a) and (b) essentially eliminates the problem we just alluded to. In (a) there is simply only finite instead infinite possible kernels and each corresponds to only a finite collection of steps required to converge; (b) introduces "uniform boundedness" through the combined condition of continuity and compactness: though there are still infinite possible kernels, the maximal number of steps required to converge among all the kernels are bounded. It can be proved using the well-known result in analysis:

Let K be a compact subset of metric space M and $f : K \to \mathbb{R}$ a continuous function on K. Then f attains its maximum value at some point in K.

Both conditions can be used to verify the simultaneous uniform ergodicity condition for certain algorithms. A noteworthy application of this is for adaptive Metropolis algorithm proposed in (Haario et al., 2001), which adapts by using different proposal distribution $Q(\cdot)$. The following Proposition can be proved using (b):

Proposition 3.3. Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, P_γ represents a Metropolis-Hastings algorithm with proposal kernel $Q_\gamma(x, dy) = f_\gamma(x, y)\lambda(dy)$ having a density $f_\gamma(x,)$ with respect to some finite reference measure λ on $\mathcal{X}$, with corresponding density g for $\pi(\cdot)$ so that $\pi(dy) = g(y)\lambda(dy)$. Finally, suppose that the $f_\lambda(x, y)$ are uniformly bounded, and for each fixed $y \in \mathcal{X}$, the mapping $(x, \gamma) \to f_\gamma(x, y)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then the adaptive algorithm is ergodic.

3.1.3 Non-Uniformly Converging Case

As mentioned in last section, the uniform ergodicity condition can be further relaxed. Notice that in the proof, condition (a) was only used to show that the auxiliary chain was sufficiently close to $\pi(\cdot)$. As we will see shortly, a weaker condition can be used to ensure convergence. This weakened condition is referred to as the“containment condition”, which can be described as the following:

To construct the containment condition, define “ϵ convergence time function” $M_\epsilon : \mathcal{X} \times \mathcal{Y} \to \mathbb{N}$ by

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

$M_\epsilon(x, \gamma)$ has the intuitive interpretation as the number of steps required for a chain with kernel P_γ and initial state x to converge to $\pi(\cdot)$ *within $\epsilon-$distance.* Notice that if each kernel is ergodic, for any $\epsilon > 0$, $M_\epsilon(x, \gamma) < \infty$.

The containment condition essentially requires that the sequence $\{M_\epsilon(X_n, Y_n)\}^\infty_{n=0}$ is bounded in probability for all ϵ for any initial state space and kernel. If containment condition and diminishing adaptation condition are both satisfied, the process is ergodic. Formally, the non-uniform convergence theorem can be stated as following:

Theorem 3.2. Consider an adaptive MCMC algorithm that satisfies Diminishing Adaptation condition, i.e.

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} ||P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}|| = 0 \text{ in probability}$$

Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{Y}$. Then $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$ provided that for all $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \Gamma_n)\}^\infty_{n=0}$ is bounded in probability given $X_0 = x_*, \Gamma_0 = \gamma_*$, i.e. for all $\delta > 0$, there is $N \in \mathbb{N}$ such that

$$P[M_\epsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta, \forall n \in \mathbb{N}$$

Before stating the proof, we would like to remark that this relaxation of the simultaneous uniform ergodicity condition is just replacement of the original “definite” bound on convergence speed across all kernels with a “probabilistic” bound.

Therefore, we can see that the weaker condition is really analogous to the convergence case (convergence in probability) whereas the original, uniform bounds on convergence speed is analogous to sure convergence. Now we will state a proof where this point can be seen more clearly.

Proof. In the proof for the uniform converging case, condition (a) essentially requires that there exists $N \in \mathbb{N}$ for each $\epsilon > 0$ such that

$$\{\omega \in \Omega | \sup_{x_* \in \mathcal{X}, \gamma_* \in \mathcal{Y}} M_\epsilon(x_*, \gamma_*) \leq N\} = \Omega$$

Now since this condition is relaxed and it was used only in deriving $P(X'_K \neq Z, E) < \epsilon$ in the final step of the proof, we have

$$T(x_*, \gamma_*, n) < 3\epsilon + P(M_\epsilon(X_n, \Gamma_n) > N | X_0 = x_*, \Gamma_0 = \gamma_*).$$

Assuming that the containment condition is satisfied, we are able to find $m \in \mathbb{N}$ such that

$$P(M_\epsilon(X_n, \Gamma_n) > m | X_0 = x_*, \Gamma_0 = \gamma_*) \leq \epsilon, \forall n \in \mathbb{N}.$$

Then let $N \geq m$, we conclude that

$$T(x_*, \gamma_*, K) < 3\epsilon + \epsilon = 4\epsilon, K \geq n * +m.$$

□

3.1.4 Relation to Quantitative Convergence Rates

In part I, we have reviewed the theory of quantitative convergence bounds with (Rosenthal, 1995) and (Rosenthal, 1996). Upon satisfying the drift condition and the Minorization condition, a bound on convergence speed of the Markov process can be established. Therefore, it provides us yet another alternative to substitute the containment condition, which in itself mainly serves to bound convergence speed across all the kernels. First, let's reiterative a simplified version of the main theorem (Theorem 5) in (Rosenthal, 1995):

Theorem 3.3. Suppose a Markov chain $P(x, dy)$ on a state space X satisfies the Drift Condition

$$E(V(X^1)|X^0 = x) \leq \lambda V(x) + b, x \in \mathcal{X} \tag{40}$$

for some $V : \mathcal{X} \to R^{\geq 0}$, and some $\lambda < 1$ and $b < \infty$; and further satisfies a Minorization Condition

$$P(x, \cdot) \geq \epsilon Q(\cdot), \forall x \in \mathcal{X} \text{ such that } V(x) \leq d, \tag{41}$$

for some $\epsilon > 0$, some probability measure $Q(\cdot)$ on X, and some $d > \frac{2b}{1-\lambda}$. Then for any $0 < r < 1$, beginning in the initial distribution v, we have

$$||L(X^k) - \pi|| \leq (1-\epsilon)^{rk} + (\alpha^{-(1-r)} A^r)^k (1 + \frac{b}{1-\lambda} + E_v(V(X_0))), \tag{42}$$

where

$$\alpha^{-1} = \frac{1 + 2b + \lambda d}{1 + d} < 1; \tag{43}$$

$$A = 1 + 2(\lambda d + b) \tag{44}$$

We can also rewrite the bound above into a more compact form,

$$||P^n(x,\cdot) - \pi(\cdot)|| \leq KV(x)\rho^n, \forall\lambda \in \mathcal{Y},$$

where $K < \infty$ and $\rho < 1$, depending only on the constants ϵ, λ, b, d.

To apply this theorem to bound convergence speed across kernels, we consider a notion called *simultaneously strongly aperiodically geometric ergodicity.*

Definition 3.1. The family of kernels $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ are referred to as simultaneously strongly aperiodically geometrically ergodic if there is $V : \mathcal{X} \to R^{\geq 0}$, $\epsilon > 0$, $b, d < \infty$ such that
(i) for each $\gamma \in \mathcal{Y}$, there exists a probability measure $Q_\gamma(\cdot)$ on $\mathcal{X}$ with $P_\gamma(x,\cdot) \geq \epsilon Q_\gamma(\cdot)$ for all $x \in \mathcal{X}$ such that $V(x) \leq d$;
(ii)$P_\gamma V \leq \lambda V + b$.

Therefore, for an adaptive algorithm with kernels being simultaneously strongly aperiodically geometrically ergodic, there convergence speed can be bounded uniformly by $KV(x)\rho^n$. Therefore, the "$\epsilon-$convergence time" can also be bounded: Since

$$KV(X_n)\rho^N < \epsilon \to \sup_{\gamma_*\in\mathcal{Y}} M_\epsilon(X_n, \gamma_*) \leq N$$

we only need $\{V(X_n)\}$ to be bounded in probability to ensure that for all $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^\infty$ is bounded in probability. By Markov inequality, we also know that

$$\sup_n E(V(X_n)) < \infty \to \{V(X_n)\} \text{ is bounded in probability.}$$

Therefore, the problem is simplified to show $\sup_n E(V(X_n)) < \infty$. This can be easily derived via induction using the Drift condition (after integrating over distribution of Γ_n), double expectation formula and the fact that $E(V(X_0)) < \infty$ and $\lambda < 1$. Based on the reasoning above, we can state the following theorem:

Theorem 3.4. Consider an adaptive MCMC algorithm with Diminishing condition, such that the family of kernels $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ are simultaneously strongly aperiodically geometrically ergodic with $E(V(X_0)) < \infty$. Then the adaptive algorithm is ergodic.

3.2 Adaptive Gibbs Samplers and related MCMC methods

In this section, I will first review results and methods used in (Łatuszyński et al., 2013). (Łatuszyński et al., 2013) discussed ergodicity properties of a series of adaptive Gibbs samplers where the general framework reviewed (Containment Condition and Diminishing Condition) in the last section were used to show convergence. This paper also presents a cautionary example of a simpling-seeming adaptive Gibbs sampler that fails to converge. The paper gives a proof that this process tends to infinity with probability larger than 0 through construction of an auxiliary chain and coupling. Their method is similar to what is used in (Roberts and Rosenthal, 2007) but has greater complexity in construction of bounds. A part of my thesis study is to strengthen their results using similar techniques.

3.2.1 Notations and Preliminaries

Gibbs samplers are often used to sample from complicated high-dimensional probability distributions that have relatively straightforward conditional distribution. Such target distributions are common in Bayesian statistics. The order by which each dimension is updated can be deterministic (updating each dimension in a predetermined, sequential manner) or random, referred to as *random scan Gibbs sampler (RSG)*. A random scan Gibbs sampler choose which coordinate to update according to a *selection probability distribution*. For a random scan Gibbs sampler, it might be beneficial to introduce adaptation to selection probability to improve mixing efficiency. This is because for certain target distribution, e.g. $\pi(x, y) \propto x^{100}(1 + sin(y)), \mathcal{X} = \{0, 1\} \times [-100, 100]$, it is more efficient to focus on sampling from just one or specific several dimensions.

Metropolis-within-Gibbs is an extension of the classic Gibbs sampler with which samples of each dimension are drawn from a Metropolis sampler instead of from conditional distribution directly. For this class of algorithms, we may also have adaptations with the Metropolis sampler, in addition to adaptation of selection probability. We will concern ourselves with mostly these two types of adaptation.

Suppose there are d dimensions to the target distribution π and the state space is $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d$ and each state $X_n = (X_{n,1}, ..., X_{n_d})$. We use following notation

$$X_{n,-i} := (X_{n,1}, ..., X_{n,i-1}, X_{n,i+1}, ..., X_{n,d})$$

$$\mathcal{X}_{-i} = \mathcal{X}_1 \times ... \times \mathcal{X}_{i-1} \times \mathcal{X}_{i+1} \times ... \times \mathcal{X}_d$$

We also used $\alpha_n = (\alpha_{n,1}, ..., \alpha_{n,d})$ to denote probability to update each coordinate at step n. We denote transition kernel of a random scan Gibbs sampler with selection probability α as P_α. The set of selection probabilities is assumed to be:

$$\mathcal{Y} := [\epsilon, 1]^d \cap \Delta_{d-1}$$

where $0 < \epsilon \leq 1/d$ and $\delta d - 1 := \{(p_1, ..., p_d) \in \mathbb{R}^d : p_i \geq 0, \sum_{i=0}^{d} p_i = 1\}$

3.2.2 Adaptive Random Scan Gibbs Sampler

The Containment Condition and Diminishing Condition proposed in (Roberts and Rosenthal, 2007) may be adapted and simplified to analyze ergodicity properties of adaptive random scan Gibbs sampler. The main result is as the following:

Theorem 3.5. Let the selection probabilities $\alpha \in \mathcal{Y}$ for all n. Given
(a) $||\alpha_n - \alpha_{n-1}|| \to 0$ in probability for fixed starting values $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$.
(b) $\exists \beta \in \mathcal{Y}$ such that the random scan Gibbs with fixed kernel β is uniform ergodic.
Then the adaptive random scan Gibbs sampler is ergodic:

$$\lim_{n\to\infty} T(x_0, \alpha_0, n) = 0$$

Moreover, if $\sup_{x_0,\alpha_0} ||\alpha_n - \alpha_{n,1n-1}|| \to 0$ in probability,

$$\sup_{x_0,\alpha_0} \lim_{n\to\infty} T(x_0, \alpha_0, n) = 0$$

Apparently, condition (a) is used to derive Diminishing condition while (b) is used to derive containment condition. Let's first look at how (b) would imply simultaneous uniform ergodicity.

Proposition 3.4. If $RSG(\beta)$ is uniformly ergodic where $\beta \in \mathcal{Y}$, then $RSG(\alpha)$ is uniformly ergodic for every $\alpha \in \mathcal{Y}$. Moreover, there exists $M < \infty, \rho < 1$ such that $\sup_{x_0,\alpha} T(x_0, \alpha, n) \leq M\rho^n \to 0$.

Proof. This proof sets yet another example of utilizing small sets to show convergence. Essentially, it is trying to identify relations between P_β, which is already known to be convergent, to any other transition kernels in $\mathcal{Y}$. The small sets provide means for such comparison.

From previous results concerning small sets, we know that the entire state space $\mathcal{X}$ is small, which means that there exists $s > 0$, a probability measure μ and a positive integer m such that,

$$P_\beta^m(x, \cdot) \geq s\mu(\cdot).$$

For any $\alpha \in \mathcal{Y}$, we let

$$r := \min_i \frac{\alpha_i}{\beta_i}$$

. By the constraints we set on $\mathcal{Y}$, we have

$$\frac{\epsilon}{1-(d-1)\epsilon} \leq r \leq 1$$

Thus, P_α may be written as a mixture of transition kernels, namely

$$P_\alpha = rP_\beta + (1-r)P_q,$$

where $q = \frac{\alpha - r\beta}{1-r}$. To see this, we can just pick any coordinate say j and check that the probability to be selected with this kernel is indeed α_j:

$$P(\text{j coordinate is updated}) = r\beta_i + (1-r)\frac{\alpha_i - r\beta_i}{1-r} = \alpha_i$$

This implies that the entire state space is small with respect to P_α as well:

$$P_\alpha^m(x,\cdot) \geq r^m P_\beta^m(x,\cdot) \geq r^m s\mu(\cdot)$$

$$\geq (\frac{\epsilon}{1-(d-1)\epsilon})^m s\mu(\cdot), \forall x \in \mathcal{X}$$

. This in turn implies uniform ergodicity:

$$||P_\alpha^n(x,\cdot) - \pi(\cdot)||_{TV} \leq \left((\frac{\epsilon}{1-(d-1)\epsilon})^m s\right)^{\lfloor n/m \rfloor}$$

□

It is more obvious that (a) will imply Diminishing Condition: (a) essentially requires adaptation of selection probabilities to diminish. We only need to show that total variance distance of the kernel can be bounded by total variance distance between the selection probabilities somehow to obtain the results.

Proposition 3.5. Let P_α and $P_{\alpha'}$ be RSG using kernel $\alpha, \alpha' \in \mathcal{Y}$. Then,

$$||P_\alpha(x,\cdot) - P_{\alpha'}(x,\cdot)|| \leq \frac{|\alpha - \alpha'|}{\epsilon + |\alpha - \alpha'|} \leq \frac{|\alpha - \alpha'|}{\epsilon}$$

Proof. Let $\delta := |\alpha - \alpha'|$. Then

$$r := \min_i \frac{\alpha_i'}{\alpha_i} \geq \frac{\epsilon}{\epsilon + \max_i |\alpha_i - \alpha_i'|} \geq \frac{\epsilon}{\epsilon + \delta}$$

The inequality here can be thought of as by choosing the smallest possible α_i' and the largest distance between α_i and α_i'. We use the same method employed in the previous proof, i.e. writing $P_{\alpha'} = rP_\alpha + (1-r)P_q$ for some q. We then have

$$||P_\alpha(x,\cdot) - P_{\alpha'}(x,\cdot)|| = ||rP_\alpha + (1-r)P_\alpha - rP_\alpha + (1-r)P_q||$$

$$= (1-r)||P_\alpha - P_q|| \leq \frac{\delta}{\epsilon + \delta}$$

□

Therefore, combining the two proposition above, we just showed that the adaptive RSG satisfies both the Containment Condition and the Diminishing Condition. We then have convergence.

An erroneous claim was made in (Levine and Casella, 2006) regarding the convergence of adaptive random scan Gibbs sampler. It states that the adaptive RGS is ergodic if it follows the following conditions:
(i) $\alpha_n \to \alpha$ a.s. for some fixed $\alpha \in (0,1)^d$; and
(ii) The random scan Gibbs sampler with fixed selection probabilities α induces an ergodic Markov chain with stationary distribution π.

These conditions, however, do not guarantee ergodicity. We notice that it differs from the result we just derived in that it lacks constraints on the kernel family and the additional requirement of fixed kernel P_α to be *uniformly* ergodic. A counter example on 2D space was cited in (Łatuszyński et al., 2013) to refute this claim. We will discuss this counter example and methods of proof in the next section.

3.2.3 Adaptive Random Scan Metropolis-within-Gibbs

Such adaptive RSG can be extended by "embedding" a Metropolis within. Such algorithms are referred to as Adaptive Random Scan Metropolis with Gibbs sampler (AdapRSMwG). Here we consider the case where the proposal distribution of the Metropolis algorithm remains fixed.

Ergodicity conditions of such algorithm are given in the following theorem:

Theorem 3.6. Assume that
(a) $|\alpha_n - \alpha_{n-1}| \to 0$ in probability for fixed starting values $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$.
(b) $\exists \beta \in \mathcal{Y}$ such that the random scan Gibbs with fixed kernel β is uniform ergodic.
(c) For every $i \in \{1, ..., d\}$, $P_{x_{-i}}$ is uniformly ergodic for every $x_{-i} \in \mathcal{X}_{-i}$. Moreover there exist $s_i > 0$ and an integer m_i such that for every $x_{-i} \in \mathcal{X}_{-i}$ there exists a probability measure $v_{x_{-i}}$ on $(\mathcal{X}_i, B(\mathcal{X}_i))$ such that

$$P_{x_{-i}}^{m_i} \geq s_i v_{x_{-i}}(\cdot), \forall x_i \in \mathcal{X}_i$$

Then the AdaptRSMwG is ergodic:

$$\lim_{n\to\infty} T(x_0, \alpha_0, n) = 0$$

Moreover, if $\sup_{x_0,\alpha_0} ||\alpha_n - \alpha_{n,1n-1}|| \to 0$ in probability,

$$\sup_{x_0,\alpha_0} \lim_{n\to\infty} T(x_0, \alpha_0, n) = 0$$

Before we proceed to the proof, we will introduce the notion *strongly uniform ergodicity*:

Definition 3.2. A transition kernel P on $\mathcal{X}$ with stationary distribution π is $(m, s)-$strongly uniform ergodic, if for some $s > 0$ and positive integer m

$$P^m(x, \cdot) \geq s\pi(\cdot), \forall x \in \mathcal{X}.$$

Similarly, a family of Markov chains $\{P_\gamma\}_{\gamma \in \Gamma}$ that share a common stationary distribution π is called $(m, s)-$simultaneously strongly uniformly ergodic, if for some $s > 0$ and positive integer m

$$P_\gamma^m(x, \cdot) \geq s\pi(\cdot), \forall x \in \mathcal{X}, \gamma \in \Gamma$$

The following is a result concerning strongly uniform ergodicity.

Proposition 3.6. Let μ be a probability measure on $\mathcal{X}$. For positive integer m and $s > 0$. If a a transition kernel P
(a) is reversible and,
(b) satisfies the following

$$P^m(x, \cdot) \geq s\mu(\cdot), \forall x \in \mathcal{X},$$

then it is $((\lfloor \frac{log(s/4)}{log(1-s)} \rfloor + 2)m, \frac{s^2}{8})-$strongly uniformly ergodic.

Proof. The key of the proof is to first utilize reversibility, the fact that the chain converges to π and the Minorization condition given, in order to identify the following relations:

$$\pi(dx)P^m(x, dy) = \pi(dy)P^m(y, dx) \geq \pi(dy)s\mu(dx)$$

and consequently

$$P^m(y, dx) \geq \pi(dy)s(\mu(dx)/\pi(dy))$$

Usually, to work with the Minorization condition and similar form, we first write out

$$P^{km+m}(x, \cdot) = \int_{\mathcal{X}} P^{km}(x, dz)P^m(z, \cdot) \geq \left(\int_{\mathcal{X}} P^{km}(x, dz)s(\mu(dx)/\pi(dy)) \right)\pi(\cdot)$$

The next step involves finding a subset $A \subseteq \mathcal{X}$ such that $\left(\int_{\mathcal{X}} P^{km}(x, dz)s(\mu(dx)/\pi(dy)) \right)$ can be bounded below. One such set could be the following:

$$A := \{x \in \mathcal{X} : \mu(dx)/\pi(dx) \geq 1/2\}$$

Now

$$\left(\int_A P^{km}(x, dz)s(\mu(dx)/\pi(dy)) \right) \geq \left(\int_A P^{km}(x, dz)s/2 \right) = (s/2)P^{km}(x, A)$$

Recall that the chain converges to π uniformly. Combining this with the Minorization condition:

$$\pi(\cdot) \geq s\mu(\cdot) \to \pi(A) \geq s\mu(A)$$

Clearly $\mu(A^c) \leq 1/2$, which implies $\mu(A) > 1/2$. We are able to get the result combining arguments above. □

Before we are able to prove the main theorem, the following Proposition is needed:

Proposition 3.7. If $RSG(\beta)$ is uniformly ergodic, then there exists $s' > 0$ and positive integer m' such that the family $\{RSG(\alpha)\}_{\alpha \in \mathcal{Y}}$ is $(m', s')-$simultaneously strongly uniformly ergodic.

Remarks. Notice that both random scan Gibbs sampler and Metropolis-Hastings algorithm are reversible. Therefore, we immediately have that $RSG(\beta)$ is strongly uniformly ergodic for some m and s_1. By previous result, for any α, there exists $s_2 \geq (\frac{\epsilon}{1-(d-1)\epsilon})^m$ such that

$$P_\alpha^m(x, \cdot) \geq s_1 s_2 \pi(\cdot)$$

The final step to the main theorem involves Theorem 2 from (Roberts and Rosenthal, 1998) (slightly modified to suit our purpose):

Theorem 3.7. Consider a *random scan hybrid sampler*

$$P = \sum_{i=1}^{d} \alpha_i P_i,$$

where $\alpha_i \in \mathcal{Y}$ denotes selection probability of coordinate i and P_i denotes a Markov kernel on $\mathcal{X}$ which fixes coordinates other than i

Assume that
(i) for any i, $P_{x_{-i}}$ has stationary distribution $\pi(\cdot|X_{-i})$ and is $(N_i, \epsilon_i)-$strongly uniformly ergodic;
(ii) the corresponding RSG, with stationary distribution $\pi(\cdot)$, is $(N', \epsilon')-$strongly uniformly ergodic.
Then this hybrid sampler is $(N_*, \epsilon_*)-$strongly uniformly ergodic, where

$$N_* = N' \max_{1 \leq i \leq k} \{N_i\};$$

$$\epsilon_* = \epsilon' \min_{1 \leq i \leq k} \{\epsilon_i^{N'} k^{-N'(\max_{1 \leq i \leq k})\{N_i\}-1}\}$$

Proof of the main Theorem: To show Containment Condition, essentially we need to show for each kernel P_α the requirements (i) (ii) in Theorem 3.7

are satisfied with the same N_* and ϵ_*. (i) is guaranteed by condition (c) in Theorem 3.6 and the fact that Metropolis-Hastings algorithms are reversible. (ii) is guaranteed by Proposition 3.7 and condition (b) in Theorem 3.6. The Diminishing Condition, on the other hand, is guaranteed by condition (a) in Theorem 3.6 as before.

3.2.4 Adaptive Random Scan adaptive Metropolis-within-Gibbs

It is possible to extend the adaptation regime to allow both adaptation of selection probabilities and proposal probability distribution of the embedded Metropolis algorithm (AdapRSadapMwG). This doubly adaptive algorithm has been used in e.g. (Richardson et al., 2010) for an application in statistical genetics. Adaptation of proposal distribution of the embedded Metropolis algorithm is motivated by results in optimal scaling for random walk Metropolis algorithms (Roberts and Rosenthal, 2001). The ergodicity conditions for this algorithm can be formulated as follows:

Theorem 3.8. Let $\alpha_n \in \mathcal{Y}$ represent choices of selection probabilities, and $\gamma_n \in \Gamma$ represent choices of proposal distribution. Assume that
(a) $|\alpha_n - \alpha_{n-1}|\to 0$ in probability for fixed starting values $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$ and $\gamma_0 \in \Gamma$.
(b) $\exists \beta \in \mathcal{Y}$ such that the random scan Gibbs with fixed kernel β is uniform ergodic.
(c) For every $i \in \{1, ..., d\}$, $P_{x_{-i}}$, $x_{-i} \in \mathcal{X}_{-i}$ and $\gamma_i \in \Gamma_i$, the transition kernel P_{x_{-i},γ_i} is uniformly ergodic. Moreover there exist $s_i > 0$ and an integer m_i such that for every $x_{-i} \in \mathcal{X}_{-i}$ and $\gamma_i \in \Gamma_i$ there exists a probability measure v_{x_{-i},γ_i} on $(\mathcal{X}_i, B(\mathcal{X}_i))$ such that

$$P^{m_i}_{x_{-i}} \geq s_i v_{x_{-i}}(\cdot), \forall x_i \in \mathcal{X}_i$$

(d) The Metropolis-within-Gibbs kernels exhibit diminishing adaptation, i.e. for every $i \in \{1, ..., d\}$,

$$\sup_{x \in \mathcal{X}} ||P_{x_{-i},\gamma_{n+1,i}}(x_i, \cdot) - P_{x_{-i},\gamma_{n,i}}(x_i, \cdot)||\to 0$$

in probability as $n \to \infty$, for fixed starting values $x_0 \in \mathcal{X}$, $\alpha_0 \in \mathcal{Y}$ and $\gamma_0 \in \Gamma$. Then the AdapRSadapMwG is ergodic:

$$\lim_{n\to\infty} T(x_0, \alpha_0, n) = 0$$

Moreover, if
(a') $\sup_{x_0,\alpha_0} ||\alpha_n - \alpha_{n,1n-1}||\to 0$ in probability,
(b') $\sup_{x_0,\alpha_0} \sup_{x\in\mathcal{X}} ||P_{x_{-i},\gamma_{n+1,i}}(x_i, \cdot) - P_{x_{-i},\gamma_{n,i}}(x_i, \cdot)||\to 0$ in probability, then

$$\sup_{x_0,\alpha_0} \lim_{n\to\infty} T(x_0, \alpha_0, n) = 0$$

Proof. The method of proof is almost identical to simpler adaptive algorithms we reviewed previously in this section. We only need to check requirements (i) and (ii) in the Theorem 3.7. Notice that (c) in 3.8 and Proposition 3.6 ensures that every adaptive transition kernel for $i-th$ coordinate, i.e. P_{x_{-i},γ_i} is strongly uniformly ergodic, which satisfies (i). And (b) in 3.8 and Proposition 3.7 ensures that (ii) is satisfied.

The Diminishing Adaptation condition can be shown by "separating" both adaptations:

$$\sup_{x\in\mathcal{X}}||P_{\alpha_{n-1},\gamma_n}(x,\cdot) - P_{\alpha_{n-1},\gamma_{n-1}}(x,\cdot)||\geq$$

$$\sup_{x\in\mathcal{X}}||P_{\alpha_n,\gamma_n}(x,\cdot) - P_{\alpha_{n-1},\gamma_n}(x,\cdot)||$$

$$+\sup_{x\in\mathcal{X}}||P_{\alpha_{n-1},\gamma_n}(x,\cdot) - P_{\alpha_{n-1},\gamma_{n-i}(x,\cdot)}||$$

The first term above in the summation converges to 0 in probability by (a) in 3.8. The second term

$$\sup_{x\in\mathcal{X}}||P_{\alpha_{n-1},\gamma_n}(x,\cdot) - P_{\alpha_{n-1},\gamma_{n-i}(x,\cdot)}||\leq$$

$$\sum_{i=1}^{d}\alpha_{n-1,i}\sup_{x\in\mathcal{X}}||P_{x_{-i},\gamma_{n+1,i}}(x_i,\cdot) - P_{x_{-i},\gamma_{n,i}}(x_i,\cdot)||$$

which is derived by "extract" the summation sign out of total variation distance using triangular inequality and out of the supremum operator. This term converges to 0 in probability using condition (d) in 3.8. □

3.3 The "Stairway to Heaven" problem

3.3.1 The Problem

Definition 3.3. Let α_n denote the "selection probability" of a $d-$dimension adaptive Gibbs sampler where $\alpha_n \in \mathcal{Y}$ and $\mathcal{Y} \subseteq [0,1]^d$. An adaptive Gibbs sampler follows the following procedure:

1. Set $\alpha_n := R_n(\alpha_0, ..., \alpha_{n-1}, X_{n-1}, ..., X_0)$
2. Choose coordinate $i \in \{1, ..., d\}$ according to selection probability α_n
3. Draw $Y \sim \pi(\cdot|X_{n-1,-i})$, $-i$ meaning fix all coordinates but i
4. Set $X_n := (X_{n-1,1}, ..., X_{n-1,i-1}, Y, X_{n-1,i+1}, ..., X_{n-1,d})$

The following statement of "stairway to heaven" problem can be found in (Łatuszyński et al., 2013):

Let $\mathcal{N} = \{1, 2....\}$ and let the state space $\mathcal{X} = \{(i,j) \in \mathcal{N}\times\mathcal{N}, i = j \text{ or } i = j+1\}$, with target distribution given by $\pi(i,j) \propto j^{-2}$. On $\mathcal{X}$, consider a class of adaptive random scan Gibbs samplers for π, as defined above with update rule given by:

$$R_n(\alpha_{n-1}, X_{n-1} = (i,j)) = \begin{cases} \{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\}, & \text{if } i = j \\ \{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\}, & \text{if } i = j+1, \end{cases} \tag{45}$$

for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$

3.3.2 Simplified Proof

In this section, we give a proof of the following theorem. The proof in this section uses similar method as in the "proof by constructing phases" presented in the last section. However, the construction is entirely different and much simpler. The two original proofs are still kept because they are essentially different.

Theorem 3.9. For any fixed $\sigma \in [0,1)$, there exists a choice of $\tilde{a}_n$ such that $P(\tilde{X}_n \to \infty) > \sigma$.

Fix any natural number $K \geq 1$.

Definition 3.4. Let S_n denote the "distance" from X_n to starting position $(1,1)$. If $X_n = (x_n, x_n)$, $S_n = 2(x_n-1)$; if $X_n = (x_n, x_n-1)$, $S_n = 2(x_n-1)-1 = 2x_n - 3$.

First we first note the following Lemma:

Lemma 3.1. If $X_n = (x_n, x_n)$, the distribution of $S_{n+1} - S_n$, taking value $\{-1, 0, 1\}$, is

$$\left((\frac{1}{2} - \frac{4}{a_n}) \frac{x_n^2}{x_n^2 + (x_n - 1)^2}, 1 - (\frac{1}{2} - \frac{4}{a_n}) \frac{x_n^2}{x_n^2 + (x_n - 1)^2} - (\frac{1}{4} + \frac{2}{a_n}), \frac{1}{4} + \frac{2}{a_n} \right) \tag{46}$$

If $X_n = (x_n, x_n - 1)$, the distribution of $S_{n+1} - S_n$, taking value $\{-1, 0, 1\}$, is

$$\left(\frac{1}{4} - \frac{2}{a_n}, 1 - (\frac{1}{4} - \frac{2}{a_n}) - (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n - 1)^2}{x_n^2 + (x_n - 1)^2}, (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n - 1)^2}{x_n^2 + (x_n - 1)^2} \right) \tag{47}$$

Proof. This follows directly from the updating routine specified in previous section. □

Definition 3.5. Fix any $M \in \mathbb{N}$ such that

$$0.01 \cdot (M - 2) - 2\sqrt{K \cdot (M - 2) \ln(M - 2)} \geq 4$$

and

$$M > \max(M_0(\sigma), M_1(\sigma))$$

where $M_0(\sigma)$ and $M_1(\sigma)$ are finite numbers that depend only on the fixed σ. We will define $M_0(\sigma)$ and $M_1(\sigma)$ later in the proof: this is mostly for more concise presentation–$M_0(\sigma)$ and $M_1(\sigma)$ can be defined right away given σ as we will see later. M apparently exists for any fixed natural number K.

Definition 3.6. Define sequence N_i such that $N_0 = 0$, $N_1 = M_0(\sigma)$ and $N_i - Ni - 1 = M - 2 + 2(i - 2), i = 2, 3, ...$

Definition 3.7. Define a_n as follows:

$$\begin{cases} a_n = 8, \text{ if } 0 \leq n < N_1 \\ a_n = 8\tfrac{2i^2+1-2i}{2i-1+0.1} - 0.001, \text{ if } N_{i-1} \leq n < N_i, \forall i \geq 2 \end{cases} \tag{48}$$

Lemma 3.2. (i) $a_n \to \infty$; (ii) $a_n \geq 8$.

Proof. (i) is true since N_i is finite for all i; (ii) is a just an algebra exercise: a_n increases for $n \geq 2$ and $a_2 > 8$. □

Proposition 3.8. For each $i \geq 2$, if $x_n \geq i$ and $N_{i-1} \leq n < N_i$, there exists a sequence of i.d.d random variable $\{Z_i\}$ such that Z_i is stochastically smaller than $S_{n+1} - S_n$ for each $n \in [N_{i-1}, N_i)$ and $Z_{i,n}$ take value $\{-1, 0, 1\}$ and $\mathbb{E}(Z_i) \geq 0.01$

Proof. Since $x_n \geq i$, and a_n is constant for $N_{i-1} \leq n < N_i$

$$\frac{1}{4} + \frac{2}{a_n} > (\frac{1}{2} + \frac{4}{a_n})\frac{(x_n - 1)^2}{x_n{}^2 + (x_n - 1)^2} \geq (\frac{1}{2} + \frac{4}{a_n})\frac{(i-1)^2}{i^2 + (i-1)^2} \tag{49}$$

$$\frac{1}{4} - \frac{2}{a_n} < (\frac{1}{2} - \frac{4}{a_n})\frac{(x_n)^2}{(x_n)^2 + (x_n - 1)^2} \leq (\frac{1}{2} - \frac{4}{a_n})\frac{i^2}{i^2 + (i-1)^2} \tag{50}$$

Solve the following inequality:

$$(\frac{1}{2} + \frac{4}{a_n})\frac{(i-1)^2}{i^2 + (i-1)^2} - (\frac{1}{2} - \frac{4}{a_n})\frac{i^2}{i^2 + (i-1)^2} > 0.1 \tag{51}$$

We obtain:

$$a_n < 8\frac{2i^2 + 1 - 2i}{2i - 1 + 0.1}$$

This means that with our choice of a_n, for all $n \in [N_{i-1}, N_i)$

$$(\frac{1}{2} + \frac{4}{a_n})\frac{(i-1)^2}{i^2 + (i-1)^2} - (\frac{1}{2} - \frac{4}{a_n})\frac{i^2}{i^2 + (i-1)^2} > 0.1$$

Note that since $i \geq 2$,

$$(\frac{1}{2} - \frac{4}{a_n})\frac{i^2}{i^2 + (i-1)^2} \leq 4/5 \cdot 1/2 = 2/5$$

and since $a_n > 10$ for $i \geq 2$

$$(\frac{1}{2} + \frac{4}{a_n})\frac{(i-1)^2}{i^2 + (i-1)^2} < 1/2 \cdot (1/2 + 4/10) = 9/20$$

Denote distribution of Z_i as $(a_i, 1 - a_i - b_i, b_i)$. For $n \in [N_{i-1}, N_i)$, we choose

$$a_i = (\frac{1}{2} - \frac{4}{a_n})\frac{i^2}{i^2 + (i-1)^2} + 0.0001 \tag{52}$$

$$b_i = (\frac{1}{2} + \frac{4}{a_n})\frac{(i-1)^2}{i^2 + (i-1)^2} - 0.0001 \tag{53}$$

Note that here a_n is constant depending on only i (thus the notation does not refer to n); $a_i < 1/2, b_i < 1/2, \forall i > 1$ and $\mathbb{E}(Z_i) > 0.1 - 0.0002 = 0.0998 > 0.01$. □

Define $I_{i,m} := \sum_{j=1}^{m} Z_i$. Since Z_i is strictly bounded by $[-1, 1]$, by Hoeffding's inequality, for any $t > 0$,

$$P(|I_{i,m} - E(I_{i,m})| \geq t) \leq 2\exp(-\frac{t^2}{2m}) \tag{54}$$

For any $n > 1$, let the "bound" t be $t_m = 2\sqrt{Kn\ln(n)}$ for each n. Then the probability of "exceeding the bound" for each m is

$$P\left(|I_{i,m} - 0.01m| \geq 2\sqrt{Km\ln(m)}\right) \leq \frac{2}{m^{2K}} \tag{55}$$

Therefore,

$$P\left(I_{i,m} > 0.01m - 2\sqrt{Km\ln(m)}\right) > 1 - \frac{2}{m^{2K}} \tag{56}$$

Now we can proceed to prove the main theorem.

Proof. When $n < N_1$, $a_n = 8$. the probability of moving one step back is 0 and there is positive probability of moving ahead. From law of large number, we know that there exists finite $M_0(\sigma)$ such that $P(S_{M_0} > M) > (1 - \sigma/2)$ for arbitrarily small $\sigma > 0$.

Let's adopt the following notation:

$$\Omega_1 = \{S_{M_0} > M\}$$

$$\Omega_i = \{S_{N_i} \geq M + 4(i-1)\}$$

We need to first prove some lemmas.

Lemma 3.3. For each $i \geq 2$, under event $\cap_{j=1}^{i-1}\Omega_i$, $x_n \geq i$ for all $n \in [N_{i-1}, N_i)$.

Proof. We know that under event $\cap_{j=1}^{i-1}\Omega_i$, $S_{N_{i-1}} \geq M + 4(i-2)$. We now claim that for $N_{i-1} \leq n < N_i$, $x_n \geq i$: assume the worst case where $S_n - S_{n-1} = -1$ always for $N_{i-1} \leq n < N_i$; since $N_i - N_{i-1} = M - 2 + 2(i-2)$, we know that the smallest S_n is achieved at S_{N_i}:

$$S_{N_i} \geq S_{N_{i-1}} - (M - 2 + 2(i-2)) \geq M + 4(i-2) - M + 2 - 2i + 4 = 2i - 2.$$

This implies that the smallest value possible for $x_n, n \in [N_{i-1}, N_i)$ is $x_{Ni} \geq i$. So $x_n \geq i, n \in [N_{i-1}, N_i)$. □

Before we prove the next lemma, we restate the following theorem concerning monotone coupling and stochastic domination:

Theorem 3.10. The real random variable X is stochastically larger than Y if and only if there is a coupling between X, Y such that

$$P(X \geq Y) = 1$$

Remark. This means that if X_i is stochastically larger than Y_i respectively for all $i > 1$, we may find a coupling for each i between X_i and Y_i such that $P(\sum X_i \geq \sum Y_i) = 1$.

Lemma 3.4. For each $i \geq 2$, under event $\cap_{j=1}^{i-1}\Omega_i$, for all $n \in [N_{i-1}, N_i)$,

$$P\{S_{N_i} \geq M + 4(i-1)\} \geq 1 - \frac{2}{(M-2+2(i-2))^{2K}}$$

Proof. Note that $N_i - N_{i-1} = M - 2 + 2(i-2) \geq M - 2, \forall i \geq 2$. And from definition of M, we know that for all $m \geq M - 2$

$$0.01m - 2\sqrt{Km\ln(m)} \geq 4.$$

We have also shown that

$$P\biggl(I_{i,m} > 0.01m - 2\sqrt{Km\ln(m)}\biggr) > 1 - \frac{2}{m^{2K}} \tag{57}$$

From the previous lemma, we know that for $n \in [N_{i-1}, N_i)$ we may couple $S_{n+1} - S_n$ with Z_i since $x_n \geq i$. This gives us that

$$P\{S_{N_i} \geq M + 4(i-1)\} \geq P\{S_{N_{i-1}} + I_{i,N_i-N_{i-1}} \geq M + 4(i-1)\}$$

$$= P\biggl\{I_{i,M-2+2(i-2)} \geq M + 4(i-2) - M - 4(i-2) = 4\biggr\}$$

$$\geq P\biggl\{I_{i,M-2+2(i-2)} \geq$$

$$0.01(M-2+2(i-2)) - 2\sqrt{K(M-2+2(i-2))\ln(M-2+2(i-2))}\biggr\}$$

$$> 1 - \frac{2}{(M-2+2(i-2))^{2K}}$$

□

Corollary 3.1. From the lemma above, we have

$$P(S_n \to \infty) \geq (1 - \sigma/2) \cdot \Pi_{j=0}^{\infty}(1 - \frac{2}{(M-2+2j)^{2K}})$$

Proof. The lemma essentially provides us with the following:

$$P\{\Omega_i | \cap_{j=1}^{i-1}\Omega_j\} \geq 1 - \frac{2}{(M-2+2j)^{2K}}$$

Since $\cap_{j=1}^{\infty}\Omega_j$ implies that $\{S_n \to \infty, n \to \infty\}$. The claim follows by induction. □

The main theorem can be proved follows. Let $K = 1$. for any $\sigma \in [0, 1)$, we may choose sufficiently large $M_1(\sigma)$ so that $\Pi_{j=0}^{\infty}(1 - \frac{2}{(M_1(\sigma)-2+2j)^{2K}}) > (1 - \sigma/2)$. Such $M_1(\sigma)$ exists because $\lim_{M\to\infty} \Pi_{j=0}^{\infty}(1 - \frac{2}{(M-2+2j)^2}) = 1$.

As a result,

$$P(S_n \to \infty) \geq (1 - \epsilon) \cdot \Pi_{j=0}^{\infty}(1 - \frac{2}{(M - 2 + 2j)^{2K}})$$

$$\geq (1 - \sigma/2)^2 > 1 - \sigma$$

This concludes the proof. □

3.3.3 Proof Using An Auxiliary Process

The main result is Theorem 3.11. We first need to construct a auxiliary process as in the following proposition.

Proposition 3.9. Define a random-walk-like space homogeneous process as following:

$$S_0 = 0 \quad and \quad S_n := \sum_{i=1}^{n} Y_i, \quad for \quad n \geq 1$$

, where $Y_1, Y_2, ...$ are independent random variables taking values in $\{-1, 0, 1\}$. Let distribution of Y_n on $\{-1, 0, 1\}$ be

$$v_n = \Big\{\frac{1}{4} - \frac{1}{a_n}, \frac{1}{2}, \frac{1}{4} + \frac{1}{a_n}\}$$

.

Then there exists a choice of $\{a_n\}$, a positive,strictly increasing, unbounded sequence, such that S_n tends to infinity in probability. That is, for any large number M and any $\epsilon > 0$, there exists some positive integer N such that

$$P(S_n > M) > 1 - \epsilon, \ \forall n > N$$

.

Proof. Since Y_i is strictly bounded by $[-1, 1]$, by Hoeffding's inequality, for any $t > 0$,

$$P(|S_n - E(S_n)| \geq t) \leq 2\exp(-\frac{t^2}{2n})$$

.

For any $n > 1$, let the "bound" t be $t_n = 2\sqrt{n \ln(n)}$ for each n. Then the probability of "exceeding the bound" for each n is

$$p_n := P\bigg(|S_n - E(S_n)| \geq 2\sqrt{n\ln(n)}\bigg) \leq \frac{2}{n^2}$$

Let's choose $a_n = \sqrt[3]{n+999}$, whereby $E(S_n) = \sum_{i=1}^{n} \frac{2}{\sqrt[3]{i+999}}$.

Define $\Omega_n = \{\omega \in \Omega : |S_n(\omega) - E(S_n)| \leq t_n\}$.

Under event $\cap_{n=m}^{\infty}\Omega_n$, we have

$$S_n > E(S_n) - t_n = \sum_{i=1}^{n} \frac{2}{\sqrt[3]{i+999}} - 2\sqrt{n\ln(n)},\ \forall n > m \tag{1}$$

Notice that

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{2}{\sqrt[3]{i+999}} - 2\sqrt{n\ln(n)} = +\infty \tag{2}$$

(since $\sum_{i=1}^{n} \frac{2}{\sqrt[3]{i+999}} > \int_{1}^{n+1} \frac{2}{\sqrt[3]{x+999}}dx = 3(n+1000)^{2/3} - 300$)

And for any integer $m > 1$,

$$P(\cap_{n=m}^{\infty}\Omega_n) = \Pi_{n=m}^{\infty}(1-p_n) \geq \Pi_{n=m}^{\infty}(1-\frac{2}{n^2}) > (\frac{m-1-\sqrt{2}}{m-1+\sqrt{2}})^{\frac{\sqrt{2}}{2}} \tag{3}$$

Therefore, by "Squeeze Theorem" and (3),

$$\lim_{m\to\infty} P(\cap_{n=m}^{\infty}\Omega_n) = 1 \tag{4}$$

By (2), for any M, we are able to find positive integer N_1 so that

$$\sum_{i=1}^{n} \frac{2}{\sqrt[3]{i+999}} - 2\sqrt{n\ln(n)} > M,\ \forall n > N_1$$

By (4), for any $\epsilon > 0$, we are able to find positive integer N_2 so that

$$P(\cap_{n=m}^{\infty}\Omega_n) > 1-\epsilon,\ \forall m > N_2$$

Choose $N = \max\{N_1, N_2\}$. By (1) and the fact that $N \geq N_1$, we know that under event $\cap_{n=N}^{\infty}\Omega_n$, $S_n \geq E(S_n) - t_n > M$, for all $n > N$, i.e.

$$P(\{\omega \in \Omega : S_n(\omega) > M, \forall n > N\}) \geq P(\cap_{n=N}^{\infty}\Omega_n) > 1-\epsilon$$

□

Theorem 3.11. For any fixed $\sigma \in [0,1)$, there exists a choice of $\tilde{a}_n$ such that $P(\tilde{X}_n \to \infty) > \sigma$.

Proof. Let's first write the distribution of $\tilde{X}_n - \tilde{X}_{n-1}$ with values $\{-1,0,1\}$:

If $X_{n-1} = (i,i)$,

$$\left((\frac{1}{2} - \frac{4}{a_n}) \frac{i^2}{i^2 + (i-1)^2}, 1 - (\frac{1}{2} - \frac{4}{a_n}) \frac{i^2}{i^2 + (i-1)^2} - (\frac{1}{4} + \frac{2}{a_n}), \frac{1}{4} + \frac{2}{a_n} \right)$$

If $X_{n-1} = (i, i-1)$,

$$\left(\frac{1}{4} - \frac{2}{a_n}, 1 - (\frac{1}{4} - \frac{2}{a_n}) - (\frac{1}{2} + \frac{4}{a_n}) \frac{(i-1)^2}{i^2 + (i-1)^2}, (\frac{1}{2} + \frac{4}{a_n}) \frac{(i-1)^2}{i^2 + (i-1)^2} \right)$$

Notice that if we plug in $a_n = 8$, the above simplifies to:

If $X_{n-1} = (i,i)$,

$$\left(0, \frac{1}{2}, \frac{1}{2} \right)$$

If $X_{n-1} = (i, i-1)$,

$$\left(0, 1 - \frac{(i-1)^2}{i^2 + (i-1)^2}, \frac{(i-1)^2}{i^2 + (i-1)^2} \right)$$

Notice that if $a_n = 8$, $\tilde{X}_n - \tilde{X}_{n-1}$ is stochastically larger than random variable $Z_n = -1, 0, 1$, which is *i.d.d* for all n with the following distribution:

$$(0, \frac{9}{10}, \frac{1}{10})$$

Therefore, for $O_n := \sum_{i=1}^{n} Z_n$, there exists a coupling between $\tilde{X}_n - \tilde{X}_{n-1}$ and Z_n such that

$$\tilde{X}_n \geq O_n, \ \forall n > 1$$

Now we will use Hoeffding's inequality to construct a "lower bound" on O_n as before. Recall for any $t > 0$,

$$P(|O_n - E(O_n)| \geq t) \leq 2 \exp(-\frac{t^2}{2n})$$

.

For any $n > 1$, let the "bound" t be $t_n = 2\sqrt{n \ln(n)}$ for each n. Then the probability of "exceeding the bound" for each n is

$$P\left(|O_n - \frac{n}{10}|\geq 2\sqrt{n\ln(n)}\right) \leq \frac{2}{n^2}$$

Define $\Omega_{O,n} = \{\omega \in \Omega : |O_n(\omega) - \frac{n}{10}|\leq t_n\}$. Similar to the case for S_n, for any integer $m > 1$,

$$P(\cap_{n=m}^{\infty}\Omega_{O,n}) \geq \Pi_{n=m}^{\infty}(1 - \frac{2}{n^2}) > (\frac{m-1-\sqrt{2}}{m-1+\sqrt{2}})^{\frac{\sqrt{2}}{2}} \tag{5}$$

Notice that under event $\cap_{n=m}^{\infty}\Omega_{O,n}$,

$$O_n \geq \frac{n}{10} - 2\sqrt{n\ln(n)},\ \forall n > m$$

Meanwhile, under event $\cap_{n=m}^{\infty}\Omega_n$, as we have shown above, for all $n > m$

$$S_n < E(S_n)+t_n = \sum_{i=1}^{n}\frac{2}{\sqrt[3]{i+999}}+2\sqrt{n\ln(n)} < 3(n+999)^{\frac{2}{3}}-300+\frac{1}{5}+2\sqrt{n\ln(n)}$$

Apparently, there exists some N_0 such that the "lower bound" of O_n (of order n) will exceed "upper bound" of S_n (of order $n^{\frac{2}{3}}$), i.e. for all $n > N_0$,

$$\frac{n}{10} - 2\sqrt{n\ln(n)} > 3(n+999)^{\frac{2}{3}} - 300 + \frac{1}{5} + 2\sqrt{n\ln(n)}$$

There exists some N_1 such that $a_n = \sqrt{n+999} > 8$ for all $n > N_1$;

By (4), as $\sigma < 1$, there exists some N_2 such that

$$P(\cap_{n=m}^{\infty}\Omega_n) > \sqrt{\sigma},\ \forall m > N_2$$

There exists some N_3 such that for all $n > N_3$, the lower bound of O_n will exceed 15, i.e.

$$\frac{n}{10} - 2\sqrt{n\ln(n)} > 15,\ \forall n > N_3$$

By (5), as $\sigma < 1$, there exists some N_4 such that

$$P(\cap_{n=m}^{\infty}\Omega_{O,n}) > \sqrt{\sigma},\ \forall m > N_4$$

There exists some N_5 such that the "lower bound" of S_n minus 6 (of order $n^{\frac{2}{3}}$) will exceed a_{n+1} (of order $n^{\frac{1}{3}}$), i.e. for all $n > N_5$

$$\sum_{i=1}^{n}\frac{2}{\sqrt[3]{i+999}}-2\sqrt{n\ln(n)}-6 > 3(n+1000)^{2/3}-300-2\sqrt{n\ln(n)}-6 > \sqrt[3]{n+1000}$$

Let $N := \max\{N_0, N_1, N_2, N_3, N_4, N_5\}$.

Define $\tilde{a}_n = 8$, if $n \leq N$ while $\tilde{a}_n = a_n$ otherwise. Notice that since $N \geq N_1$, $\tilde{a}_n$ remains an increasing sequence tending to infinity.

For all $n \leq N$, $\tilde{X}_n - \tilde{X}_{n-1}$ is stochastically larger than Z_n. Therefore, there exists a coupling between $\tilde{X}_n - \tilde{X}_{n-1}$ and Z_n such that

$$\tilde{X}_N \geq O_N \tag{6}$$

Under event $(\cap_{i=N}^{\infty}\Omega_i) \cap (\cap_{i=N}^{\infty}\Omega_{O,i})$, since $N > N_0$, the "lower bound" of O_N has already exceeded "upper bound" of S_N,

$$O_N \geq S_N \tag{7}$$

By (6) and (7),

$$\tilde{X}_N \geq S_N \tag{8}$$

Now we seek to use induction to show that $\tilde{X}_n \geq S_n$ for all $n \geq N$. We want to show the following: under event $(\cap_{i=N}^{\infty}\Omega_i) \cap (\cap_{i=N}^{\infty}\Omega_{O,i})$, for all $n > N$, if given $\tilde{X}_{n-1} \geq S_{n-1}$, $\tilde{X}_n \geq S_n$.

Indeed, since $N \geq N_5$, for all $n > N$, the "lower bound" of S_n minus 6 will be already larger than a_{n+1}, i.e. under event $(\cap_{i=N}^{\infty}\Omega_i) \cap (\cap_{i=N}^{\infty}\Omega_{O,i})$, for all $n > N > N_5$,

$$S_n > 3(n+1000)^{2/3} - 300 - 2\sqrt{n\ln(n)} - 6 > \sqrt[3]{n+1000} = a_{n+1}$$

Since $\tilde{X}_{n-1} \geq S_{n-1}$,

$$\tilde{X}_{n-1} - 6 \geq S_{n-1} - 6 > a_n$$

By Lemma 6.4(b), there exists coupling of $\tilde{X}_n - \tilde{X}_{n-1}$ and Y_n such that if $\tilde{X}_{n-1} - 6 \geq a_n$ then $\tilde{X}_n - \tilde{X}_{n-1} \geq Y_n$, given that $\tilde{X}_n - \tilde{X}_{n-1}$ and Y_n follow the same a_n. In our case, this rule applies for all $n > N$. Therefore,

$$\tilde{X}_{n-1} - 6 > a_n \Longrightarrow \tilde{X}_n \geq S_n$$

.

Thus, by induction, under event $(\cap_{i=N}^{\infty}\Omega_i) \cap (\cap_{i=N}^{\infty}\Omega_{O,i})$,

$$\tilde{X}_n \geq S_n, \; \forall n \geq N$$

Therefore, since event $\cap_{i=N}^{\infty}\Omega_i$ and event $\cap_{i=N}^{\infty}\Omega_{O,i}$ are independent,

$$P(\tilde{X}_n \to \infty) > P((\cap_{i=N}^{\infty}\Omega_i) \cap (\cap_{i=N}^{\infty}\Omega_{O,i})) = P(\cap_{i=N}^{\infty}\Omega_i)P(\cap_{i=N}^{\infty}\Omega_{O,i}) = \sigma$$

□

3.3.4 Proof by constructing "phases"

Here we provide another proof to Theorem 3.11:
Theorem 3.12. For any fixed $\sigma \in [0, 1)$, there exists a choice of $\{a_n\}$ such that $P(\tilde{X}_n \to \infty) > \sigma$.

Proof. We will construct countably many "phases" for the process $\{X_n\}$. Phase i will have N_i steps where $N_i < \infty$ (N_i will be specified later). Let $S_i := \sum_{n=1}^{i} N_n, \forall i > 0$ and $S_0 = 0$. Let $o_i := \tilde{X}_{S_i} - \tilde{X}_{S_{i-1}}$, $\forall i > 1$ and $o_1 := \tilde{X}_{N_1}$. Intuitively o_i can be interpreted as the increment of $\{\tilde{X}_n\}$ during phase i.

Before we move on to define N_i, let's first start to construct our choice of $\{a_n\}$:

$$\begin{cases} a_1 = 8 \\ a_n = 8^{\frac{2i^2+1-2i}{(2i-1)}} - 0.1, \text{ if } S_{i-1} < n \leq S_i \end{cases} \tag{58}$$

Let's write the distribution of $\tilde{X}_{n+1} - \tilde{X}_n$ with values $\{-1, 0, 1\}$:

If $X_n = (x_n, x_n)$,

$$\left((\frac{1}{2} - \frac{4}{a_n}) \frac{x_n^2}{x_n^2 + (x_n-1)^2}, 1 - (\frac{1}{2} - \frac{4}{a_n}) \frac{x_n^2}{x_n^2 + (x_n-1)^2} - (\frac{1}{4} + \frac{2}{a_n}), \frac{1}{4} + \frac{2}{a_n} \right) \tag{59}$$

If $X_n = (x_n, x_n - 1)$,

$$\left(\frac{1}{4} - \frac{2}{a_n}, 1 - (\frac{1}{4} - \frac{2}{a_n}) - (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n-1)^2}{x_n^2 + (x_n-1)^2}, (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n-1)^2}{x_n^2 + (x_n-1)^2} \right) \tag{60}$$

Notice that during phase i, if $x_n > i - 1$, with our choice of a_n

$$\frac{1}{4} + \frac{2}{a_n} > (\frac{1}{2} + \frac{4}{a_n}) \frac{(x_n-1)^2}{{x_n}^2 + (x_n-1)^2} > (\frac{1}{2} - \frac{4}{a_n}) \frac{(x_n)^2}{(x_n)^2 + (x_n-1)^2} > \frac{1}{4} - \frac{2}{a_n} \tag{61}$$

This means that if $x_n > i - 1$ for all phase i, the "balance" for the next step is tipped: the probability that $\tilde{X}_{n+1} - \tilde{X}_n$ takes value 1 is larger than it takes value -1.

Claim: if $x_n > i - 1$ during phase i, there exists an *i.d.d* random variable $Z_{i,n} = -1, 0, 1$ that is stochastically smaller than $\tilde{X}_{n+1} - \tilde{X}_n$, $\forall n$ such that $S_{i-1} < n \leq S_i$ with the following distribution

$$\{a_i, 1 - a_i - b_i, b_i\}, \text{ where } b_i > a_i \tag{62}$$

and as a result, $\tilde{X}_{n+1} - \tilde{X}_n$ can be coupled to $Z_{i,n}$ in a way that $\tilde{X}_{n+1} - \tilde{X}_n > Z_{i,n}$.

Proof. Since $x_n > i-1$, and a_n is constant during phase i as defined above

$$\frac{1}{4}+\frac{2}{a_n} > (\frac{1}{2}+\frac{4}{a_n})\frac{(x_n-1)^2}{{x_n}^2+(x_n-1)^2} \geq (\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2} \tag{63}$$

$$\frac{1}{4}-\frac{2}{a_n} < (\frac{1}{2}-\frac{4}{a_n})\frac{(x_n)^2}{(x_n)^2+(x_n-1)^2} \leq (\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2} \tag{64}$$

With our choice of a_n, there exists some $\sigma_i > 0$, which is constant during phase i,

$$(\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2} - (\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2} > \sigma_i \tag{65}$$

Thus, an obvious choice of a_i and b_i is:

$$a_i = (\frac{1}{2}-\frac{4}{a_n})\frac{i^2}{i^2+(i-1)^2} + \sigma_i/3 \tag{66}$$

$$b_i = (\frac{1}{2}+\frac{4}{a_n})\frac{(i-1)^2}{i^2+(i-1)^2} - \sigma_i/3 \tag{67}$$

Note that we can always choose σ_i small enough to make sure $a_i < 1$. □

For every $\{Z_{i,n}\}$ as specified above, $e_i := E(Z_{i,n}) > 0$.

Now we seek to construct a positive integer sequence M_i as the following, for later use.

Construction of $\{M_i\}$: Before we construct $\{M_i\}$, let's first define some variables and establish some relations. Define $I_{i,n} := \sum_{j=1}^{n} Z_{i,j}$. Since Z_i is strictly bounded by $[-1,1]$, by Hoeffding's inequality, for any $t > 0$,

$$P(|I_{i,n} - E(I_{i,n})| \geq t) \leq 2\exp(-\frac{t^2}{2n}) \tag{68}$$

For any $n > 1$, let the "bound" t be $t_n = 2\sqrt{n\ln(n)}$ for each n. Then the probability of "exceeding the bound" for each n is

$$P\bigg(|I_{i,n} - ne_i| \geq 2\sqrt{n\ln(n)}\bigg) \leq \frac{2}{n^2} \tag{69}$$

Therefore,

$$P\bigg(I_{i,n} > ne_i - 2\sqrt{n\ln(n)}\bigg) > 1-\frac{2}{n^2} \tag{70}$$

We choose M_i for each i such that

$$M_i > 10, \forall i > 0 \tag{71}$$

$$ne_i - 2\sqrt{n\ln(n)} > 2, \forall n > M_i \tag{72}$$

$$\prod_{n=M_i}^{\infty}(1-\frac{2}{n^2}) > p_i := 1-\frac{1}{2i^2} \tag{73}$$

We choose N_i, i.e. the number of steps for each phase i such that

$$N_i > 2, \forall i > 0 \tag{74}$$

$$ne_i - 2\sqrt{n\ln(n)} > 2 + M_i, \forall n \geq N_i \tag{75}$$

Define event $\Omega_{M,i} = \{\omega \in \Omega : o_i(\omega) > 2 + M_i\}$. If we choose N_i as total steps of each phase, by construction above,

$$P(\Omega_{M,1}) = P(\tilde{X}_{N_1} > 2+M_1) \geq P(I_{1,N_1} > 2+M_1) > 1-\frac{2}{N_1^2} > 1-\frac{1}{2\cdot 1^2} \tag{76}$$

Now assume

$$P(\cap_{j=1}^{i-1}\Omega_{M,j}) > \prod_{n=1}^{i-1}(1-\frac{1}{2\cdot n^2}) \tag{77}$$

We want to prove the following, so that the result will be true for all $i > 0$ by induction:

$$P(\cap_{j=1}^{i}\Omega_{M,j}) > \prod_{n=1}^{i}(1-\frac{1}{2\cdot n^2}) \tag{78}$$

Under event $\cap_{j=1}^{i-1}\Omega_{M,j}$,

$$\tilde{X}_{S_{i-1}} > 2i-2+M_{i-1} \Longrightarrow x_{S_{i-1}} > i-1+\frac{M_{i-1}}{2} \tag{79}$$

Claim: under event $\cap_{j=1}^{i-1}\Omega_{M,j}$, $\tilde{X}_n - \tilde{X}_{n-1} \geq_{st} Z_{i,n}$, for all n such that $S_{i-1} < n \leq S_i$, if

$$I_{i,m} > me_i - 2\sqrt{m\ln(m)}, \text{ for all } m > M_{i-1} \tag{80}$$

Proof. For the first M_{i-1} steps during phase i, by (79), we know $x_n > i-1$, i.e.

$$x_n > i-1, \forall n,\ S_{i-1} < n \leq S_{i-1}+M_{i-1} \tag{81}$$

And in case where $n = m+S_{i-1} > M_{i-1}+S_{i-1}$, we have $I_{i,m}$ bounded by (80). With this we can prove the claim by induction:

Assume for some $n \geq M_{i-1}+S_{i-1}$,

$$x_k > i-1, \forall k \leq n \tag{82}$$

Then we can couple $\tilde{X}_n - \tilde{X}_{n-1}$ and $Z_{i,n}$ as suggested in the claim so that:

$$\begin{aligned} x_{n+1} = x_{m+S_{i-1}} &> x_{S_{i-1}} + \frac{\tilde{X}_{m+S_{i-1}} - \tilde{X}_{S_{i-1}}}{2} - 1 \\ &> x_{S_{i-1}} + \frac{I_{i,m}}{2} - 1 \\ &> i - 1 + \frac{M_{i-1}}{2} + 1 - 1 \\ &> i - 1 \end{aligned} \tag{83}$$

By our construction of a_n, during phase i, if $x_{n-1} > i - 1$, $\tilde{X}_n - \tilde{X}_{n-1} \geq_{st} Z_{i,n}$. □

For each phase $i > 1$, define the event where $I_{i,n}$ is indeed "bounded" after M_{i-1} steps:

$$\Omega_{I,i} := \{\omega \in \Omega : I_{i,n}(\omega) > ne_i - 2\sqrt{n\ln(n)}, \text{ for all } n > M_{i-1}\} \tag{84}$$

The direct result of the claim above is the following:

Under event $\Omega_{I,i} \cap (\cap_{j=1}^{i-1} \Omega_{M,j})$

$$o_i > I_{i,N_i} > N_i e_i - 2\sqrt{N_i \ln(N_i)} > 2 + M_i \tag{85}$$

This means,

$$\begin{aligned} &\Omega_{I,i} \cap (\cap_{j=1}^{i-1} \Omega_{M,j}) \subseteq \Omega_{M,i} \\ \Longrightarrow &\Omega_{I,i} \cap (\cap_{j=1}^{i-1} \Omega_{M,j}) \subseteq (\cap_{j=1}^{i-1} \Omega_{M,j}) \cap \Omega_{M,i} \\ \Longrightarrow &P(\Omega_{I,i} | \cap_{j=1}^{i-1} \Omega_{M,j}) \leq P(o_i > 2 + M_i | \cap_{j=1}^{i-1} \Omega_{M,j}) \end{aligned} \tag{86}$$

Therefore,

$$\begin{aligned} P(\cap_{j=1}^{i} \Omega_{M,j}) &= P(o_i > 2 + M_i | \cap_{j=1}^{i-1} \Omega_{M,j}) P(\cap_{j=1}^{i-1} \Omega_{M,j}) \\ &\geq P(\Omega_{I,i} | \cap_{j=1}^{i-1} \Omega_{M,j}) P(\cap_{j=1}^{i-1} \Omega_{M,j}) \\ &= \prod_{n=M_i}^{\infty} (1 - \frac{2}{n^2}) \prod_{n=1}^{i-1} (1 - \frac{1}{2 \cdot n^2}) \\ &> (1 - \frac{1}{2i^2}) \prod_{n=1}^{i-1} (1 - \frac{1}{2 \cdot n^2}) \\ &= \prod_{n=1}^{i} (1 - \frac{1}{2 \cdot n^2}) \end{aligned} \tag{87}$$

Therefore, if we choose N_i as the steps of each phase i, the probability of x-coordinate of X_n will increase by at least 1 during phase i, is larger than p_i. Therefore, we can write

$$P(X_n \to \infty) > P(\cap_{i=1}^{\infty} \Omega_{M,i}) > \prod_{i=1}^{\infty} (1 - \frac{1}{2i^2}) > 0.5 \tag{88}$$

As a matter of fact, we can push 0.5 above infinitely close towards 1 by choosing p_i that converges faster. □

4 APPENDIX I. Inequalities in Probability

Inequalities are very important tools to develop results in probability. In this section, we will record some useful, well-known inequalities in probability. Some of them are useful in my studies of the "stairway to heaven" example and the quantitative convergence rates.

Markov's Inequality can be used to bound random variables when the expectation are known (or known to be finite).

Theorem 4.1. (Markov's inequality) Suppose that $E|X|^r < \infty$ for some $r > 0$ and let $x > 0$. Then,

$$P(|X| > x) \leq \frac{E|X|^r}{x^r}$$

The Hoeffding's inequality is instrumental in my "stairway to heaven" proof. It can be generally used to bound partial sums to the expectation of partial sums, which is essentially a deterministic sequence. It is very useful to gain information regarding "speed of growth" of the partial sum S_n.

Theorem 4.2. (Hoeffding's inequality) Let $X_1, X_2, ..., X_n$ be independent random variables, such that $P(a_k \leq X_k \leq b_k) = 1$ for $k = 1, 2, ..., n$, and let $S_n, n \geq 1$, denote the partial sums. Then,

$$P(S_n - ES_n > x) \leq \exp[-\frac{2x^2}{\sum_{k=1}^{n}(b_k - a_k)^2}],$$

$$P(|S_n - ES_n| > x) \leq 2\exp[-\frac{2x^2}{\sum_{k=1}^{n}(b_k - a_k)^2}].$$

5 APPENDIX II. Useful Mathematics

In this section, we will record some useful mathematical results from real and complex analysis. Some of them are useful for my studies of asymptotic behaviors of random process or derivation of quantitative rates of convergence of specific Markov chains.

A common method to estimate functions is to use Taylor expansion. Here are a few such estimates.

Proposition 5.1. We have

$$e^x \leq 1 + x + x^2, |x| \leq 1$$

$$-(\frac{1}{1-\delta})x < \log(1-x) < -x$$

$$|e^z - 1| \leq |z|e^{|z|}$$

$$|e^z - 1 - z| \leq |z|^2$$

$$|\log(1-z) + z| \leq |z|^2$$

The function $e^{-x^2/2}$ often appears due to its close relation to normal distribution. However, as it does not have a primitive function expressible of elementary functions, we need to estimate its values via approximation. The following inequality is often useful.

Proposition 5.2.

$$(1 - 1/x^2)\frac{\phi(x)}{x} < 1 - \Phi(x) < \frac{\phi(x)}{x}.$$

In particular,

$$\lim_{x\to\infty} \frac{x(1-\Phi(x))}{\phi(x)} = 1$$

When working with infinite sums, it is often easier to derive some bounds with integrals. It is a very effective way to obtain order of decreasing/increasing speed for partial sums. Some typical estimates are listed in the following proposition.

Proposition 5.3. For $\alpha > 0, n \geq 2$

$$\frac{1}{\alpha n^\alpha} \leq \sum_{k=n}^{\infty} \frac{1}{k^{\alpha+1}} \leq \frac{1}{\alpha(n-1)^\alpha} \leq \frac{2^\alpha}{\alpha n^\alpha}.$$

This is proven to be a very useful result in probability (see my proof)–with it we know how fast the partial sum of a "quasi-harmonic series"(not exactly a harmonic series since $\alpha > 0$) is vanishing in reverse order. The decreasing speed is on par with $1/(n^\alpha)$

Similarly, for $\beta > 0$

$$\lim_{n\to\infty} n^{-\beta} \sum_{k=1}^{n} k^{\beta-1} = 1/\beta$$

The "rate of growth" of partial sum of harmonic series can be bounded as the following:

$$\log n + 1/n \leq \sum_{k=1}^{n} 1/k \leq \log n + 1$$

So "rate of growth" of partial sum of harmonic series is of the same order as $\log n$.

References

Yan Bai. An adaptive directional Metropolis-within-Gibbs algorithm. *Preprint*, 2009.

Radu V Craiu, Lawrence Gray, Krzysztof Łatuszyński, Neal Madras, Gareth O Roberts, and Jeffrey S Rosenthal. Stability of adversarial Markov chains, with an application to adaptive MCMC algorithms. *The Annals of Applied Probability*, 25(6):3592–3623, 2015.

Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85 (410):398–409, 1990.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

MI Jordan. The conjugate prior for the normal distribution. *Lecture notes on Stat260: Bayesian Modeling and Inference*, 2010.

Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

Krzysztof Łatuszyński and Jeffrey S Rosenthal. The containment condition and AdapFail algorithms. *Journal of Applied Probability*, 51(4):1189–1195, 2014.

Krzysztof Łatuszyński, Gareth O Roberts, and Jeffrey S Rosenthal. Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability*, 23(1):66–98, 2013.

Richard A Levine and George Casella. Optimizing random scan Gibbs samplers. *Journal of Multivariate Analysis*, 97(10):2071–2100, 2006.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1(2σ2):16, 2007.

Vygantas Paulauskas. Skorokhod space. `https://www.encyclopediaofmath.org/index.php/Skorokhod_spac` 2011.

Sylvia Richardson, Leonardo Bottolo, and Jeffrey S Rosenthal. Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics*, 9:539–569, 2010.

Gareth Roberts and Jeffrey Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, 2:13–25, 1997.

Gareth O Roberts and Jeffrey S Rosenthal. Two convergence properties of hybrid samplers. *The Annals of Applied Probability*, 8(2):397–407, 1998.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

Gareth O Roberts and Jeffrey S Rosenthal. Coupling and ergodicity of adaptive Markov Chain Monte Carlo algorithms. *Journal of applied probability*, 44(2): 458–475, 2007.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

Gareth O Roberts and Jeffrey S Rosenthal. Complexity bounds for MCMC via diffusion limits. *arXiv preprint arXiv:1411.0712*, 2014.

Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association*, 90(430): 558–566, 1995.

Jeffrey S Rosenthal. Analysis of the Gibbs sampler for a model related to James-Stein estimators. *Statistics and Computing*, 6(3):269–275, 1996.

Jeffrey S Rosenthal and Jinyoung Yang. Ergodicity of Combocontinuous Adaptive MCMC Algorithms. *Methodology and Computing in Applied Probability*, pages 1–17, 2017.

Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.