Markov Chain Monte Carlo Algorithm and Its Rate of Convergence

Yihui Tian
University of Toronto

December 25, 2014

STA 496H Readings in Statistics
Fall 2014
Supervisor: Prof. Jefferey S. Rosenthal

Preface

This report is an overview and summary of the supervised reading course I took under the supervision of Professor Rosenthal at Unversity of Toronto in fall 2014.

The first part of this report summarizes my study about the Markov chains on general state space. I focused on papers by Rosenthal [5], [7] and the textbook by Meyn and Tweeide [3]. The second part discusses the Markov Chain Monte Carlo(MCMC) algorithms. With the background from the first part, the way the MCMC algorithm works on general state space can be better understood. The third part is about the comparison of two theorems on Markov chain's ergodicity. One theorem is from Rosenthal[5] and the other is from Rosenthal[9]. Two examples of MCMC from [9] are developed for the comparison, and I used the software Mathematica for computing the numerical results in the examples. In the Appendices, I also included some essential theorems about probability theory that I studied from Rosenthal[2].

Many theorems, propositions and their proofs from [5] were included in this report. I also added some details and remaks to the original proofs in this report after I carefully studied them. Some examples from [5] were also used to demonstrate the notions introduced in this report.

Throughout this semester I had regular meetings with Professor Rosenthal. I would like to thank him for his careful guidance. He gave me inspiring advice and encouraged me to develop the topics in MCMC that I am interested in. My thanks also goes to Department of Statistics at Univeristy of Toronto who provided me the computing facilities.

Yihui Tian
December, 2014

Contents

1 General State Space Markov Chains

A Markov chain is a common stochastic process with the property that the next state depends only on the current state. It has been applied as a statistical model for many real-world processes. The uses of Markov chain often cover cases where the process follows a continuous state space. In this section, we will dicuss the concepts of the Markov chain on general (non-countable) state spaces, with emphasis on its asymptotic convergence and its ergodicity.

1.1 Fundementals

Definition 1. *General State Space*
The state space χ is called ***general*** *if it is equipped with a (countably generated) σ-algebra $\sigma(\chi)$.*

Throughout this report, we only focus on time-homogeneous Markov chains.

Definition 2. *Time-Homogeneity*
A Markov chain $\{X_n\}$ is called ***time-homogeneous*** *if*

$$P(X_k \in A | X_{k-1} = x) = P(X_1 \in A | X_0 = x), \qquad \forall x \in \chi, A \subseteq \chi, k \in \boldsymbol{N}^+$$

Definition 3. *Transition Probability Kernel*
If a function $P = \{P(x, A), x \in \chi, A \in \sigma(\chi)\}$ is called a ***Markov chain kernel****, it satisfies*
(1) for each $A \in \sigma(\chi)$, $P(\cdot, A)$ is a non-negative measurable function on χ.
(2) for each $x \in \chi$, $P(x, \cdot)$ is a probability measure on $\sigma(\chi)$.

For convenience, we denote n- step transitional kernel as P^n such that

$$P^n(x, \cdot) = \int_{y_1 \in \chi} \int_{y_2 \in \chi} \cdots \int_{y_{n-1} \in \chi} P(x, dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, \cdot), \qquad \forall x \in \chi.$$

Theorem 1.1. *Chapman-Kolomogorv Equation*

$$P^{n+m}(x, \cdot) = \int_{y \in \chi} P^n(x, dy) P^m(y, \cdot), \qquad \forall n, m \in \boldsymbol{N} \cup \{0\}$$

Definition 4. *Stationary Measure(Distribution)*
A probability measure $\pi(\cdot)$ on $(\chi, \sigma(\chi))$ is a ***stationary measure(distribution)*** *for a Markov chain with transition probability P if*

$$\pi(A) = \int_{x \in \chi} P(x, A) \pi(dx) \qquad \forall x \in \chi, \forall A \subseteq \chi$$

The notion of "irreducibility" from discrete Markov chain is impossible when χ is uncountable, since often $P(x, \{y\}) = 0$ for all x and y. Therefore we introduce a weaker condition of irreducibility for general state space:

Definition 5. *ϕ-irreducibility*
A chain is ϕ-irreducible if there exists a non-zero σ-finite measure[1]*, such that for all $A \subseteq \chi$ with $\phi(A) > 0$, and for all $x \in \chi$, there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$.*

[1] see Appendices

Note that when χ is countable, ϕ-irreducibility does not necessarily imply irreducibility. For instance, let $\chi = \mathbf{N}$. If $\phi(A) = \delta_{x^*}(A)$ and $P(x, x^*) = 1$, then it is ϕ-irreducible: If $\phi(A) > 0$, then A must contain x^*; so for any $x \in \chi$, $P(x, A) \geq P(x, x^*) = 1$. However, this chain is not irreducible.

Unlike discrete Markov chains with individual state having its own period, we need to examine the period of a subset of the general state space.

Definition 6. *Periodicity*
*A Markov chain with stationary distribution $\pi(\cdot)$ is **aperiodic** if there do not exist $d \geq 2$ and disjoint subsets $\chi_1, \chi_2, \ldots, \chi_d \subseteq \chi$ with $P(x, \chi_{i+1}) = 1$ for all $x \in \chi_i (1 \leq i \leq d-1)$, and $P(x, \chi_1) = 1$ for all $x \in \chi_d$, such that $\pi(\chi_1) > 0$(and hence $\pi(\chi_i) > 0$ for all i). Otherwise, the chain is **periodic** with period d, and period decomposition $\chi_1, \ldots, \chi_d$.*

Definition 7. *First Hitting Time and First Return Time*
For any set $A \subseteq \chi$, if the variable τ_A is defined as:

$$\min\{n \geq 0, X_n \in A\},$$

it is interpreted as the first hitting time. If τ_A is defined as:

$$\min\{n \geq 1, X_n \in A\},$$

it is interpreted as the first return time.

Definition 8. *Reversibility*
*A Markov chain on a state space χ is **reversible** with respect to a probability distribution $\pi(\cdot)$ on χ, if*

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx), \qquad x, y \in \chi.$$

A very important property of reversibiity is the following:

Proposition 1.1. *If Markov chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is stationary for the chain.*

Proof. $\int_\chi \pi(dx)P(x, dy) = \int_\chi \pi(dy)P(y, dx) = \pi(dy)$. □

Suppose $\nu(\cdot)$ is a probability measure on χ, P is a transition kernel on χ and h is any measurable function $\chi \to \mathbf{R}$, then for any $A \subseteq \chi$,

$$(\nu P)(A) := \int_\chi \nu(dx)P(x, A)$$
$$(Ph)(x) := \int_\chi P(x, dy)h(y).$$

Thus, $(\nu P)(A)$ is the expectation of $P(X, A)$, where $X \sim \nu(\cdot)$ and $(Ph)(x)$ is the conditional expected value of $h(X_{n+1})$, given $X_n = x$.

1.2 Convergence of Markov Chains

We measure the “distance” between to two probability measures in terms of total variation distance, which is defined as the following:

Definition 9. *Total Variation Distance*
*The **total variation distance** between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is*

$$||\nu_1(\cdot) - \nu_2(\cdot)|| = \sup_A |\nu_1(A) - \nu_2(A)| \qquad A \subseteq \chi.$$

The following properties of total variation distance are useful in the proofs of the convergence theorems later.

Proposition 1.2. *Properties of Total Variation Distance*
Suppose $\nu_1(\cdot)$ and $\nu_2(\cdot)$ are two probability measures. Then

(a) $||\nu_1(\cdot) - \nu_2(\cdot)|| = \sup\limits_{f:\chi\to[0,1]} |\int f d\nu_1 - \int f d\nu_2|$.
(b) For any $a < b$,

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

In particular,

$$||\nu_1(\cdot) - \nu_2(\cdot)|| = \frac{1}{2} \sup_{f:\chi\to[-1,1]} |\int f d\nu_1 - \int f d\nu_2|.$$

(c) If $\pi(\cdot)$ is stationary for a Markov chain kernel P, then $||P^n(x,\cdot) - \pi(\cdot)||$ is non-increasing in n, i.e. $||P^n(x,\cdot) - \pi(\cdot)|| \leq ||P^{n-1}(x,\cdot) - \pi(\cdot)||$ for $n \in \boldsymbol{N}$.
(d) More generally, we always have

$$||(\nu_1 P)(\cdot) - (\nu_2 P)(\cdot)|| \leq ||\nu_1(\cdot) - \nu_2(\cdot)||.$$

(e) Let $t_n = 2\sup\limits_{x\in\chi}||P^n(x,\cdot) - \pi(\cdot)||$, where $\pi(\cdot)$ is stationary. Then t is sub-multipicative, i.e.

$$t(m+n) \leq t(m)t(n) \qquad \forall m,n \in \boldsymbol{N}.$$

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

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

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

Proof. (a) and (b)
Note that (b) This part is a generalized version of (a), so we show proof for (b) first. According to Radon-Nikodym theorem[2], we can let $\rho(\cdot)$ be any σ-finite measure such that

[2] See Appendices.

$\nu_1 \ll \rho$ and $\nu_2 \ll \rho$. Let $g = \frac{d\nu_1}{d\rho}$ and $h = \frac{d\nu_2}{d\rho}$, then they are the Radon-Nikodym derivatives of ν_1 and ν_2 with respect to ρ. For any function $f : \chi \to [a, b]$, we have

$$|\int f d\nu_1 - \int f d\nu_2| = |\int f(g-h)d\rho| = |\int_{\{g \le h\}} f(g-h)d\rho + \int_{\{g>h\}} f(g-h)d\rho|$$

Since $g - h \ge 0$ on $\{g > h\}$ and $a \le f(x) \le b$, it follows that

$$a \int_{\{g>h\}} (g-h)d\rho \le \int_{\{g>h\}} f(g-h)d\rho \le b \int_{\{g>h\}} (g-h)d\rho. \tag{1}$$

Similarly, since $h - g \le 0$ on $\{g \le h\}$,

$$b \int_{\{g \le h\}} (g-h)d\rho \le \int_{\{g \le h\}} f(g-h)d\rho \le a \int_{\{g \le h\}} (g-h)d\rho. \tag{2}$$

From (1) and (2), we have

$$\begin{aligned}
&(b-a)\int_{\{g \le h\}} (g-h)d\rho + a\int_{\chi} (g-h)d\rho = b\int_{\{g \le h\}} (g-h)d\rho + a\int_{\{g>h\}} (g-h)d\rho \\
&\le \int_{\chi} f(g-h)d\rho \le b\int_{\{g>h\}} (g-h)d\rho + a\int_{\{g \le h\}} (g-h)d\rho \\
&= (b-a)\int_{\{g>h\}} (g-h)d\rho + a\int_{\chi} (g-h)d\rho = (b-a)\int_{\{g>h\}} (g-h)d\rho,
\end{aligned}$$

where the last equality comes from the fact $\int_{\chi}(g-h)d\rho = \nu_1(\chi) - \nu_2(\chi) = 0$.
So since $(b-a)\int_{\{g \le h\}}(g-h)d\rho \le \int_{\chi} f(g-h)d\rho \le (b-a)\int_{\{g>h\}}(g-h)d\rho$, we have

$$\frac{1}{(b-a)} \sup_{f:\chi \to [a,b]} |\int_{\chi} f(g-h)d\rho| = \max\{\int_{\{g>h\}} (g-h)d\rho, |\int_{\{g \le h\}} (g-h)d\rho|\},$$

On the other hand, for any $A \subseteq \chi$ we have
$|\nu_1(A) - \nu_2(A)| = |\int_A (g-h)d\rho| = |\int_{A \cap \{g>h\}}(g-h)d\rho + \int_{A \cap \{g \le h\}}(g-h)d\rho|$. Hence $|\nu_1(A) - \nu_2(A)|$ might be maximized either when $A = \{g > h\}$ or $A = \{g \le h\}$. Thus

$$\begin{aligned}
||\nu_1(\cdot) - \nu_2(\cdot)|| &= \max\{\int_{\{g>h\}} (g-h)d\rho, |\int_{\{g \le h\}} (g-h)d\rho|\} \\
&= \frac{1}{b-a} \sup_{f:\chi \to [a,b]} |\int_{\chi} f(g-h)d\rho|,
\end{aligned}$$

which verifies (b). When $f : \chi \to [0, 1]$, we have (a). □

Proof. (c) and (d)
For any $n \in \mathbf{N}$ and $A \subseteq \chi$, we let $f(y) = P(y, A)$, then

$$\begin{aligned}
|P^{n+1}(x, A) - \pi(A)| &= |\int_{y \in \chi} P^n(x, dy)P(y, A) - \int_{y \in \chi} \pi(dy)P(y, A)| \\
&= |\int_{y \in \chi} P^n(x, dy)f(y) - \int_{y \in \chi} \pi(dy)f(y)| \\
&\le ||P^n(x, \cdot) - \pi(\cdot)||,
\end{aligned}$$

where the last inequality comes from (a). Thus

$$||P^{n+1}(x,\cdot)-\pi(\cdot)|| = \sup_A |P^{n+1}(x,A)-\pi(A)| \leq ||P^n(x,\cdot)-\pi(\cdot)||.$$

Very similarly, for (d), we have

$$\begin{aligned}|(\nu_1 P)(A)-(\nu_2 P)(A)| &= |\int \nu_1(dx)P(x,A) - \int \nu_2(dx)P(x,A)| \\ &= |\int f(x)\nu_1(dx) - \int f(x)\nu_2(dx)| \\ &\leq ||\nu_1(\cdot)-\nu_2(\cdot)||.\end{aligned}$$

□

Proof. (e)
Let $\hat{P}(x,\cdot) = P^n(x,\cdot)-\pi(\cdot)$ and $\hat{Q}(x,\cdot) = P^m(x,\cdot)-\pi(\cdot)$, then

$$\begin{aligned}&(\hat{P}\hat{Q}f)(x) \equiv \int_{y\in\chi} f(y)\int_{z\in\chi}[P^n(x,dz)-\pi(dz)][P^m(z,dy)-\pi(dy)] \\ &= \int_{y\in\chi} f(y)[\int_{z\in\chi} P^n(x,dz)P^m(z,dy) - \int_{z\in\chi}\pi(dz)P^m(z,dy) - \pi(dy)\int_{z\in\chi}P^n(x,dz) - \pi(dy)\int_{z\in\chi}\pi(dz)] \\ &= \int_{y\in\chi} f(y)[P^{n+m}(x,dy)-\pi(dy)-\pi(dy)+\pi(dy)] \qquad (since\ \pi\ stationary) \\ &= \int_{y\in\chi} f(y)[P^{n+m}(x,dy)-\pi(dy)]\end{aligned}$$

Let $f:\chi\to[0,1]$ and $g(x) = (\hat{Q}f)(x) \equiv \int_{y\in\chi}\hat{Q}(x,dy)f(y)$. Let $g^* = \sup_{x\in\chi}|g(x)|$, then

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

If $g^*=0$, then for almost every $x\in\chi$, $\int_{y\in\chi} f(y)(P^m(x,dy)-\pi(dy))=0$, thus

$$(\hat{P}\hat{Q}f)(x) = \int_{z\in\chi}[P^n(x,dz)-\pi(dz)]\int_{y\in\chi}f(y)[P^m(z,dy)-\pi(dy)]=0.$$

If $g^*\neq 0$,

$$2\sup_{x\in\chi}|(\hat{P}\hat{Q}f)(x)| = 2g^*\sup_{x\in\chi}|(\hat{P}[\frac{g}{g^*}])(x)| \leq t(m)\sup_{x\in\chi}|(\hat{P}[\frac{g}{g^*}])(x)| \qquad (3)$$

Since $-1 \le \frac{g}{g^*} \le 1$, we have $(\hat{P}[\frac{g}{g^*}])(x) \le 2||P^n(x,\cdot) - \pi(\cdot)||$ by (b), so $\sup\limits_{x\in\chi}(\hat{P}[\frac{g}{g^*}])(x) \le t(n)$. Then

$$\begin{aligned}
t(n+m) &= 2\sup_{x\in\chi}||P^{n+m}(x,\cdot) - \pi(\cdot)|| \\
&= 2\sup_{x\in\chi}\sup_{f:\chi\to[0,1]} |\int (f dP^{m+n} - \int f d\pi)| \qquad (by\ (b)) \\
&= 2\sup_{x\in\chi}\sup_{f:\chi\to[0,1]} |(\hat{P}\hat{Q}f)(x)| \qquad \big(since\ (\hat{P}\hat{Q}f)(x) = \int_{y\in\chi} f(y)[P^{n+m}(x,dy) - \pi(dy)]\big) \\
&= 2\sup_{f:\chi\to[0,1]}\sup_{x\in\chi} |(\hat{P}\hat{Q}f)(x)| \\
&\le t(m)\sup_{f:\chi\to[0,1]}\sup_{x\in\chi}(\hat{P}[\frac{g}{g^*}])(x) \qquad (by\ (3)) \\
&\le t(m)t(n)
\end{aligned}$$

□

Proof. (f)
The first equality follows since as in proof of (b) with $a = -1$ and $b = 1$, for either $f = 1$ on $\{g - h > 0\}$ and $f = -1$ on $\{g - h \le 0\}$ or vice versa, we both have

$$||\mu(\cdot) - \nu(\cdot)|| = \frac{1}{2}(\int_{\{g>h\}}(g-h)d\rho + \int_{\{g\le h\}}(g-h)d\rho) = \frac{1}{2}\int_{\chi}(M-m)d\rho.$$

The second equality follows since $M + m = g + h$, so that $\int_\chi (M+m)d\rho = 2$, and hence

$$\begin{aligned}
\frac{1}{2}\int_\chi (M-m)d\rho &= 1 - \frac{1}{2}(2 - \int_\chi (M-m)d\rho) = 1 - \frac{1}{2}\int_\chi (M + m - (M-m))d\rho \\
&= 1 - \int_\chi m d\rho.
\end{aligned}$$

□

Proof. (g)
Let $a = \int_\chi m d\rho$, $b = \int_\chi (g-m)d\rho$, and $c = \int_\chi (h-m)d\rho$. First assume a, b, c are all positive. We jointly construct random variables Z, U, V, I such that Z has density $\frac{m}{a}$, U has density $\frac{(g-m)}{b}$, V has density $\frac{(h-m)}{b}$, and I is independent of Z, U, V with $P(I = 1) = a$ and $P(I = 0) = 1 - a$. We then let $X = Y = Z$ if $I = 1$, and $X = U$ and $Y = V$ if $I = 0$. For any $A \subseteq \chi$,

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

thus $X \sim \mu(\cdot)$. Similarly, $Y \sim \nu(\cdot)$ because

$$\begin{aligned}P(Y \in A) &= P(V \in A, I = 0) + P(Z \in A, I = 1) = P(V \in A)P(I = 0) + P(Z \in A)P(I = 1)\\ &= \frac{1-a}{b}\int_A (h - m)d\rho + \int_A m d\rho = \int_A h d\rho.\end{aligned}$$

Furthermore, U has support $\{g-h < 0\}$ and V has support $\{g-h \geq 0\}$, thus $P(U = V) = 0$. By the first equality of (f), $P(X = Y) = P(I = 1) = a = 1 - ||\mu(\cdot) - \nu(\cdot)||$.

If any of b or c equals 0, without loss of generality, we assume $b = 0$. In this case, we have $g = \min(g, h)$ almost everywhere; but since $\int_\chi g d\rho = \int_\chi h d\rho = 1$, indeed $g = h$ almost everywhere. Then $a = 1$, $X = Y = Z \sim \mu(\cdot)$, $||\mu(\cdot) - \nu(\cdot)|| = 0$ and $P(X = Y) = 1 = 1 - ||\mu(\cdot) - \nu(\cdot)||$; if $a = 0$, then $X = U$, $Y = V$ and the rest of the proof follows the same as the above paragraph. Therefore, the statement also holds true if any of a, b, c equals zero. □

Definition 10. *Small Set and Minorisation Condition*
*A subset $C \subseteq \chi$ is **small** (or (n_0, ϵ, ν)-small) if there exists a positive integer $n_0, \epsilon > 0$, and a probability measure $\nu(\cdot)$ on χ such that the following **minorisation condition** holds:*

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

i.e. $P^{n_0}(x, A) \geq \epsilon\nu(A)$ for all $x \in C$ and $A \in \sigma(\chi)$.

Since the proofs of the Markov chain convergence theorem and the theorems about the Markov chain's ergodicity(we will present) use direct coupling constructions, we first introduce the notion of coupling:

Theorem 1.2. *The Coupling Inequality*
Suppose we have two random variables X and Y, defined jointly on some state space χ. Let $L(X)$ and $L(Y)$ for their respective probability distributions, then $||L(X) - L(Y)|| \leq P(X \neq Y)$.

Proof.

$$\begin{aligned}||L(X) - L(Y)|| &= \sup_A |P(X \in A) - P(Y \in A)|\\ &= \sup_A |P(X \in A, X = Y) + P(X \in A, X \neq Y)\\ &\quad - P(Y \in A, X = Y) - P(Y \in A, X \neq Y)|\\ &= \sup_A |P(X \in A, X \neq Y) - P(Y \in A, X \neq Y)|\\ &\leq P(X \neq Y),\end{aligned}$$

where the last inequality follows since both $P(X \in A, X \neq Y)$ and $P(Y \in A, X \neq Y)$ are nonnegative and $\leq P(X \neq Y)$. □

The Coupling Construction:
Suppose we run two copies $\{X_n\}$ and $\{X'_n\}$ of the Markov chains on χ following the instructions below, and C is a small set((n_0, ϵ, ν)-small). We start with $X_0 = x$ and $X'_0 \sim \pi(\cdot)$, and repeat the following loop forever:

Beginning of Loop Given X_n and X'_n:
1. If $X_n = X'_n$, choose $X_n = X'_{n+1} \sim P(X_n, \cdot)$, and replace n by $n+1$.
2. Else, if $(X_n, X'_n) \in C \times C$, then:
(a)with probability ϵ, choose $X_{n+n_0} = X'_{n+n_0} \sim \nu(\cdot)$;
(b)else, with probability $1-\epsilon$, conditionally independently choose
$X_{n+n_0} \sim \frac{1}{1-\epsilon}[P^{n_0}(X_n, \cdot) - \epsilon\nu(\cdot)]$ and $X'_{n+n_0} \sim \frac{1}{1-\epsilon}[P^{n_0}(X'_n, \cdot) - \epsilon\nu(\cdot)]$.
Replace n with $n_0 + n$
3. Else, conditionally independently choose $X_{n+1} \sim P(X_n, \cdot)$ and $X'_{n+1} \sim P(X'_n, \cdot)$, and replace n by $n+1$.
Return to Beginning of Loop.

Note that from such construction, the two chains marginally follows the updating rules $P(x, \cdot)$: It is obvious for condition 1 and condition 3; when $(X_n, X'_n) \in C \times C$, we have $X_{n+n_0} \sim \epsilon\nu(\cdot) + \frac{1-\epsilon}{1-\epsilon}[P^{n_0}(X_n, \cdot) - \epsilon\nu(\cdot)] = P^{n_0}(X_n, \cdot)$. Thus, it follows that $P(X_n \in A) = P^n(x, A)$ and $P(X'_n \in A) = \pi(A)$(since π is stationary). Moreover, their joint construction(using small set C) gives them a high probability of becoming equal to each other. In the case of $n_0 > 1$, if $(X_n, X'_n) \in C \times C$f, or completeness we go back and construct $X_{n+1}, \cdots, X_{n+n_0-1}$ from their conditional distributions given X_n and X_{n+n_0}. Specifically, suppose $n_0 = 3$ and we know $X_8 = 8$ and $X_{11} = 11$. $P(X_9 \in A | X_8 = 8, X_{11} = 11) = \int_{z \in \chi} \int_{y \in A} P(8, dy)P(y, dz)P(z, 11)$; once X_9 is generated, say X_9 is a, we then have $P(X_{10} \in A | X_8 = 8, X_9 = a, X_{11} = 11) = P(X_{10} \in A | X_9 = a, X_{11} = 11) = \int_{y \in A} P(a, dy)P(y, 11)$. In the same way we construct $X'_{n+1}, \ldots, X'_{n+n_0-1}$ from their conditional distribution given X'_n and X'_{n+n_0}.

Theorem 1.3. *Markov Chain Convergence Theorem*
If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e. [3]*$x \in \chi$,*

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

In particular, $\lim_{n\to\infty} P^n(x, A) = \pi(A)$ for all measurable $A \subseteq \chi$.
In fact, if h:$\chi \to R$ with $\pi(|h|) < \infty$, then a "strong law of large numbers" also holds as follows:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \pi(h) \qquad w.p.\ 1.$$ [4]

The coupling approach requires a small set existing in the state space but we are not sure about its existence in the Markov chain described in Theorem (1.3). However, it was proved by Jain and Jameson[1] that

Theorem 1.4. [5] *Every ϕ-irreducible Markov chain, on a state space with countably generated σ-algebra, contains a small set $C \subseteq \chi$ with $\phi(C) > 0$; furthermore, the minorisation measure $\nu(\cdot)$ may be taken to satisfy $\nu(C) > 0$.*

[3] "a.e." stands for almost everywhere.
[4] "w.p.1" stands for with probability 1
[5] The proof can be found in [1].

Our plan for proving Theorem (1.3) is to show that the pair (X_n, X'_n) will hit $C \times C$ infinitely often, then they will have probability $\geq \epsilon > 0$ of coupling each time, then $P_{(x,y)}(\tau_{C\times C} < \infty) = 1 - P_{(x,y)}(\tau_{C\times C} = \infty) \geq 1 - \lim_{n\to\infty}(1-\epsilon)^n = 1$, i.e. they will eventually couple with probability 1, and we finish the proof. In proving the assumption that the chain will hit $C \times C$ infinitely often, we first prove the following claim:

Claim 1. *Consider a Markov chain on a stae space χ, having stationary distribution $\pi(\cdot)$. Suppose that for some $A \subseteq \chi$, we have $P_x(\tau_A < \infty) > 0$ for all $x \in \chi$. Then for π-a.e. $x \in \chi$, $P_x(\tau_A < \infty) = 1$.*

Proof. We prove the claim with a contradiction. Suppose the conclusion in the claim does not hold, i.e.

$$\pi\{x \in \chi : P_x(\tau_A = \infty) > 0\} > 0. \tag{4}$$

Then we have the following results:

Result 1. *There are constants $l, l_0 \in \boldsymbol{N}$, $\delta > 0$ and $B \subseteq \chi$ with $\pi(B) > 0$, such that*

$$P_x(\tau_A = \infty, \sup\{k \geq 1; X_{kl_0} \in B\} < l) \geq \delta, \qquad \forall x \in B. \tag{5}$$

Proof. Result 1
From (4), there must exist $\delta_1 > 0$ such that $\pi\{x \in \chi : P_x(\tau_A = \infty) \geq \delta_1)\} > 0$. Let $B_1 = \{x \in \chi : P_x(\tau_A = \infty) > \delta_1\}$, then for all $x \in B_1$, $P_x(\tau_A < \infty) \leq 1 - \delta_1$. By the assumption that $P_x(\tau_A < \infty) > 0$ for all $x \in \chi$, we can find $l_0 \in \mathbf{N}$ and $\delta_2 > 0$ and $B_2 \subseteq B_1$, with $\pi(B_2) > 0$ and with $P^{l_0}(x, A) \geq \delta_2$ for all $x \in B_2$. Indeed, B_1 can be written as $B_1 = \bigcup_{m=1}^{\infty} \bigcup_{j=1}^{\infty} B_{mj}$, where $B_{mj} = \{x \in B_1 | P^m(x, A) \geq \frac{1}{j}\}$. Since $\pi(B_1) > 0$, by sub-additivity of π, there must exist mj such that $\pi(B_{mj}) > 0$, and thus m is the l_0 and $\frac{1}{j}$ is the δ_2 we are looking for. Set $\eta = \#\{k \geq 1; X_{kl_0} \in B_2\}$. Then for any $r \in \mathbf{N}$ and $x \in \chi$, $P_x(\tau_A = \infty, \eta = r) \leq (1 - \delta_2)^r$. Thus,

$$P_x(\tau_A = \infty, \eta = \infty) = P_x(\tau_A = \infty, \bigcup_{r=1}^{\infty}\{\eta = r\}) = \lim_{r\to\infty} P_x(\tau_A = \infty, \eta = r) = 0.$$

Hence for $x \in B_2$, we have

$$P_x(\tau_A = \infty, \eta < \infty) = 1 - P_x(\tau_A = \infty, \eta = \infty) - P_x(\tau_A < \infty) \geq \delta_1.$$

then there exist $l \in \mathbf{N}$, $\delta > 0$, and $B \subseteq B_2$ with $\pi(B) > 0$, such that for any $x \in B$,

$$P_x(\tau_A = \infty, \sup\{k \geq 1; X_{kl_0} \in B_2\} < l) \geq \delta.$$

Since $B \subseteq B_2$, we have $sup\{k \geq 1; X_{kl_0} \in B_2\} \geq sup\{k \geq 1; X_{kl_0} \in B\}$, we thus have

$$P_x(\tau_A = \infty, \sup\{k \geq 1; X_{kl_0} \in B\} < l) \geq P_x(\tau_A = \infty, \sup\{k \geq 1; X_{kl_0} \in B_2\} < l) \geq \delta.$$

□

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

$$\int_{x\in\chi} \pi(dx) P_x[S = r, X_{jL} \notin A] \geq \pi(B)\delta.$$

Proof. By Result (1) and π is a stationary distribution, we have

$$\begin{aligned}\int_{x\in\chi}\pi(dx)P_x(S=r,X_{jL}\notin A) &= \int_{x\in\chi}\pi(dx)\int_{y\in B}P^{rL}(x,dy)P_y[S=-\infty,X_{(j-r)L\notin A}]\\ &= \int_{y\in B}\int_{x\in\chi}\pi(dx)P^{rL}(x,dy)P_y[S=-\infty,X_{(j-r)L\notin A}]\\ &= \int_{y\in B}\pi(dy)P_y[S=-\infty,X_{(j-r)L\notin A}]\\ &\geq \int_B\pi(dy)\delta\\ &= \pi(B)\delta.\end{aligned}$$

□

With all the above two results, we now complete the proof for the claim. For any $j\in\mathbf{N}$,

$$\begin{aligned}\pi(A^c) = \int_{x\in\chi}\pi(dx)P^{jL}(x,A^c) &= \int_{x\in\chi}\pi(dx)P_x(X_{jL}\notin A)\\ &\geq \sum_{r=1}^{j}\int_{x\in\chi}\pi(dx)P_x[S=r,X_{jL}\notin A]\\ &\geq \sum_{r=1}^{j}\pi(B)\delta = j\pi(B)\delta,\end{aligned}$$

when $j>\frac{1}{\pi(B)\delta}$, this gives $\pi(A^c)>1$,which is impossible, therefore we reach a contradiction. So we finish the proof of Claim (1). □

Lemma 1.5. *Consider an aperiodic Markov chain on a state space χ, with stationary distribution $\pi(\cdot)$. Let $\nu(\cdot)$ be any probability measure on χ. Assume that $\nu(\cdot)\ll\pi(\cdot)$, and that for all $x\in\chi$, there is $n=n(x)\in\boldsymbol{N}$ and $\delta=\delta(x)>0$ such that $P^n(x,\cdot)\geq\delta\nu(\cdot)$. Let $T=\{n\geq 1;\exists\delta_n>0\ s.t.\int\nu(dx)P^n(x,\cdot)\geq\delta_n\nu(\cdot)\}$, and assume that T is non-empty. Then there is $n^*\in\boldsymbol{N}$ with $\{n^*,n^*+1,n^*+2,\dots\}\subseteq T$.*

Proof. Since for all $x\in\chi$, $P^{(n(x))}(x,\cdot)\geq\delta(x)\nu(\cdot)$, then $\int\nu(dx)P^{(n(x))}(x,\cdot)\geq\left(\int\nu(dx)\delta(x)\right)\nu(\cdot)$, so T is nonempty.
If $n,m\in T$, then

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

Thus if $n,m\in T$, then $n+m\in T$.
We now show that $gcd(T)=1$. Suppose $gcd(T)=d>1$, then for $1\leq i\leq d$, let

$$\chi_i=\{x\in\chi;\ \exists l\in\mathbf{N}\ and\ \delta>0\ s.t.\ P^{ld-i}(x,\cdot)\geq\delta\nu(\cdot)\}.$$

Then by the assumption, $\bigcup_{i=1}^{d} \chi_i = \chi$. Let $S = \bigcup_{i \neq j}(\chi_i \cap \chi_j)$ and let

$$\bar{S} = S \cup \{x \in \chi; \exists m \in \mathbf{N} \; s.t. \; P^m(x, S) > 0\}.$$

Let $\chi_i' = \chi_i \backslash \bar{S}$. Then since removing S, we have $\chi_1', \chi_2', \cdots, \chi_d'$ disjoint to each other. If $x \in \chi_i'$, then $P(x, \bar{S}) = 0$, and since $\bigcup_{j=1}^{d} \chi_j' = \chi \backslash \bar{S}$, we have $P(x, \bigcup_{j=1}^{d} \chi_j') = 1$. Furthermore, we must have $P(x, \chi_{i+1}') = 1$ in the case $i < d$ and $P(x, \chi_1') = 1$ for $i = d$ (Suppose y is in χ_1'. if y moves to x after one step and x is some state other than χ_2', then y would not be in χ_1'). For all $m \geq 0$, we next show that $\nu P^m(\chi_i \cap \chi_j) = 0$ whenever $i \neq j$. Assume not, i.e. there exists i and j such that $\nu P^m(\chi_i \cap \chi_j) > 0$, i.e. $\int_\chi \nu(dx) P^m(x, \chi_i \cap \chi_j) > 0$, then there would be $S' \subseteq \chi$, and $l_1, l_2 \in \mathbf{N}$, and $\delta > 0$ such that for all $x \in S'$, $P^{l_1 d + i}(x, \cdot) \geq \delta \nu(\cdot)$ and $P^{l_2 d + j}(x, \cdot) \geq \delta \nu(\cdot)$, implying that $l_1 d + i + m \in T$ and $l_2 d + j + m \in T$, contradict the assumption that $gcd(T) = d$. Thus $\nu P^m(\chi_i \cap \chi_j) = 0$ for all $i \neq j$ and $m \geq 0$. Thus let $m = 0$, we then have $\nu(\chi_i \cap \chi_j) = 0$. By sub-additivity of measures we $\nu(\bar{S}) \leq \bigcup_{i \neq j} \nu(\chi_i \cap \chi_j) = 0$. Therefore, $\nu(\bigcup_{i=1}^{d} \chi_i') = \nu(\bigcup_{i=1}^{d} \chi_i') + \nu(\bar{S}) = \nu(\bigcup_{i=1}^{d} \chi_i) = \nu(\chi) = 1$. There exist some i such that $\nu(\chi_i') > 0$. Since $\nu \ll \pi$, we have $\pi(\cup_{i=1}^{d} \chi_i') = \sum_{i=1}^{d} \pi(\chi_i') > 0$. On the other hand, $\pi(\chi_2') = (\pi P)(\chi_2') = \int_\chi \pi(dx) P(x, \chi_2') = \int_{\chi_1'} \pi(dx) = \pi(\chi_1')$, and similarly we have $\pi(\chi_i') = \pi(\chi_j')$ for any i and j, thus $\chi_1', \ldots, \chi_d'$ are subsets of positive π-measure and the Markov chain is periodic of period d, contradicting the assumption of aperiodicity of the chain, so $gcd(T) = 1$. Since T is non-empty, additive and $gcd(T) = 1$, by the fact(see [2], p. 92), there is $n^* \in \mathbf{N}$ such that $\{n^*, n^* + 1, n^* + 2, \ldots\} \subseteq T$. □

Let C be the small set found in Theorem (1.4). In the context of the coupling construction of $\{X_n, Y_n\}$, let $G \subseteq \chi \times \chi$ be the set of (x, y) such that $P_{(x,y)}(\exists n \geq 1; X_n = Y_n) = 1$. If $(X_0, X_0') \equiv (x, X_0') \in G$, then $\lim_{n \to \infty} P[X_n = X_n'] = 1$, since once they are coupled, they will be equal forever from the coupling construction, so that $\lim_{n \to \infty} ||P^n(x, \cdot) - \pi(\cdot)|| = 0$. Therefore it suffices to show $P[(x, X_0') \in G] = 1$. Let $G_x = \{y \in \chi; (x, y) \in G\}$ for $x \in \chi$, and let $\bar{G} = \{x \in \chi; \pi(G_x) = 1\}$. With the results from Claim (1) and Lemma (1.5), we can now finish the proof of Theorem (1.3) by proving the following claim:

Claim 2. $\pi(\bar{G} = 1)$.

Proof. First we prove that $(\pi \times \pi)(G) = 1$. From the way of finding C in Theorem (1.4), $\nu(C) > 0$. From Lemma (1.5), we know for any $x \in \chi$, there exist $n(x)$ s.t. $P^{n(x)}(x, C) > 0$. Thus for any $x \in \chi$, the chain has positive probability of eventually hitting C, i.e. $P_x(\tau_C < \infty) \geq P^n(x, C) > 0$. Since when $(x, y) \notin C \times C$, the chains are updated either according to the same distribution, or being updated independently, we can apply the result from Claim (1) for a single chain. Thus for $(\pi \times \pi)$-a.e.$(x, y) \notin C \times C$, the joint chain will reach $C \times C$ with probability 1. Once the chain reaches $C \times C$, then conditional on not coupling, the joint chain will update from $\frac{1}{(1-\epsilon)^2}[P^{n_0}(X_n, \cdot) - \epsilon\nu(\cdot)][P^{n_0}(X_n', \cdot) - \epsilon\nu(\cdot)]$ which must be absolutely continuous with respect to $\pi \times \pi$. Again by Claim (1), the chain will return to $C \times C$ with probability 1. Hence the joint chain will repeatedly return to $C \times C$ with probability 1, until such time $X_n = X_n'$(since within $C \times C$, the probability of coupling is positive $\geq \epsilon$). Hence eventually we will have $X_n = X_n'$, thus proving that $(\pi \times \pi)(G) = 1$. If $\pi(\bar{G}) < 1$, then

$$(\pi \times \pi)(G^c) = \int_\chi \pi(dx) \pi({G_x}^c) \geq \int_{\bar{G}^c} \pi(dx)[1 - \pi(G_x)] > 0,$$

where the last inequality comes from the fact that a positive function ($1-\pi(G_x)$ is positive on $\bar{G}^c$) integrated over a set of positive measure ($\pi(\bar{G}^c)$) is positive. This result contradicts the fact that $(\pi \times \pi)(G) = 1$. □

Since periodic chains sometimes occur in MCMC, the following corollary is another version of Theorem (1.3) with periodicity.

Corollary 1. *If a Markov chain is ϕ-irreducible, with period $d \geq 2$, and has a stationary distribution $\pi(\cdot)$, then for π a.e. $x \in \chi$,*

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

Proof. Suppose the chain has periodic decomposition $\chi_1, \ldots, \chi_d \subseteq \chi$, let $P' = P^d$, then P' is ϕ-irreducible on χ_1: Suppose $A \subseteq \chi_1$ with $\phi(A) > 0$, then there exists $n \in \mathbf{N}$ such that for any $x \in \chi_1$, $P^n(x, A) > 0$; periodicity implies that $n = dm$ for some $m \in \mathbf{N}$, and $(P')^m(x, A) > 0$. Suppose that the stationary distribution on χ_1 is $\pi'(\cdot)$, then again by periodicity, the stationary distribution on χ_i is $\pi' P^{i-1}(\cdot)$ for $1 \leq i \leq d$, and the stationary distribution on χ is therefore

$$\pi(\cdot) = \frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\cdot), \qquad \forall \pi \text{ } a.e. \text{ } x \in \chi.$$

From Proposition 1.2 (c), it suffices to prove when $n = md$ with $m \to \infty$. Without loss of generality we assume $x \in \chi_1$. From Proposition 1.2 (d), we have $||P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)|| \leq ||P^{md}(x,\cdot) - \pi'(\cdot)||$ for $j \in \mathbf{N}$. Then by the triangle inequality of total variation norm,

$$\begin{aligned} ||\frac{1}{d}\sum_{i=md}^{md+d-1} P^i(x,\cdot) - \pi(\cdot)|| &= ||\frac{1}{d}\sum_{j=0}^{d-1} P^{md+j}(x,\cdot) - \frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)|| \\ &\leq \frac{1}{d}\sum_{j=0}^{d-1} ||P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)|| \\ &\leq \frac{1}{d}\sum_{j=0}^{d-1} ||P^{md}(x,\cdot) - \pi'(\cdot)||. \end{aligned}$$

Applying Theorem 1.3 to P', we have $\lim_{m\to\infty} ||P^{md}(x,\cdot) - \pi'(\cdot)|| = 0$ for π' a.e. $x \in \chi_1$. Moreover, since

$$\forall A \subseteq \chi_i, \;\; (\pi' P^{i-1})(A) = \int_{\{x\in\chi_1\}} \pi'(dx) P^{i-1}(x, A) \qquad \& \qquad \pi(\cdot) = \frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\cdot),$$

the union of null sets in $\chi_1, \ldots, \chi_d$ is also a null set with respect to π, thus proving the result for π'-a.e. $x \in \chi_1$ is equivalent to proving for π-a.e. $x \in \chi$. □

Note that since Theorem (1.3) might not hold for a null set of χ with respect to $\pi(\cdot)$, it is worth exploring the behavior of such set in terms of convergence. The following is an example illustrating the failure of convergence of the null set.

Example 1. *Let $\chi = \{1, 2, \dots\}$. Let $P(1, \{1\}) = 1$, and for $x \geq 2$, $P(x, \{1\}) = \frac{1}{x^2}$ and $P(x, \{x+1\}) = 1 - \frac{1}{x^2}$. Then the chain has stationary distribution $\pi(\cdot) = \delta_1(\cdot)$, and it is π-irreducible and aperiodic. if $X_0 = x \geq 2$, then $P[X_n = x + n \text{ for all } n] = \Pi_{j=x}^{\infty}(1 - (\frac{1}{j^2})^2) > 0$, so that $||P^n(x, \cdot) - \pi(\cdot)|| \not\to 0$. Hence Theorem 1.3 holds only for the set $\{1\}$, which is π a.e.*

Claim 3. *If a Markov chain satisfies the assumptions in Theorem (1.3), the chain still has postive probability of escaping the null set.*

Proof. We prove the claim with a contradiction. If the probability were zero of escaping the null set, then the state space would break up into two pieces, the null set S and its complement S^c with respect to $\pi(\cdot)$, and neither of which could reach the other. Indeed, let $M = \{x \in S^c : P(x, S) > 0\}$. Then $\int_{x \in M} \pi(dx)P(x, S) = 0$ since $\int_{x \in M} \pi(dx)P(x, S) \leq \pi(S)$ as π is stationary; then $\pi(x)P(x, S) = 0$ for almost every $x \in M$, thus $\pi(M) = 0$. Therefore $P(x, S) = 0$ for any $x \in S^c$ except for possbily a null set. Moving that possible null set from S^c to S, we form $S^{c\prime}$ and S' that could not reach each other. One of S' and $S^{c\prime}$ must have positive ϕ measure, but there is no n such that $P^n(x, S^{c\prime}) > 0$ for $x \in S'$ nor m such that $P^m(x, S') > 0$ for $x \in S^{c\prime}$, which contradicts with the assumption of ϕ-irreducibility of the chain. □

Following the proof of Claim (3), for $x \in S'$, $P(x, S^{c\prime}) > 0$. Moreover, if $\inf_{x \in S'} P(x, S^{c\prime}) = r > 0$, for any $x \in S'$, we have

$$P_x(\tau_{S^{c\prime}} < \infty) = 1 - P_x(\tau_{S^{c\prime}} = \infty) = 1 - P_x(X_n \in S' \; \forall n) \geq 1 - \lim_{n \to \infty}(1 - r)^n = 1,$$

thus with probability 1 the chain would eventually exit the null set, and would thus converge to $\pi(\cdot)$ from the null set after all. However, if $\inf_{x \in S'} P(x, S^{c\prime}) = 0$, the probability that the chain may not eventually exit the null set. For example, in Example (1), we have $\inf_{x \in \mathbf{N} \setminus \{0\}} P(x, \{1\}) = 0$ and we may fail to converge to the stationary distribution $\pi(\cdot) = \delta_1(\cdot)$ from the null set.

Here is an example that Theorem (1.3) holds for all $x \in \chi$:

Example 2. *If the transition kernels $P(x, \cdot)$ are absolutely continuous with respect to $\pi(\cdot)$(i.e. $P(x, dy) = p(x, y)\pi(dy)$ for some function $p : \chi \times \chi \to [0, \infty)$), then Theorem (1.3) holds true for all $x \in \chi$. Indeed, let S is the complement of the π a.e. set in Theorem 1.3. Since $\pi(S) = P(x, S) = 0$, $P(x, S^c) = 1$. Hence the chain will enter the π a.e. set within at most one step.*

More generally, Theorem (1.3) holds for all $x \in \chi$ when the chain is Harris recurrent:

Definition 11. *Harris Recurrent*
A Markov chain is ***Harris recurrent*** *if for all $B \subseteq \chi$ with $\pi(B) > 0$, and all $x \in \chi$, the chain will eventually reach B from x with probability 1, i.e. $P_x(\tau_B < \infty) = 1$.*

Theorem 1.6. *For a ϕ-irreducible Markov chain with stationary probability distribution $\pi(\cdot)$, if the chain is Harris recurrent, then Theorem (1.3) holds for all $x \in \chi$.*

Proof. Suppose that there is some $A \subseteq \chi$ with $\pi(A) > 0$, and $x \in \chi$ and $N \in \mathbf{N}$, such that $P_x(X_n \notin A \; \forall n \geq N) > 0$. Integrating over choices of X_i, where $0 \leq i \leq N$, this implies there is some $y \in \chi$, with $P_y(\tau_A = \infty) > 0$, contradicting the defintion of Harris recurrent. Thus for all $A \subseteq \chi$ with $\pi(A) > 0$, and all $x \in \chi$, we have $P_x(X_n \in A\ i.o.^6) = 1$. Let the π-a.e. set in Theorem (1.3) be G. Then once the chain reaches G, conditional on the first hitting time τ_G and the corresponding chain value of X_{τ_G}. Then the chain will converge and Theorem (1.3) follows. □

Note that the chain in Example 1 is not Harris recurrent, since $P_x(\exists n : X_n \in \{1\}) < 1$ for $x \geq 2$.

1.3 Ergodicity of Markov Chains

In practice, we always concern how fast the chain converges if it converges to its stationary distribution, thus we introduce several notions about the rate of convergence for Markov chains here:

Definition 12. *Uniform Ergodicity*
*A Markov chain having stationary distribution $\pi(\cdot)$ is **uniformly ergodic** if*

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

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

One equivalence of uniform ergodicity is

Proposition 1.3. *A Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic if and only if $\sup_{x\in\chi} ||P^n(x,\cdot) - \pi(\cdot)|| < \frac{1}{2}$ for some $n \in \boldsymbol{N}$.*

Proof. If the chain is uniformly ergodic, then

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

thus for n that is sufficiently large, $\sup\limits_{x\in\chi} ||P^n(x,\cdot)-\pi(\cdot)|| < \frac{1}{2}$. Conversely, suppose $\sup\limits_{x\in\chi} ||P^n(x,\cdot)-\pi(\cdot)|| < \frac{1}{2}$ for some $n \in \mathbf{N}$, then by Proposition (1.2) (e), we have $t(n) \equiv \beta < 1$, such tht for all $j \in \mathbf{N}$, we have $t(jn) \leq (t(n))^j = \beta^j$. By Proposition (1.2) (c),

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

Thus the chain is uniformly ergodic with $M = \frac{1}{2}\beta^{-1}$ and $\rho = \beta^{\frac{1}{n}}$. □

One significant condition for ergodicity of Markov chains is Doeblin's Condition.

Definition 13. *Doeblin's Condition*
Suppose there exists a probability measure $\nu(\cdot)$ with the property that for some $m \in \boldsymbol{N}, \epsilon < 1, \delta > 0$

$$\nu(A) > \epsilon \Rightarrow P^m(x, A) \geq \delta$$

*for every $x \in \chi$ and $A \in \sigma(\chi)$, it satisfies **Doeblin's Condition**.*

[6] "i.o." stands for infinitely often

We present a theorem showing the equivalence of uniform ergodicity and Doeblin's Condition. The proof can be found in Meyn and Tweedie(pg. 395)[3]

Theorem 1.7. *An aperiodic ϕ-irreducible chain satisfies Doeblin's condition if and only if it is uniformly ergodic.*

Note that When the state space is finite, if a chain is ϕ-irreducible and aperiodic, we can show it automatically satisfies Doeblin's condition. Accordingly, the chain is uniformly ergodic.

Proposition 1.4. *If the state space χ is finite, then an irreducible and aperiodic Markov chain is uniformly ergodic.*

Proof. Suppose χ is finite and the chain $\{X_n\}$ is irreducible and aperiodic. If a Markov chain is irreducible and aperiodic, then for each pair (i, j) of the states, there is a number $N = N(i, j)$ such that $P^n(i, j) > 0$ for all $n \geq N$(see pg. 92 [2]). Since the state is finite, among each pair of possible states, we can take $M = \max\{N(i, j)|(i, j) \in \chi \times \chi\}$ such that $P^M(i, j) \geq \delta > 0$ for all $i, j \in \chi$, where δ is $\min\{P^M(i, j)|(i, j) \in \chi \times \chi\}$, thus Doeblin's condition is satisfied. □

Note that from the fact that if a Markov chain satisfies Doeblin's condition, then it possesses a stationary distribution. (see pg.110 [4]), we know $\{X_n\}$ has a stationary distribution. Since the state is finite, $\lim_{n\to\infty} P^n(i, j) = 0$ for all $i, j \in \chi$ is impossible, thus all states are positive recurrent (see pg. 96 [2]). Thus the existence and uniqueness of the chain's stationary distribution is ensured(see pg. 97[2]). It guarantees that the stationary distribution from Doeblin's condition is unique in the circumstance of Proposition (1.4).

Definition 14. *Geometric Ergodicity*
*A Markov chain having stationary distribution $\pi(\cdot)$ is **geometrically ergodic** if*

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

for some $\rho < 1$ and $M(x) < \infty$ for π-a.e. $x \in \chi$.

Definition 15. *Drift Condition*
*A Markov chain satisfies a **drift condition**(univariate drift condition) if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \chi \to [1, \infty]$ such that for some $C \subseteq \chi$,*

$$PV \leq \lambda V + b\boldsymbol{I}_C,$$

i.e. such that $\int_\chi P(x, dy)V(y) \leq \lambda V(x) + b\boldsymbol{I}_C(x)$ for all $x \in \chi$.

The following three theorems(Theorem (1.8), Theorem (1.9) and Theorem (1.12)) guarantee the uniform ergodicity and the geometric ergodicity of a Markov chain, respectively:

Theorem 1.8. *Consider a Markov chain with invariant probability distribution $\pi(\cdot)$. Suppose the minorisation condition is satisfied for some $n_0 \in \boldsymbol{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, in the special case $C = \chi$(i.e. the entire state space is small). Then the chain is uniformly ergodic, and in fact*

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

Proof. Consider the coupling construction. Since $C = \chi$, every n_0 iterations we have probability of at least ϵ of making $X_n = X'_n$, thus if $n = n_0 m$, we have $P(X_n \neq X'_n) \leq (1-\epsilon)^m$. From the coupling inequality, $||P^n(x,\cdot) - \pi(\cdot)|| \leq (1-\epsilon)^m = (1-\epsilon)^{\frac{n}{n_0}}$. Then from Proposition 1.2(c), we can find the greatest $n' \in \mathbf{N}$ such that $n' \leq n$ and $n' = n_0 k$ for some k, then

$$||P^n(x,\cdot) - \pi(\cdot)|| \leq ||P^{n'}(x,\cdot) - \pi(\cdot)|| \leq (1-\epsilon)^{\frac{n'}{n_0}} = (1-\epsilon)^{\left\lfloor \frac{n}{n_0} \right\rfloor}.$$

□

Note that Theorem (1.8) is a refinement of Theorem 16.2.4 in Meyn and Tweedie[3] for the case that $m > 1$. From the expression of the quantitative bound in Theorem (1.8), we want ϵ large and n_0 small the minorisation condition.

Definition 16. *Bivariate Drift Condition*
Two indenpendent copies of a Markov chain on χ satisfies a ***bivariate drift condition*** *if*

$$\bar{P}h(x,y) \leq \frac{h(x,y)}{\alpha}, \qquad (x,y) \notin C \times C$$

for some function $h : \chi \times \chi \to [1,\infty)$, $C \subseteq \chi$ and some $\alpha > 1$, where

$$\bar{P}h(x,y) \equiv \int_\chi \int_\chi h(z,w) P(x,dz) P(y,dw).$$

On the other hand, suppose the bavariate drift condition is satiesfied. For $(x,y) \in C \times C$, define

$$\bar{R}h(x,y) \equiv \int_\chi \int_\chi (1-\epsilon)^{-2} h(z,w)(P^{n_0}(x,dz) - \epsilon\nu(dz))(P^{n_0}(y,dw) - \epsilon\nu(dw)).$$

The next theorem about geometric ergodicity which deals with a quantitative bound on convergence rate, has the form $||P^n(x,\cdot) - \pi(\cdot)|| \leq g(x,n)$ for some explicit function $g(x,n)$. Such quantitative bound helps us better understand how fast the chain converges to the stationary distribution from any initial values, if the chain satisfies certain conditions:

Theorem 1.9. *For a Markov chain on χ with transition kernel P, suppose there is $C \subseteq \chi$, $h : \chi \times \chi \to [1,\infty)$, a probability distribution $\nu(\cdot)$ on χ, $\alpha > 1$, $n_0 \in N$, and $\epsilon > 0$, such that the minorisation and the bivariate drift condition are satisfied on C. Define*

$$B_{n_0} \equiv \max[1, \alpha^{n_0}(1-\epsilon) \sup_{C \times C} \bar{R}h]. \tag{6}$$

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

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

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

Proof. First assume $n_0 = 1$ in the minorisation condition. Denote B_{n_0} as B. Let

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

and let $\tau_1, \tau_2, \dots$ be the times of the successive visits of $\{(X_n, X'_n)\}$ to $C \times C$. For any integer j such that $1 \leq j \leq k$,

$$P(X_k \neq X'_k) = P(X_k \neq X'_k, N_{k-1} \geq j) + P(X_k \neq X'_k, N_{k-1} < j). \tag{7}$$

The event $\{X_k \neq X'_k, N_{k-1} \geq j\}$ implies that for the first j times that $(X_n, X'_n) \in C \times C$, the chains are updated through (b) of condition 2 in the coupling construction. Thus $P(X_k \neq X'_k, N_{k-1} \geq j) \leq (1-\epsilon)^j$, which bounds the first term in (7).
For the second term in (7), we first define

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

where $k \in \mathbf{N} \cup \{0\}$. We are now going to prove $\{M_k\}$ is a supermartingale.
If $(X_k, X'_k) \notin C \times C$, then $N_k = N_{k-1}$, so

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

Similarly, if $((X_k, X'_k) \in C \times C)$, then $N_k = N_{k-1} + 1$, assuming $X_k \neq X'_k$ ($X_k = X'_k$is a trivial case), we have

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

where the last inequality comes from the definition of B and h.
Combining the two cases, we conclude $\{M_k\}$ is a supermartingle. Next since $B \geq 1$,

$$\begin{aligned} P(X_k \neq X'_k, N_{k-1} < j) &= P(X_k \neq X'_k, N_{k-1} \leq j-1) \\ &\leq P(X_k \neq X'_k, B^{-N_{k-1}} \leq B^{-(j-1)}) \\ &= P(\mathbf{I}(X_k \neq X'_k) B^{-N_{k-1}} \leq B^{-(j-1)}) \\ &\leq B^{(j-1)} E[(\mathbf{I}(X_k \neq X'_k) B^{-N_{k-1}}] \qquad (by\ Markov's\ inequality) \\ &\leq B^{(j-1)} E[(\mathbf{I}(X_k \neq X'_k) B^{-N_{k-1}} h(X_k, X'_k)] \\ &= \alpha^{-k} B^{(j-1)} E[M_k] \leq \alpha^{-k} B^{(j-1)} E[M_0] \qquad (\{M_k\}\ is\ supermartingale) \\ &= \alpha^{-k} B^{(j-1)} E[h_0]. \end{aligned}$$

When $n_0 > 1$, suppose $(X_n, X'_n) \in C \times C$, we do not count the visits to $C \times C$ for "filling in" times $(X_{n+1}, X'_{n+1}), \dots, (X_{n+n_0-1}, X'_{n+n_0-1})$ in condition 2 of the coupling construction. Accordingly, N_k and $\{\tau_i\}$ only records the times to $C \times C$ that are not "filling in" visits. In addition, N_{k-1} becomes N_{k-n_0} in (7) and the definition of $\{M_k\}$. And $\{M_{t(k)}\}$ is a supermartingale ($t(k)$ means the latest time $\leq k$ such that $X_{t(k)}$ is not a "filling in" time), not $\{M_k\}$. With such modifications, the proof follows in the same way as before. □

Lemma 1.10. *Given a small set C and drift function V satisfying the minorisation condition and the univariate drift condition , we can find a small set $C_0 \subseteq C$ such that the minorisation condition and the univariate drift condition still hold true (with the same n_0 and ϵ and b, but with λ replaced by some $\lambda_0 < 1$), and such that the following inequality holds:*

$$\sup_{x \in C_0} V(x) < \infty \tag{8}$$

Proof. Let λ and b be as in $PV \leq \lambda V + b\mathbf{I}_c$. We can choose δ with $0 < \delta < 1 - \lambda$. Let $\lambda_0 = 1 - \delta$, let $K = \frac{b}{1-\lambda-\delta}$, which is positive, and set

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

Then the minorisation condition $P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot)$ still holds for C_0, which is a subset of C, thus C_0 is a small set. For $x \in C_0$, we have $(PV)(x) \leq \lambda V(x) + b = (1-\delta)V(x) - (1-\lambda-\delta)V(x) + b \leq (1-\delta)V(x) + b = \lambda_0 V(x) + b$; and for $x \notin C$, we have$(PV)(x) \leq \lambda V(x) = (1-\delta)V(x) - (1-\lambda-\delta)V(x) \leq (1-\delta)V(x) = \lambda_0 V(x)$. For $x \in C \backslash C_0$, we have $V(x) \geq K$, so we have

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

thus the univariate drift condition holds true with C replaced by C_0 and λ replaced by λ_0. By the construction of C_0, we have $\sup_{x \in C_0} V(x) < \infty$. □

Definition 17. *A subset $C \subseteq \chi$ is petite(or, (n_0, ϵ, ν)-petite) if there exists a positive integer $n_0, \epsilon > 0$, and a probability measure $\nu(\cdot)$ on χ such that*

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

Lemma 1.11. *For an aperiodic ϕ-irreducible Markov chain, all petite sets are small sets.*

The proof can be found in the appendix of [5]

Proposition 1.5. *Suppose the univariate drift condition is satisfied for some $V : \chi \to [1, \infty]$, $C \subseteq \chi$, $\lambda < 1$ and $b < \infty$. Let $d = \inf_{x \in C^c} V(x)$. If $d > \frac{b}{1-\lambda} - 1$, then the bivariate drift condition is satisfied for the same C, with $h(x, y) = \frac{1}{2}[V(x) + V(y)]$ and $\alpha^{-1} = \lambda + \frac{b}{d+1} < 1$.*

Proof. When $(x, y) \notin C \times C$, then either $x \notin C$ or $y \notin C$ or both. Without loss of generality, we can assume $x \notin C$, then $V(x) \geq d$ and $V(y) \geq 1$ for all $y \in \chi$, thus $h(x, y) =$

$\frac{1}{2}(V(x)+V(y)) \geq \frac{1}{2}(1+d)$ and $PV(x)+PV(y) \leq \lambda V(x)+\lambda V(y)+b$. Then

$$\begin{aligned}
\bar{P}h(x,y) &= \int_\chi \int_\chi \frac{1}{2}(V(z)+V(w))P(x,dz)P(y,dw) \\
&= \frac{1}{2}\int_\chi \int_\chi V(z)P(x,dz)P(y,dw) + \frac{1}{2}\int_\chi \int_\chi V(w)P(x,dz)P(y,dw) \\
&= \frac{1}{2}\int_\chi V(z)P(x,dz) + \frac{1}{2}\int_\chi V(w)P(y,dw) \\
&= \frac{1}{2}[PV(x)+PV(y)] \\
&\leq \frac{1}{2}[\lambda V(x)+\lambda V(y)+b] \\
&= \lambda h(x,y) + \frac{b}{2} \leq \lambda h(x,y) + (\frac{b}{2})\frac{h(x,y)}{(1+d)/2} \\
&= [\lambda + \frac{b}{1+d}]h(x,y)
\end{aligned}$$

Furthermore, $d > \frac{b}{1-\lambda} - 1$ implies that $\lambda + \frac{b}{1+d} = \alpha^{-1} < 1$. □

Theorem 1.12. *Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Suppose the minorisation condition is satisfied for some $C \subseteq \chi$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose further that the drift condition is satisfed for some constants, $0 < \lambda < 1$ and $b < \infty$, and a function $V : \chi \to [1, \infty]$ with $V(x) < \infty$ for at least one $x \in \chi$. Then the chain is geometrically ergodic.*

Proof. We set $h(x,y) = \frac{1}{2}(V(x)+V(y))$. From the above Lemma (1.10), we can shrink C so that (8) holds. By the univariate drift condition and the definition of h, we have

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

and thus B_{n_0} in (6) is finite. Let $d = \inf_{C^c} V$. If further assuming $d > \frac{b}{(1-\lambda)} - 1$ holds, then Proposition (1.5) holds, and all the conditions of Theorem (1.9) are satisfied, and therefore we have geometric ergodicity by specifying the initial distribution of X_n and X'_n and updating with the coupling approach. Suppose insdead, $d \leq \frac{b}{(1-\lambda)} - 1$, we then cannot directly apply Proposition (1.5), but we can still enlarge C so that the new value of d satisfies $d > \frac{b}{(1-\lambda)} - 1$, and use aperiodicity to show that C remains a small set. Specifically, we choose any $d' > \frac{b}{(1-\lambda)} - 1$, and let $S = \{x \in \chi; V(x) \leq d'\}$, and set $C' = C \cup S$. Then $\inf_{x \in C'^c} V(x) \geq d' > \frac{b}{(1-\lambda)} - 1$, so the bivariate drift condition holds with C'. By construction of S, $\sup_{x \in C'} V(x) < \infty$, thus $B_{n_0} < \infty$ when replacing C with C'. Next if we can show that C' is a small set, then with Proposition (1.5) and Theorem (1.9) we can still get geometric ergodicity of the chain. By Lemma (1.11), we just need to show that C' is a petite set. We choose N large enough that $r \equiv 1 - \lambda^N d > 0$. Let $\tau_C = \inf\{n \geq 1; X_n \in C\}$ be the first return time to C. Let $Z_n = \lambda^{-n}V(X_n)$, and let $W_n = Z_{\min(n,\tau_C)}$. Then the univariate drift condition implies that W_n is a supermartingale. Indeed if $\tau_C \leq n$, then

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

If $\tau_C > n$, then $X_n \notin C$, so using the univariate drift condition,

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

For $x \in S$, we have

$$\begin{aligned} P[\tau_C \geq N|X_0 = x] &= P[\lambda^{-\tau_C} \geq \lambda^N|X_0 = x] \\ &\leq \lambda^N E[\lambda^{-\tau_C}|X_0 = x] \quad (by\ Markov's\ inequality) \\ &\leq \lambda^N E[\lambda^{-\tau_C}V(X_{\tau_C})|X_0 = x] \quad (by\ V \geq 1) \\ &= \lambda^N E[Z_{\tau_C}|X_0 = x] \\ &\leq \lambda^N E[Z_0|X_0 = x] \quad (by\ \{W_n\}\ is\ supermartingale) \\ &= \lambda^N E[V(X_0)|X_0 = x] \\ &= \lambda^N V(x) \\ &\leq \lambda^N d, \end{aligned}$$

thus $P[\tau_c < N|X_0 = x] \geq 1 - \lambda^N d = r$. On the other hand, since C is small, so that $P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot)$ for $x \in C$. For $x \in S$,

$$\begin{aligned} \sum_{i=1+n_0}^{N+n_0} P^i(x, \cdot) &\geq \sum_{i=1}^{N} \int_C P^i(x, dy)P^{n_0}(y, \cdot) \\ &= \sum_{i=1}^{N} P_x(X_i \in C)\epsilon\nu(\cdot) \\ &\geq P_x[\bigcup_{i=1}^{N}\{X_i \in C\}])\epsilon\nu(\cdot) \\ &= P_x(\tau_c \leq N)\epsilon\nu(\cdot) \\ &\geq r\epsilon\nu(\cdot). \end{aligned}$$

Also recall when $x \in C$, $P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot) \geq r\epsilon\nu(\cdot)$. Therefore, for $x \in S \cup C$, $\sum_{i=1}^{N+n_0} P^i(x, \cdot) \geq \sum_{i=n_0}^{N+n_0} P^i(x, \cdot) \geq r\epsilon\nu(\cdot)$. Thus C' is petite thus small. □

2 Markov Chains Monte Carlo Algorithms

Markov chain Monte Carlo(MCMC) algorithm is a popular way of approximately sampling from complicated probability distributions in high dimensions. The following is a case where MCMC algorithms are considered to be applied.

2.1 Motivations

Given a density function π_u with respect to Lebesgue measure, on some state space χ, which is possibly unnormalised but at least satisfies $0 < \int_\chi \pi_u < \infty$. Then the probability measure $\pi(\cdot)$ derived from this density is

$$\pi(A) = \frac{\int_A \pi_u(x)dx}{\int_\chi \pi_u(x)dx}.$$

We want to estimate expectations of a function $f : \chi \to \mathbf{R}$ with respect to $\pi(\cdot)$, i.e.

$$\pi(f) = E_\pi[f(X)] = \frac{\int_\chi f(x)\pi_u(x)dx}{\int_\chi \pi_u(x)dx}.$$

In the context of Bayesian statistical inference, this problem can be described as:

Let $L(\mathbf{y}|\theta)$ be the likelihood function (i.e., density of data $\mathbf{y}$ given unknown parameters θ) of a statistical model, for $\theta \in \chi$. Let the prior density of θ be $p(\theta)$, then the posterior distribution of θ given $\mathbf{y}$ is the density

$$\pi_u(\theta) \propto L(\mathbf{y}|\theta)p(\theta),$$

and we want the posterior mean of any functional f, i.e.

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

A usual way of solving the above problem is the classical Monte Carlo method: We simulate i.i.d random variables $Z_1, Z_2, \ldots, Z_N \sim \pi(\cdot)$, and then estimate $\pi(f)$ by

$$\hat{\pi}_1(f) = \frac{1}{N}\sum_{i=1}^{N} f(Z_i).$$

$\hat{\pi}_1(f)$ is an unbiased estimate of $\pi(f)$, which has $Var\left(\hat{\pi}_1(f) - \pi(f)\right) = \frac{{\sigma_f}^2}{N}$, where ${\sigma_f}^2 = Var_\pi(f)$. Then by classical Central Limit Theorem, we have the error $\hat{\pi}_1(f) - \pi(f)$ converges to a normal distribution. However, if π_u is complicated or χ is high-dimensional, then direct integration for the normalising constant is very difficult. Thus, it is infeasible to directly simulate i.i.d random variables from $\pi(\cdot)$.

The difficulty in the practice of the direct approach can be avoided by MCMC algorithm. The MCMC solution construct a Markov chain on χ which can be easily run on a computer, and which has $\pi(\cdot)$ as a stationary distribution. More Specifically, we define a easily-simulated Markov chain transition probability kernel $P(x, y)$ for $x, y \in \chi$, such that $\int_{x\in\chi} \pi(dx)P(x, dy) = \pi(dy)$. Suppose the chain satisfies certain conditions (for example, the conditions in Theorem (1.3)), if we run the Markov chain for a long time (started from anywhere on χ), then for a large n, the distribution of X_n will be approximately stationary: $P^n(x,\cdot) \approx \pi(\cdot)$. We can then set $Z_1 = X_n$, and then restart and return the Markov chain to obtain Z_2, Z_3, etc., and then do estimates as in the direct approach. In practice, by the 'strong law of large numbers" described in Theorem (1.3)), we can also estimate $\pi(f)$ by

$$\hat{\pi}_2(f) = (N-B)^{-1} \sum_{i=B+1}^{N} f(X_i).$$

where B is the "burn-in" time such that $P^B(x,\cdot) \approx \pi(\cdot)$. $\hat{\pi}_2(f)$ can be computed more efficiently than $\hat{\pi}_1(f)$, but finding an appropriate "burn-in" time is difficult(for example, pseudo-convergence).

2.2 Metropolis Hastings Algorithm

Suppose again there is $\pi(\cdot)$ having a density π_u with respect to Lebesgue measure defined χ,

Definition 18. ***Metropolis Hastings Algorithm***

Let $Q(x,\cdot)$ be an easily-simulated Markov chain, whose transition kernel also has a density with respect to Lebesgue measure, i.e.

$$Q(x,dy) \propto q(x,y)dy.$$

Suppose $\pi(\cdot)$ is not concentrated at a single state, then Metropolis-Hastings algorithm proceeds as follows:

1. Choose some X_0(by an initial distribution defined on χ or just any point on χ)
2. Given X_n, generate a proposal Y_{n+1} from $Q(X_n,\cdot)$.
3. Flip an independent coin, whose probability of heads equals $\alpha(X_n, Y_{n+1})$, where

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

To avoid ambiguity, we set $\alpha(x,y) = 1$ whenever $\pi_u(x)q(x,y) = 0$.
4. If the coin is a head, we accept the proposal by setting $X_{n+1} = Y_{n+1}$; if the coin is tail, we then reject the proposal by setting $X_{n+1} = X_n$.
5. Replace n by $n+1$ and repeat.

Note that in practice, one can never toss a coin with probability α defined above; instead, we replace the above Step 3 and Step 4 as the following:

3'. Choose $U_{n+1} \sim$ Uniform$[0,1]$
4'. If $U_{n+1} < \alpha$, then set $X_{n+1} = Y_{n+1}$(accept). Otherwise set $X_{n+1} = X_n$(reject).

Since $P(U < \alpha) = \alpha$, this replacement has the same effect as the theoretical approach of tossing the special coin.

It is worth taking a closer look at $\alpha(x,y)$: If $\pi_u(x)q(x,y) = 0$, since the proposal y must satisfy $q(x,y) > 0$ w.p.1, because $q(x,\cdot)$ is the conditional density of y given x, we must have $\pi_u(x) = 0$ w.p.1. In this case, since $\alpha = 1$, the updating rule automatically almost surely rejects the "bad" x (since $\pi_u(x) = 0 w.p.1$) and accepts the new proposal. However, to avoid the situation that the new proposal y gets stuck in the set $\{x \in \chi : \pi_u(x) = 0\}$ too long, Tierney [6] set $E^+ = \{x : \pi_u(x) > 0\}$ and regulated that $Q(x, E^+) = 1$ for all $x \in E^{+^c}$. Such restriction makes sure that the chain will enter E^+ after at most one step. Furthermore, if $\pi_u(x)q(x,y) > 0$ and $\pi_u(y) = 0$, then $\alpha(x,y) = 0$. Thus once the chain is in set E^+, then almost surely it will not leave the set. Another property of E^+ is that $\pi(E^{+^c}) = 0$, since $\pi_u(x) = 0$ for all x in E^{+^c}.

The Metropolis transition kernel is

$$P(x, A) = \int_{y \in A} \alpha(x,y)q(x,dy) + \delta_x(A) \int_{u \in \chi} (1 - \alpha(x,u))q(x,du), \ x \in \chi, \ A \subseteq \chi, \tag{9}$$

or equivalently,

$$P(x, A) = (1 - r(x))M(x, A) + r(x)\delta_x(A), \ x \in \chi, \ A \subseteq \chi, \tag{10}$$

where $\delta_x(\cdot)$ is a point-mass at x, $\int_\chi (1 - \alpha(x,u))q(x,du)$ is the probability of rejecting when starting at $X_n = x$ and $M(x, \cdot)$ is the kernel conditional on moving.

Such Metropolis procedure ensures that the transition kernel we construct has $\pi(\cdot)$ as a stationary distribution and its reason is the following:

Proposition 2.1. *The Metropolis-Hastings algorithm produces a Markov chain $\{X_n\}$ which has stationary distribution $\pi(\cdot)$.*

Proof. By Proposition (1.1), it suffices to show $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$. Assume $x \neq y$ and set $c = \int_\chi \pi_u(x)dx$,

$$\begin{aligned}\pi(dx)P(x,dy) &= [c^{-1}\pi_u(x)dx][q(x,y)\alpha(x,y)dy] \\ &= c^{-1}\pi_u(x)q(x,y)\min[1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}]dxdy \\ &= c^{-1}\min[\pi_u(x)q(x,y), \pi_u(y)q(y,x)]dxdy\end{aligned}$$

which is symmetric in x and y. If $x = y$, automatically $\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$. □

Therefore, if the Markov chain from Metropolis-Hastings algorithm converges to $\pi(\cdot)$ in the end, one obvious advantage of MCMC is that we only need to compute the ratios of densities, i.e. $\frac{\pi_u(y)}{\pi_u(x)}$, so we do not need to compute the normalising constants $c = \int_\chi \pi_u(x)dx$.

The common classes of proposal densities with respect to Lebesgue measure are the following:

- Symmetric Metropolis Algorithm. $q(x,y) = q(y,x)$
- Random Walk Metropolis-Hastings. $q(x,y) = q(y - x)$
- Independence Sampler $q(x,y) = q(y)$, i.e. $Q(x, \cdot)$ does not depend on x.

Yet, not every Markov chain we build can converge to the desired $\pi(\cdot)$, certain properties of the chains are required. For instance, if the Markov chain from Metropolis Hastings algorithm satisfies the conditions(ϕ-irreducibility and aperiodicity) in Theorem (1.3), then it will converges to $\pi(\cdot)$. We now present an example with common context of running a Metropolis-Hastings algorithm.

Example 3. *Suppose that $\pi(\cdot)$ is a probability measure having unnormalised density function π_u with respect to d-dimension Lebesgue measure. Consider the Metropolis-Hastings algorithm for π_u with proposal density $q(x,\cdot)$ with respect to d-dimensional Lebesgue measure. Then if $q(\cdot,\cdot)$ is positive and continuous on $\boldsymbol{R}^d \times \boldsymbol{R}^d$, and π_u is finite everywhere, then the algorithm is π-irreducible and aperiodic.*

Proof. Pick a set A such that $\pi(A) > 0$. Then there exists $R > 0$ such that $\pi(A_R) > 0$, where $A_R = A \cap B_R(\mathbf{0})$, and $B_R(\mathbf{0})$ is the ball of radius R centered at $\mathbf{0}$. For any $x \in R^d$, since B_R is compact and $q(\cdot,\cdot)$ is continuous and positive, we have

$$\inf_{y\in A_R} \min\{q(x,y), q(y,x)\} \geq \inf_{y\in B_R} \min\{q(x,y), q(y,x)\} = \min_{y\in B_R}\{q(x,y), q(y,x)\},$$

thus $\inf_{y\in A_R} \min\{q(x,y), q(y,x)\} \geq \epsilon$ for some $\epsilon > 0$. Also, by (9) we have

$$P(x,A) \geq P(x,A_R) \geq \int_{A_R} q(x,y) \min[1, \frac{\pi_u(y)q(x,y)}{\pi_u(x)q(x,y)}]dy,$$

$$\begin{aligned}\int_{A_R} q(x,y) \min[1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}]dy &= \int_{A_R} \frac{1}{\pi_u(x)} \min[\pi_u(x)q(x,y), \pi_u(y)q(y,x)]dy \\ &\geq \frac{\epsilon}{\pi_u(x)} \int_{A_R} \min[\pi_u(x), \pi_u(y)]dy.\end{aligned}$$

When $\pi_u(y) \geq \pi_u(x)$,

$$\frac{\epsilon}{\pi_u(x)} \int_{A_R \cap \{y:\pi_u(y)\geq\pi_u(x)\}} \pi_u(x)dy = \epsilon Leb(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}).$$

When $\pi_u(y) < \pi_u(x)$, let $K = \int_{\chi} \pi_u(x)dx > 0$,

$$\begin{aligned}\frac{\epsilon}{\pi_u(x)} \int_{A_R \cap \{y\in A_R:\pi_u(y)<\pi_u(x)\}} \pi_u(y)dy &= \frac{\epsilon K}{\pi_u(x)} \frac{\int_{\{y\in A_R:\pi_u(y)<\pi_u(x)\}} \pi_u(dy)}{\int_{\chi} \pi_u(x)dx} \\ &= \frac{\epsilon K}{\pi_u(x)} \pi(\{y \in A_R : \pi_u(y) < \pi_u(x)\})\end{aligned}$$

Combining the two cases, we get $\int_{A_R} q(x,y) \min[1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}]dy \geq$

$$\epsilon Leb(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}) + \frac{\epsilon K}{\pi_u(x)} \pi(\{y \in A_R : \pi_u(y) < \pi_u(x)\}), \tag{11}$$

Since $\pi(\cdot)$ is absolutely continuous with respect to Lebesgue measure, $\pi(A_R) > 0$ implies $Leb(A_R) > 0$. Assume $Leb(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}) = 0$. Then if $\pi(\{y \in A_R : \pi_u(y) < \pi_u(x)\}) = 0$, we have $\pi(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}) > 0$ and $Leb(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}) = 0$, contradiction. Similarly, assume $\pi(\{y \in A_R : \pi_u(y) < \pi_u(x)\}) = 0$. Then if $Leb(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}) = 0$, then we have $\pi(\{y \in A_R : \pi_u(y) \geq \pi_u(x)\}) = 0$

and $\pi(\{y \in A_R : \pi_u(y) < \pi_u(x)\}) > 0$, thus $Leb(\{y \in A_R : \pi_u(y) < \pi_u(x)\}) > 0$, again contradiction. So it follows that the two terms in (11) cannot be both 0, so we must have $P(x, A) > 0$, and thus the chain is thus π-irreducible.

We now show the Markov chain in this example is aperiodic. We prove it by contradiction. Suppose that χ_1 and χ_2 are disjoint subsets of χ both of positive π measure, with $P(x, \chi_2) = 1$ for all $x \in \chi_1$. But for any $x \in \chi_1$, by (9), we have

$$P(x, \chi_1) \geq \int_{y \in \chi_1} q(x, y)\alpha(x, y)dy > 0,$$

which is a contradiction. □

Note that Theorem (1.3) does not require a specifitc choice ϕ. As long as the chain is ϕ-irreducible for some non zero σ finite measure ϕ. For instance, in the above example, we have π-irreducible.
Here is another example of the Metropolis-Hasting algorithm whose Markov chain is Harris recurrent:

Example 4. *If any ϕ-irreducible Metropolis-Hastings algorithm whose propsal distributions $Q(x, \cdot)$ are absolutely continuous with respect to $\pi(\cdot)$, then its Markov chain is Harris recurrent.*

Proof. Indeed, since the chain is ϕ-irreducible and $\pi(\{x\}) < 1$ for all $x \in \chi$, we have the probability of rejecting $r(x) < 1$ when starting at $X_n = x$ for any $x \in \chi$. Suppose $\pi(A) = 1$. By absolute continuity $\int_{A^c} q(x, y)\alpha(x, y)\pi(dy) \leq \pi(A^c) = 0$, thus $\int_A q(x, y)\alpha(x, y)\pi(dy) = 1$. Since $r(x) < 1$, $\lim_{n\to\infty} r(x)^n = 0$, this means that the chain will eventually move according to $Q(x, \cdot)$ and at which point it will necessarily enters A. Thus $P_x(\tau_A < \infty) = 1$. By the equivalence statement of Harris recurrence in Roberts and Rosenthal(Theorem (6) [7]), the Markov chain is Harris recurrent. Moreover, by Theorem (1.6), thus Theorem (1.3) holds for all $x \in \chi$. □

Note that the proof used in the Example 4 applies the proof in Roberts and Rosenthal(Theorem (8),[7]). The difference between the assumptions in Example 4 and Theorem (8)[7] is that, instead of directly stating $Q(x, \cdot) \ll \pi(\cdot)$ as in the example, Theorem (8)[7] proposed some reference measure $\nu(\cdot)$ that $\pi(\cdot)$ is absolutely continuous to, such that for a function $f : \chi \to \mathbf{R}$

$$\int_\chi f(x)\nu(dx) < \infty \qquad and \qquad \pi(A) = \frac{\int_A f(x)\nu(dx)}{\int_\chi f(x)\nu(dx)}, \qquad \forall A \subseteq \chi.$$

Furthermore,Theorem (8)[7] assumed that $f > 0$ on χ. If $\pi(A) = 0$, since $f > 0$ then $\nu(A) = 0$, thus $\nu \ll \pi$. Then If $Q(x, \cdot) \ll \nu(\cdot)$, then $Q(x, \cdot) \ll \pi(\cdot)$ by the transitive property of absolute continuity.

2.3 Gibbs Sampler

Another MCMC method is Gibbs sampler. Suppose that $\pi_u(\cdot)$ is d-dimensional density with χ an open subset of R^d, and $x = (x_1, \ldots, x_d)$. The i^{th} component of Gibbs sampler is defined such that P_i leaves all components besides i unchanged, and replaces the i^{th}

component by a draw from the full conditional distribution of $\pi(\cdot)$ conditional on all the other components. Specifically, let

$$S_{x,i,a,b} = \{y \in \chi; y_j = x_j \ for \ j \neq i, \ and \ a \leq y_i \leq b\}$$

Then

$$P_i(x, S_{x,i,a,b}) = \frac{\int_a^b \pi_u(x_1, \dots, x_{i-1}, t, x_{i+1}, \dots, x_n)dt}{\int_{-\infty}^{\infty} \pi_u(x_1, \dots, x_{i-1}, t, x_{i+1}, \dots, x_n)dt}.$$

P_i has $\pi(\cdot)$ as a stationary distribution for any $i \in \{1, \dots, d\}$, since

Proposition 2.2. *P_i is reversible with respect to $\pi(\cdot)$,*

Proof.

$$\pi(dx)P_i(x, dy) = \frac{\pi_u(x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_n)dx}{\int_\chi \pi_u(x)dx} \frac{\pi_u(x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots, x_n)dy}{\int\limits_{-\infty}^{\infty} \pi_u(x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n)dt} \tag{12}$$

$$\pi(dy)P_i(y, dx) = \frac{\pi_u(x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots, x_n)dy}{\int_\chi \pi_u(x)dx} \frac{\pi_u(x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_n)dx}{\int\limits_{-\infty}^{\infty} \pi_u(x_1, \cdots, x_{i-1}, t, x_{i+1}, \cdots, x_n)dt}, \tag{13}$$

thus $\pi(dx)P_i(x, dy) = \pi(dy)P_i(y, dx)$. □

The full Gibbs sampler is constructed out of the various P_i, by combining them in one of the following two ways:

- The Deterministic-scan Gibbs Sampler: $P = P_1P_2, \dots, P_d$, i.e. the algorithm applies first the chain P_1, then P_2, until P_d, then goes back to P_1 and start the loop again.
- The Random-scan Gibbs Sampler: $P = \frac{1}{d}\sum_{i=1}^{d} P_i$, i.e. the algorithm does one of the d different Gibbs sampler components, chosen uniformly at random.

In both above two methods, the combined chains P also has π as a stationary distribution.

Gibbs sampler is usually called a special case of the Metropolis-Hastings algorithm in the sense that P_i as the proposal distribution for Metropolis algorithm, has $\alpha(x, y) = 1$. From equations (12) and (13), $\frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)} = 1$.

Note that it may not always hold true that if all P_is are reversible with resepct to π, then P is also reversible with respect to π. Yet this is not a serious problem, since as long as P and P_is have stationary distribution $\pi(\cdot)$, we can check if the assumptions in Theorem (1.3) are satiesfied to examine if the Markov chain in Gibbs sampler converges to $\pi(\cdot)$. The following is an example [7]of Gibbs sampler that satisfies ϕ-irreducibility and aperiodicity in its Markov chain.

Example 5. *$Y_1, \cdots, Y_m$ are iid $N(\mu, \theta)$ and the prior for (μ, θ) is proportional to $\frac{1}{\sqrt{\theta}}$. The goal is to find, $\pi(\cdot, \cdot)$, the joint posterior distribution of μ and θ.*

[7] This example comes from Jones and Hobert[8]

We first find each component of the Gibbs sampler in our circumstance. Since the posterior density is proportional to the product of the prior distribution and likelihood function,

$$\pi(\mu,\theta|\mathbf{y}) \propto \theta^{-\frac{(m+1)}{2}} exp\left(-\frac{1}{2\theta}\sum_{j=1}^{m}(y_j-\mu)^2\right), \qquad \mathbf{y}=(y_1,\ldots,y_m)^T.$$

The Gibbs sampling requires conditional distribution: $\pi(\mu|\theta,\mathbf{y})$ and $\pi(\theta|\mu,\mathbf{y})$. We calculate them as the following:

$$\pi(\mu|\theta,\mathbf{y}) \propto exp\left(-\frac{1}{2\theta}\sum_{j=1}^{m}(y_j-\mu)^2\right) \propto exp\left(-\frac{m}{\theta}(\mu^2-2\mu\bar{y})\right),$$

thus $\pi(\mu|\theta,\mathbf{y}) \sim N(\bar{y},\frac{\theta}{m})$.

$$\pi(\theta|\mu,\mathbf{y}) \propto \theta^{-\frac{(m+1)}{2}} exp\left(-\frac{1}{2\theta}\sum_{j=1}^{m}(y_j-\mu)^2\right) = \theta^{-\frac{(m+1)}{2}} exp\left(\frac{\sum_{j=1}^{m}(y_j-\bar{y})^2+m(\bar{y}-\mu)^2}{2}\right),$$

thus $\pi(\theta|\mu,\mathbf{y}) \sim IG(\frac{m-1}{2},\frac{\sum_{j=1}^{m}(y_j-\bar{y})^2+m(\bar{y}-\mu)^2}{2})$.[8]

We regulate that we update θ first, i.e. if (θ',μ') denote the the current state and (θ,μ) denote the next state, then $(\theta',\mu')\to(\theta,\mu')\to(\theta,\mu)$. In this case, the state space is $\chi = R^+\times R$ and the transition kernel is

$$P_{(\theta',\mu')}(\theta,\mu) = \pi(\theta|\mu',\mathbf{y})\pi(\mu|\theta,\mathbf{y}).$$

By such construction we have the desired distribution π as a stationary distribution of the transition kernel, which is verified as the following:

$$\begin{aligned}
&\int_{R^+}\int_R P_{(\theta',\mu')}(\theta,\mu)\pi(\theta',\mu'|\mathbf{y})d\mu'd\theta'\\
&=\int_{R^+}\int_R \pi(\theta|\mu',\mathbf{y})\pi(\mu|\theta,\mathbf{y})\pi(\theta',\mu'|\mathbf{y})d\mu'd\theta'\\
&=\pi(\mu|\theta,\mathbf{y})\int_{R^+}\int_R \pi(\theta|\mu',\mathbf{y})\pi(\theta',\mu'|\mathbf{y})d\mu'd\theta'\\
&=\pi(\mu|\theta,\mathbf{y})\pi(\theta|\mathbf{y})\\
&=\pi(\mu,\theta|\mathbf{y})
\end{aligned}$$

Suppose set $A\subseteq\chi$ such that $\pi(A)>0$. Since π is assumed to be absolutely continuous with respect to Lebesgue measure, $Leb(A)>0$. Thus for any $(\theta',\mu')\in\chi$, since P is strictly positive on χ, we have

$$P\left((\theta',\mu'),A\right) = \int_A P_{(\theta',\mu')}(\theta,\mu)d\mu d\theta > 0.$$

Therefore, the probability of moving from any point in the state space to any set with positive π-measure in one step is positive, thus the chain is π-irreducible and aperiodic.

[8] $W \sim IG(\alpha,\beta)$ if its density is proportional to $w^{-(\alpha+1)}e^{-\frac{\beta}{w}}\mathbf{I}(w>0)$.

3 Applications

3.1 Another Quantitative Convergence Rate Theorem

Before presenting Theorem (1.9) in 2002, Rosenthal presented another theorem about quantitative convergence rate in 1993(see [9]), which was published in 1995.

Theorem 3.1. *Suppose Markov chain $P(x, dy)$ on χ satisfies the minorisation condition on $C \subseteq \chi$(R is (k_0, ϵ, Q)-small), i.e. $P^{k_0}(x, \cdot) \geq \epsilon Q(\cdot)$ for all $x \in C$. Under the context of the coupling approach, there is $\alpha > 1$ and a function $h : \chi \times \chi \to C$ such that $h \geq 1$ and*

$$E[h(X^{(1)}, Y^{(1)})|X^{(0)} = x, Y^{(0)} = y] = \bar{P}h(x, y) \leq \frac{h(x, y)}{\alpha},$$

for all $(x, y) \notin C \times C$, i.e. it satisfies the bivariate drift condition. Set

$$A \equiv \sup_{(x,y)\in C\times C} E(h(X^{k_0}, Y^{k_0})|X^{(0)} = x, Y^{(0)} = y) = \sup_{(x,y)\in C\times C} \bar{P}h(x, y).$$

Then if $\nu = L(X^{(0)})$ is the initial distribution and π is a stationary distribution, then for any $j > 0$, the total variational distance of π after k steps satisfies

$$||L(X^{(k)} - \pi)|| \leq (1 - \epsilon)^{\left\lfloor \frac{j}{k_0} \right\rfloor} + \alpha^{-k+jk_0-1} A^{j-1} E_{\nu\times\pi}(h(X^{(0)}, Y^{(0)})).$$

Proof. Let

$$t_1 = \inf\{m : (X^{(m)}, Y^{(m)}) \in C\times C\};\ t_i = \inf\{m : m \geq t_{i-1}+k_0, (X^{(m)}, Y^{(m)}) \in C\times C\}\ (i > 1).$$

$$N_k = \max\{i : t_i < k\}. \quad r_i = t_i - t_{i-1}\ (r_1 = t_1).$$

Then for any $\alpha > 1$,

$$P(N_k < j) = P(r_1 + \cdots + r_j > k) = P(\alpha^{r_1+\cdots+r_j} > \alpha^k) \leq \alpha^{-k} E\left(\Pi_{i=1}^{j} \alpha_i^r\right),$$

where the last inequality uses Markov's inequality. Define $g_i(k)$

$$\begin{cases} \alpha^k h(X^{(k)}, Y^{(k)}) & k \leq t_i \\ 0 & k > t_i. \end{cases}$$

Then for $t_{i-1} + k_0 \leq k \leq t_i$,

$$\begin{aligned} E[g_i(k)] = \alpha^k E[h(X^{(k)}, Y^{(k)})] &= \alpha^k E[E[h(X^{(k)}, Y^{(k)}|X^{(k-1)}, Y^{(k-1)})]] \\ &\leq \alpha^{k-1} E[h(X^{(k-1)}, Y^{(k-1)})] \\ &= E[g_i(k-1)], \end{aligned}$$

where the inequality comes from the assumption of the theorem. Thus $g_i(k)$ has non-increasing expectation as a function of k, at least for $k \geq t_{i-1} + k_0$. Since $r_1 = t_1 \geq t_0 + k_0$ and $h \geq 1$, we then have

$$E[\alpha^{r_1}] \leq E[g_1(r_1)] \leq E[g_1(0)],$$

and

$$\begin{aligned}
E\left(\alpha^{r_i}|X^{(t_{i-1})}, Y^{(t_{i-1})}\right) &= E\left(\alpha^{t_i - t_{i-1}}|X^{(t_{i-1})}, Y^{(t_{i-1})}\right) \\
&\leq E\left(\alpha^{t_i} g_i(t_i)|X^{(t_{i-1})}, Y^{(t_{i-1})}\right) \qquad (by\ h \geq 1) \\
&\leq E\left(\alpha^{t_i} g_i(t_i + k_0)|X^{(t_{i-1})}, Y^{(t_{i-1})}\right) \qquad (by\ t_i \geq t_{i-1} + k_0) \\
&= \alpha^{k_0} E\left(h(X^{(t_{i-1}+k_0)}, Y^{(t_{i-1}+k_0)})|X^{(t_{i-1})}, Y^{(t_{i-1})}\right) \\
&\leq \alpha^{k_0} \sup_{(x,y)\in C\times C} E\left(h(X^{(1)}, Y^{(1)})|X^{(0)} = x, Y^{(0)} = y\right).
\end{aligned}$$

Then based on the result of Theorem (1) of Rosenthal[9], we have

$$||L(X^{(k)}) - L(Y^{(k)})|| \leq (1-\epsilon)^{\left\lfloor \frac{j}{k_0} \right\rfloor} + P(N_{k-k_0+1} < j).$$

With the above results, we have

$$\begin{aligned}
P(N_{k-k_0+1} < j) &\leq \alpha^{-k+k_0-1} E\left(\Pi_{i=1}^{j} \alpha_i^{r}\right) = \alpha^{-k+k_0-1} E(\alpha^{r_1}) \Pi_{i=2}^{j} E(\alpha^{r_i}|r_1, \ldots, r_{i-1}) \\
&\leq \alpha^{-k+k_0-1} \alpha^{(j-1)k_0} E\left(h(X^{(0)}, Y^{(0)})\right) A^{j-1} \\
&= \alpha^{-k+jk_0-1} A^{j-1} E\left(h(X^{(0)}, Y^{(0)})\right).
\end{aligned}$$

□

Note that the assumptions in Theorem (1.9) and Theorem (3.1) are the same with their upper bound of convergence in different expressions. It is interesting to investigate which bound behaves better. Thus the goals of the following two examples are to demonstrate how Theorem (1.9) and Theorem (3.1) can be applied and to compare the two upper bounds from the two theorems.

Before presenting the two examples, we first introduce two lemmas that are applied in our examples.

Lemma 3.2. *Given a positive integer k_0 and a subset $R \subseteq \chi$, then there exists a probability measure $Q(\cdot)$, so that $P^{k_0}(x, \cdot) \geq \epsilon Q(\cdot)$ for all $x \in R$, where $\epsilon = \int_\chi \left(\inf_{x\in R} P^{k_0}(x, dy)\right)$.*

Proof. Define $Q'(\cdot)$ on χ by

$$Q'(A) = \int_{y\in A} \left(\inf_{x\in R} P^{k_0}(x, dy)\right), \qquad A \in \chi.$$

Obviously, Q' is a measure on χ. For any $A \in \chi$, for $x \in R$,

$$P^{k_0}(x, A) = \int_{y\in A} P^{k_0}(x, dy) \geq \int_{y\in A} \inf_{x\in R} P^{k_0}(x, dy) = Q'(A).$$

Assuming $Q'(\chi) > 0$, we set $Q(\cdot) = \frac{Q'(\cdot)}{Q'(\chi)}$ and set $\epsilon = Q'(\chi)$, so that $P^{k_0}(x, \cdot) \geq \epsilon Q(\cdot)$. If $Q'(\chi) = 0$, the result is trivially true. □

Lemma 3.3. *Consider a sequentially-updated Gibbs sampler with n components and the state space is $\chi = \chi_1 \times \cdots \times \chi_n$. Suppose that for some d, conditional on values for $X_1^{(k)}, \ldots, X_d^{(k)}$, the random variables $X_{d+1}^{(k)}, \ldots, X_n^{(k)}$ are independent of all $X_i^{(k')}$ for all $k' < k$. Suppose further that there is $R \subseteq \chi$, $\epsilon' > 0$ and a probability measure $Q'(\cdot)$ on $\chi_1 \times \cdots \times \chi_d$ such that*

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

then there is a probability measure $Q(\cdot)$ on χ such that $P^{k_0}(x, \cdot) \geq \epsilon' Q(\cdot)$ for $x \in R$.

Proof. Define measure $Q(\cdot)$ as follows. Marginally on the first d coordinates, $Q(\cdot)$ agrees with $Q'(\cdot)$. Conditional on the first d coordinates, $Q(\cdot)$ is defined by

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

Then by the independence hypothesis, we have

$$\begin{aligned}
&L\left(({X_1}^{(k_0)}, \ldots, {X_n}^{(k_0)}) | ({X_1}^{(0)}, \ldots, {X_n}^{(0)}) = x\right) \\
&= L\left(({X_1}^{(k_0)}, \ldots, {X_d}^{(k_0)}) | ({X_1}^{(0)}, \ldots, {X_n}^{(0)}) = x\right) L\left(({X_{d+1}}^{(k_0)}, \ldots, {X_n}^{(k_0)} | {X_1}^{(k_0)}, \ldots, {X_d}^{(k_0)}\right) \\
&\geq \epsilon' Q({X_1}^{(k_0)}, \ldots, {X_d}^{(k_0)}) Q({X_{d+1}}^{(k_0)}, \ldots, {X_n}^{(k_0)} | {X_1}^{(k_0)}, \ldots, {X_d}^{(k_0)}) \\
&= \epsilon' Q({X_1}^{(k_0)}, \ldots, {X_n}^{(k_0)})
\end{aligned}$$

□

3.2 Bivariate Normal Model

Suppose we have bivariate normal model $(X_1, X_2) \sim N\left(\binom{\mu}{\mu}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$. Then the conditional distribution is

$$\begin{aligned}
L(X_1 | X_2 = x) &= N(x, 1), \\
L(X_2 | X_1 = x) &= N(\frac{x + \mu}{2}, \frac{1}{2})
\end{aligned}$$

When running the Gibbs sampler, we regulate updating the first component first. We now start to build function h such that all the assumptions in the theorems can be satisfied. First, notice that

$$\begin{aligned}
E[(X_2^{(1)} - \mu)^2 | X_2^{(0)} = x_2] &= E[E\left((X_2^{(1)} - \mu)^2 | X_1^{(1)}\right) | X_2^{(0)} = x_2] \\
&= E\left((\frac{X_1^{(1)} - \mu}{2})^2 + \frac{1}{2} | X_2^{(0)} = x_2\right) \\
&= \frac{1}{4}(x_2 - \mu)^2 + \frac{3}{4}.
\end{aligned}$$

Also, since at each iteration the old value $X_1^{(k)}$ is discarded, our small set C and function h should only refer to the second component. Suppose the two chains we have in the coupling construction are $\{X_n\}$ and $\{Y_n\}$. We can let $h(\mathbf{x}, \mathbf{y}) = 1 + (x_2 - \mu)^2 + (y_2 - \mu)^2$. Then we have

$$\begin{aligned} E[h(X^{(1)}, Y^{(1)})|X_2^{(0)} = x_2, Y_2^{(0)} = y_2] &= 1 + E[(Y_2^{(1)} - \mu)^2|Y_2^{(0)} = y_2] + E[(X_2^{(1)} - \mu)^2|X_2^{(0)} = x_2] \\ &= \frac{1}{4}(x_2 - \mu)^2 + \frac{1}{4}(y_2 - \mu)^2 + \frac{10}{4} \\ &= \frac{1}{4}h(x, y) + \frac{9}{4}. \end{aligned}$$

Let $C = \{\mathbf{x} \in \mathbf{R}^2 | (x_2 - \mu)^2 \leq b\}$ where b is a constant. Then if $(x, y) \notin C \times C$, we have $h(x, y) \geq 1 + b$ and Thus by multiplying both sides with $\frac{9}{4(1+b)}$, we get

$$\frac{9}{4(1+b)}h(x, y) \geq \frac{9}{4}$$

Adding $\frac{1}{4}h(x, y)$ on both sides,

$$E[h(X^{(1)}, Y^{(1)})|X^{(0)} = \mathbf{x}, Y^{(0)} = \mathbf{y}] \leq \frac{10 + b}{4 + 4b}h(x, y)$$

Thus $\alpha = \frac{4+4b}{10+b}$. Since we need $\alpha > 1$, we have $b > 2$. According to Lemma (3.2), we find a way to find the ϵ in the minorisation condition. Assume $k_0 = 1$, we have

$$\begin{aligned} P(x_2, x_2') &= \int_{y \in \mathbf{R}} P(x_2, dy)P(y, x_2') \\ &= N(\frac{x_2 + \mu}{2}, \frac{3}{4}; x_2'), \end{aligned}$$

where $N(a, b; y) = \frac{1}{\sqrt{2\pi b}}e^{\frac{-(y-a)^2}{2b}}$ is the density function of $N(a, b)$ evaluated at y. Thus since $x_2 \in [\mu - \sqrt{b}, \mu + \sqrt{b}]$, let $w = \frac{x_2+\mu}{2} \in [\frac{2\mu-\sqrt{b}}{2}, \frac{2\mu+\sqrt{b}}{2}]$

$$\begin{aligned} \epsilon &= \int_{-\infty}^{\infty} (\inf_{x \in C} N(\frac{x_2 + \mu}{2}, \frac{3}{4}; y))dy \\ &= \int_{-\infty}^{\infty} \left(\inf_{w \in [\frac{2\mu-\sqrt{b}}{2}, \frac{2\mu+\sqrt{b}}{2}]} N(w, \frac{3}{4}; y) \right) dy \\ &= \int_{-\infty}^{\mu} N(\mu + \frac{\sqrt{b}}{2}, \frac{3}{4}; y)dy + \int_{\mu}^{\infty} N(\mu - \frac{\sqrt{b}}{2}, \frac{3}{4}; y)dy \\ &= \int_{-\infty}^{0} N(\frac{\sqrt{b}}{2}, \frac{3}{4}; y)dy + \int_{0}^{\infty} N(-\frac{\sqrt{b}}{2}, \frac{3}{4}; y)dy, \end{aligned}$$

which is independent of μ. Since the stationary distribution for Y_2 is $N(\mu, 1)$, we have $E_\pi(Y_2-\mu)^2 = 1$, so $E_{\nu\times\pi}(h(X^{(0)}, Y^0)) = 2 + E_\nu(X_2-\mu)^2$.

$$\begin{aligned}
A &= \sup_{(x,y)\in C\times C} E(h(X^{(1)}, Y^{(1)})|X^{(0)} = x, Y^{(0)} = y) \\
&= \sup_{(x,y)\in C\times C} \frac{1}{4}h(x,y) + \frac{9}{4} \\
&= \frac{1+2b}{4} + \frac{9}{4} \\
&= \frac{5+b}{2}
\end{aligned}$$

Therefore, from the theorem, we can obtain the quantitative bound for this model, with ϵ, α and A all depending only on b. Futhermore, to make $E_\nu(X_2-\mu)^2$ small, we can let $\nu \sim N(\mu, 0)$, so $E_\nu(X_2-\mu)^2 = 0$. The quantitative bound is then

$$||L(X^{(k)}) - \pi|| \leq (1-\epsilon(b))^j + 2\left(\frac{4+4b}{10+b}\right)^{-k+j-1}\left(\frac{5+b}{2}\right)^{j-1}, \tag{14}$$

Note that in Rosenthal(1993)[9], only one example of set C is given, it is worth finding the optimal subset C(or b) so that the upper bound can be as small as possible. In order to get the minimum quantitative upper bound in Theorem (3.1) in equation (14), we optimize the quantitative bound function f over b and j with given steps k, which is

$$f(j,b) = (1-\epsilon(b))^j + 2\left(\frac{4+4b}{10+b}\right)^{-k+j-1}\left(\frac{5+b}{2}\right)^{j-1},$$

with the restrictions: $j \in \mathbf{N}$, $0 < j \leq k$ and $b > 2$.

In this example we are actually able calculate the total variation distance between the Markov chain and the target distribution. Since

$$P(x_2, x_2') = N(\frac{x_2+\mu}{2}, \frac{3}{4}; x_2'),$$

we have $X_2^{(1)} = \frac{X_2^{(0)}+\mu}{2} + Z_1$ where $Z_1 \sim N(0, \frac{3}{4})$. Then

$$\begin{aligned}
X_2^{(2)} &= \frac{X_2^{(1)}+\mu}{2} + Z_2 \\
&= \frac{(X_2^{(0)}+\mu)/2 + Z_1}{2} + \frac{\mu}{2} + Z_2 \\
&= \frac{X_2^{(0)}}{4} + \frac{3\mu}{4} + \frac{Z_1}{2} + Z_2,
\end{aligned}$$

where Z_1 and Z_2 are i.i.d $N(0, \frac{3}{4})$. By induction, we have

$$X_2^{(k)} = \frac{X_2^{(0)}}{2^k} + \left(1 - \frac{1}{2^k}\right)\mu + \sum_{i=1}^{k} \frac{Z_i}{2^{k-i}},$$

where Z_is are all i.i.d $N(0, \frac{3}{4})$. Since we set $X_2^{(0)} = \mu$, the k-th step transition kernel $P^{(k)}(x_2, x_2')$ is $N(\mu, 1 - \frac{1}{4^k}; x_2')$. By Proposition (1.2)($f$), the total variation distance for each component after k steps of our problem is

$$\begin{aligned}||L(X^{(k)}) - \pi|| &= 1 - \int_{-\infty}^{\infty} \min(N(\mu, 1 - \frac{1}{4^k}; y), N(\mu, 1; y))dy \\ &= 1 - \int_{-\infty}^{\infty} \min(N(0, 1 - \frac{1}{4^k}; y), N(0, 1; y))dy\end{aligned}$$

which is equivalent to

$$||L(X^{(k)}) - \pi|| = 1 - \int_{-\infty}^{\infty} \frac{1}{2}\left[N(0, 1 - \frac{1}{4^k}; y) + N(0, 1; y) - \left|N(0, 1 - \frac{1}{4^k}; y) - N(0, 1; y)\right|\right] dy.$$

Since the assumptions in Theorem (3.1) and Theorem (1.9) are the same, for obtaining the related constants of the bound in Theorem 1.9, we keep our function h the same as before, and thus ϵ and α are unchanged. For computing $\bar{R}h(x, y)$, we need

$$\begin{aligned}\epsilon\nu(A) &= \int_{y \in A} \left(\inf_{x \in C} N(\frac{x_2 + \mu}{2}, \frac{3}{4}; y)\right) dy \\ &= \int_{y \in A} N(\mu + \frac{\sqrt{b}}{2}, \frac{3}{4}; y)\mathbf{I}_{[-\infty,\mu]} + N(\mu - \frac{\sqrt{b}}{2}, \frac{3}{4}; y)\mathbf{I}_{[\mu,\infty]}dy \\ &= \int_{y \in A} N(\frac{\sqrt{b}}{2}, \frac{3}{4}; y)\mathbf{I}_{[-\infty,0]} + N(-\frac{\sqrt{b}}{2}, \frac{3}{4}; y)\mathbf{I}_{[0,\infty]}dy\end{aligned}$$

Thus

$$\epsilon\nu(dy) = \left(N(\frac{\sqrt{b}}{2}, \frac{3}{4}; y)\mathbf{I}_{[-\infty,0]} + N(-\frac{\sqrt{b}}{2}, \frac{3}{4}; y)\mathbf{I}_{[0,\infty]}\right) dy.$$

The following table shows the total variation distance, the bound from Rosenthal(1993), the optimized bound in Theorem (3.1) and the optimized bound in Theorem (1.9) when $k = 10, 50, 100, 200, 500, 750, 1000$ and 2000, computed with the software Mathematica:

k	1	2	3	4
10	4.7283×10^{-4}	0.9408	7.7892×10^{-1}	7.4540×10^{-1}
50	6.1062×10^{-16}	1.928×10^{-1}	1.1813×10^{-1}	8.1953×10^{-2}
100	3.8176×10^{-31}	2.7724×10^{-2}	1.0978×10^{-2}	5.0490×10^{-3}
200	3.0116×10^{-61}	6.2458×10^{-4}	9.5793×10^{-5}	1.9074×10^{-5}
500	0	8.7726×10^{-9}	6.3404×10^{-11}	1.0246×10^{-12}
750	0	8.1983×10^{-13}	4.4933×10^{-16}	8.9543×10^{-19}
1000	0	7.6721×10^{-17}	3.1829×10^{-21}	7.8249×10^{-25}
2000	0	5.8860×10^{-33}	8.0255×10^{-42}	4.5620×10^{-49}

1. represents the total variation distance;
2. represents the bound from Rosenthal (1993);
3. represents the optimal bound from Theorem (3.1);
4. represents the optimal bound from Theorem (1.9).

From the table, we can see that the bound from Theorem (1.9) decreases faster than the bound from Theorem (3.1). In terms of the minimum steps such that the corresponded value to be ≤ 0.01, it takes 6 steps for the total variation distance, 130 steps for the bound from Rosenthal(1993), 102 steps for the optimal bound from Theorem (3.1) and 88 steps for the optimal bound from Theorem (1.9).

In many other problems, numerically computing $\bar{R}h$ in Theorem (1.9) is expensive. Since $\bar{R}h \leq (1-\epsilon)^{-2}\bar{P}h$ and $\bar{P}h$ is relatively easier to compute, we replace B_{n_0} in Theorem (1.9) with

$$B_{n_0} = \max[1, \frac{\alpha^{n_0}}{(1-\epsilon)} \sup_{C \times C} \bar{P}h],$$

and keep the other components of the bound in Theorem (1.9) unchanged. Thus expression for this proposed new(weaker) bound from Theorem (1.9) is

$$||L(X^{(k)}) - \pi|| \leq (1-\epsilon)^j + 2\alpha^{-k}(\max[1, \frac{\alpha}{(1-\epsilon)}A])^{j-1}. \tag{15}$$

The following table shows how well this new(weaker) bound performs compared to the original bound in Theorem (1.9)(the values from the third column of the table below are copied from the fifth column of the table on the last page).

k	1	2
10	8.0722×10^{-1}	7.4540×10^{-1}
50	1.4522×10^{-1}	8.1953×10^{-2}
100	1.7672×10^{-2}	5.0490×10^{-3}
200	2.6126×10^{-4}	1.9074×10^{-5}
500	8.4312×10^{-10}	1.0246×10^{-12}
750	2.2378×10^{-14}	8.9543×10^{-19}
1000	5.9393×10^{-19}	7.8249×10^{-25}
2000	2.9467×10^{-37}	4.5620×10^{-49}

1. represents the optimal new (weaker) bound from Theorem (1.9).
2. represents the optimal originalbound from Theorem (1.9).
This new(weaker) bound takes at least 114 steps to have its value ≤ 0.01.

3.3 Hierarchical Possion Model

The Gibbs sampler for this model is a Markov chain $(\beta^{(k)}, \theta_1^{(k)}, \ldots, \theta_{10}^{(k)})$ on $\chi = (\mathbf{R}^{\geq 0})^{11}$, with updating scheme given by

$$L(\beta^{(k+1)}|\{\theta_j^{(k)}\}) = G\left(\gamma + 10\alpha_0, \delta + \sum_{j=1}^{10}\theta_j^{(k)}\right),$$

$$L(\theta_i^{(k+1)}|\beta^{(k+1)}, \{\theta_j^{(k+1)}\}_{j<i}, \{\theta_j^{(k)}\}_{j>i}) = G\left(\alpha_0 + s_i, t_i + \beta^{(k+1)}\right), \qquad (1 \leq i \leq 10)$$

where $G(a,b)$ denotes the gamma distribution with density $b^a x^{a-1} e^{(-bx)}/\Gamma(a)$, where $\alpha_0 = 1.802$, $\gamma = 0.01$ and $\delta = 1$, with the data s_i and t_i as in Gelfand and Smith (1990, Table 3). Let $S^{(k)}$ respresent $\sum_i \theta_i^{(k)}$. Then

$$S^{(k)}|\beta^{(k)} = \sum_{i=1}^{10} Z_i, \qquad Z_i \sim G\left(\alpha_0 + s_i, t_i + \beta^{(k+1)}\right), \qquad Z_i s \textit{ are i.i.d.}$$

Since the t_is are distinct from each other, the scale parameters in the distribution of Z_is are different, and thus the density function for the sum of Z_is is complicated, which can be found in P.G. Moschopoulos[10], which is hard to compute numerically. Thus the transition kernel for $S^{(k)}$ becomes too expensive to compute, so we do not discuss the total variation distance in this example. Since the transition kernel for $S^{(k)}$ is also required in computing $\bar{R}h$ in Theorem (1.9), we only apply the new(weaker) bound from Theorem (1.9), which is introduced in the bivariate normal model example.

We first construct function h. Since at each iteration the old value $\beta^{(k)}$ is discarded , our subset C and function h should only refer to the remaining components $\theta_1^{(k)}, \ldots, \theta_{10}^{(k)}$. Indeed from the expression of the conditional distributions, it is sufficient to refer only to their sum $S^{(k)}$. Since

$$E\left(S^{(k+1)}|\beta^k\right) = \sum_i \frac{\alpha_0 + s_i}{t_i + \beta^{(k)}},$$

A cursory numerical examination of the conditional mean of $S^{(k)}|\beta^{(k)}$ suggests that the value of $S^{(k)}$ roughly approches the value 6.5. Thus, let $X^{(k)} = (\beta^{(k)}, \theta_1^{(k)}, \ldots, \theta_{10}^{(k)})$, $Y^{(k)} = ({\beta'}^{(k)}, {\theta'_1}^{(k)}, \ldots, {\theta'}_{10}^{(k)})$, $S^{(k)} = \sum_i \theta_i^{(k)}$ and $S'^{(k)} = \sum_i {\theta'}_i^{(k)}$. Set

$$h(X^{(k)}, Y^{(k)}) = 1 + (S^{(k)} - 6.5)^2 + (S'^{(k)} - 6.5)^2.$$

Accordingly, define function $e(w)$ by

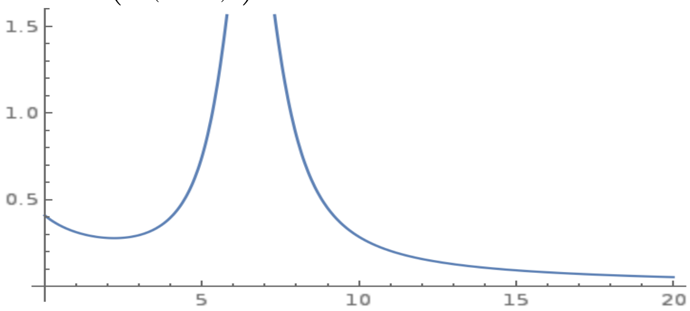
$$\begin{aligned} e(w) &= E((S^{(1)} - 6.5)^2 | S^{(0)} = w) \\ &= \int_0^\infty E\left((S^{(1)} - 6.5)^2 | \beta^{(1)} = x\right) P(\beta^{(1)} = x | S^{(0)} = w) dx \\ &= \int_0^\infty [\left(\sum_i (\frac{\alpha_0 + s_i}{t_i + x}) - 6.5\right)^2 + \sum_i (\frac{\alpha + s_i}{(t_i + x)^2})] G(\gamma + 10\alpha_0, \delta + w; x) dx, \end{aligned}$$

where the last equality uses the fact $Var\left(S^{(k+1)}|\beta^k\right) = \sum\limits_i \frac{\alpha_0 + s_i}{(t_i + \beta^{(k)})^2}$.

Let $C = \{X^{(k)} : 6.5 - r \leq S^{(k)} \leq 6.5 + r\}$, where r is a constant with $r \geq 6.5$(since $w \geq 0$). Define

$$\phi(r) = \sup_{w \notin [6.5 - r, 6.5 + r]} \left(\frac{1 + e(w)}{1 + (w - 6.5)^2} \right).$$

The following plot is w vs. $\left(\frac{1+e(w)}{1+(w-6.5)^2}\right)$:



The plot suggests if $r \geq 2.69$, then $\sup_{w \notin [6.5-r, 6.5+r]} \left(\frac{1+e(w)}{1+(w-6.5)^2}\right) = \frac{1+e(0)}{1+(0-6.5)^2} = 0.4093$; if

$r < 2.69$, the supremum is either obtained on $w = 6.5 - r$ or $w = 6.5 + r$. For finding α, since

$$\begin{aligned}
&\sup_{(x,y)\notin C\times C}\left(\frac{E[h(X^{(1)},Y^{(1)}|X^{(0)}=x,Y^{(0)}=y)]}{h(x,y)}\right)\\
&=\sup_{\{(w_1,w_2)|w_1\notin[6.5-r,6.5+r]\}}\left(\frac{1+e(w_1)+e(w_2)}{1+(w_1-6.5)^2+(w_2-6.5)^2}\right)\\
&\leq\sup_{w_2}\left(\frac{\sup_{w_1\notin[6.5-r,6.5+r]}\frac{1+e(w_1)}{1+(w_1-6.5)^2}}{1+(w_2-6.5)^2/(r^2+1)}+\frac{e(w_2)}{r^2+1+(w_2-6.5)^2}\right)\\
&=\sup_{w_2}\left(\frac{\phi(r)}{1+(w_2-6.5)^2/(r^2+1)}+\frac{e(w_2)}{r^2+1+(w_2-6.5)^2}\right).
\end{aligned}$$

Thus

$$\alpha = 1/\sup_{w}\left(\frac{\phi(r)}{1+(w-6.5)^2/(r^2+1)}+\frac{e(w)}{r^2+1+(w-6.5)^2}\right).$$

For ϵ, by using Lemma (3.2) and Lemma (3.3)(with $d = 1$), we have

$$\begin{aligned}
\epsilon &= \int_0^\infty\left(\inf_{w\in[6.5-r,6.5+r]}G(\gamma+10\alpha_0,\delta+w;x)\right)dx\\
&=\int_0^\infty\min\left(G(\gamma+10\alpha_0,\delta+6.5-r;x),G(\gamma+10\alpha_0,\delta+6.5+r;x)\right)dx
\end{aligned}$$

$A = 1 + 2\sup_{w\in[6.5-r,6.5+r]} e(w)$. Furthermore, set $V(X) = 1 + (S - 6.5)^2$. Since

$$1+e(w)\leq(1+(w-6.5)^2)\sup_{w\notin[6.5-r,6.5+r]}\left(\frac{1+e(w)}{1+(w-6.5)^2}\right)+\sup_{w\in[6.5-r,6.5+r]}(1+e(w)),$$

by setting

$$\lambda=\sup_{w\notin[6.5-r,6.5+r]}\left(\frac{1+e(w)}{1+(w-6.5)^2}\right),\qquad b=\sup_{w\in[6.5-r,6.5+r]}(1+e(w)).$$

Since $b \geq 1$, if $\lambda < 1$(this may not be true if r is very small), then the chain satisfies

$$E[V(X^{(1)}|X^{(0)}=x)]\leq\lambda V(x)+b. \tag{16}$$

By taking expectations of both sides of (16) with respect to π, we have $E_\pi(S'-6.5)^2 \leq \frac{b}{1-\lambda}$. Setting $S^{(0)} = 6.5$, we have

$$E_{\nu\times\pi}[h(X^{(0)},Y^{(0)})]<\frac{b}{1-\lambda}.$$

Note that $\epsilon, \alpha, A, \lambda$ and b all depend on r, so the bound from Theorem (3.1) is indeed a function j, k and r. Because of the difficulty of computation in minimizing the bound over r and j for given k, we instead minimize j given k for a set of different values of r. The following table shows the values of $\epsilon, \alpha, A, \lambda$ and b corresponding to the chosen values of r.

r	ϵ	α	A	λ	b
0.5	7.7782×10^{-1}	3.1938×10^{-1}	4.0182	2.0073	2.5091
1	5.7090×10^{-1}	7.5284×10^{-1}	4.3131	1.3283	2.6566
1.5	3.9205×10^{-1}	7.6547×10^{-1}	4.6752	8.7311×10^{-1}	2.8376
2	2.4906×10^{-1}	1.1217	5.0925	6.0925×10^{-1}	3.0462
2.5	1.4429×10^{-1}	1.5450	5.555	4.5208×10^{-1}	3.2776
3	7.4778×10^{-2}	1.8147	6.0552	4.0930×10^{-1}	3.5276
3.5	3.3756×10^{-2}	1.9365	6.8438	4.0930×10^{-1}	3.9219
4	1.2775×10^{-2}	2.0291	8.5478	4.0930×10^{-1}	4.0704
4.5	3.8269×10^{-3}	2.1002	10.8882	4.0930×10^{-1}	5.9441
5	8.2839×10^{-4}	2.1554	14.1012	4.0930×10^{-1}	7.5506
5.5	1.1075×10^{-4}	2.1988	18.5608	4.0930×10^{-1}	9.7804
6	6.7580×10^{-6}	2.2335	24.9110	4.0930×10^{-1}	12.9555
6.5	9.3370×10^{-8}	2.2615	34.4040	4.0930×10^{-1}	17.7020

Note that when $r = 0.5$ and $r = 1$, $\lambda > 1$ thus (16), thus we do not consider the bounds for these two r values.
The following table provides the optimal bound from Theorem (3.1) for chosen values of r, and $k = 10, 50, 100, 200, 500, 750, 1000$, and 2000.

k	r=2	r=2.5	r=3	r=3.5	r=4
10	3.2223	9.3291×10^{-1}	9.4065×10^{-1}	9.7520×10^{-1}	9.9406×10^{-1}
50	7.0650×10^{-1}	2.9484×10^{-1}	4.4291×10^{-1}	6.9036×10^{-1}	8.7670×10^{-1}
100	3.2390×10^{-1}	6.6223×10^{-2}	1.7363×10^{-1}	4.4719×10^{-1}	7.4675×10^{-1}
200	6.2331×10^{-2}	3.5347×10^{-3}	2.6709×10^{-2}	1.8784×10^{-1}	5.4435×10^{-1}
500	4.7989×10^{-4}	5.2487×10^{-7}	9.7836×10^{-5}	1.3973×10^{-2}	2.0990×10^{-1}
750	8.4974×10^{-6}	3.4285×10^{-10}	9.1713×10^{-7}	1.5984×10^{-3}	9.5191×10^{-2}
1000	1.5029×10^{-7}	2.1711×10^{-13}	8.5033×10^{-9}	1.8303×10^{-4}	4.2945×10^{-2}
2000	1.2995×10^{-14}	3.7002×10^{-26}	6.3437×10^{-17}	3.1681×10^{-8}	1.7951×10^{-3}

k	r=4.5	r=5	r=5.5	r=6	r=6.5
10	1.0022	1.0051	1.0062	1.0071	1.0086
50	9.6371×10^{-1}	9.9277×10^{-1}	9.9930×10^{-1}	1	1
100	9.2120×10^{-1}	9.8365×10^{-1}	9.9800×10^{-1}	1	1
200	8.4175×10^{-1}	9.6547×10^{-1}	9.9562×10^{-1}	1	1
500	6.4113×10^{-1}	9.1296×10^{-1}	9.8866×10^{-1}	1	1
750	5.1127×10^{-1}	8.7152×10^{-1}	9.8283×10^{-1}	1	1
1000	4.0678×10^{-1}	8.3185×10^{-1}	9.7706×10^{-1}	1	1
2000	1.6418×10^{-1}	6.9045×10^{-1}	9.9562×10^{-1}	1	1

Since when $r = 1.5$, the bound increases to infinity as k increases, so we do not present the results of $r = 1.5$ in the tables. In terms of the minimum steps such that the corresponded value ≤ 0.01, it takes 314 steps when $r = 2$, 165 steps when $r = 2.5$, 253 steps when $r = 3$, 540 steps when $r = 3.5$, 1460 steps when $r = 4$, 5083 steps when $r = 4.5$,24780 steps when $r = 5$, about 196423 steps when $r = 5.5$, about 263170000 steps for $r = 6.5$. For $r = 6$, the requires more than 1 billion steps or even more to be ≤ 0.01.

To apply the new(weaker) bound from Theorem (1.9), we use the bound

$$||L(X^{(k)}) - \pi|| \leq (1-\epsilon)^j + \frac{b}{1-\lambda}\alpha^{-k}(\max[1, \frac{\alpha}{(1-\epsilon)}A])^{j-1}.$$

We minimize this bound in the same way as we do for the bound from Theorem (3.1). The following table provides the new(weaker) bound from Theorem (1.9) for chosen values of r, and $k = 10, 50, 100, 200, 500, 750, 1000,$ and 2000.

k	r=2	r=2.5	r=3	r=3.5	r=4
10	3.2223	9.3291×10^{-1}	9.4065×10^{-1}	9.7520×10^{-1}	9.9406×10^{-1}
50	7.5379×10^{-1}	3.0935×10^{-1}	4.6291×10^{-1}	6.9239×10^{-1}	8.7788×10^{-1}
100	3.5318×10^{-1}	7.8764×10^{-2}	1.8381×10^{-1}	4.5532×10^{-1}	7.4764×10^{-1}
200	8.5124×10^{-2}	5.1103×10^{-3}	3.0395×10^{-2}	1.9312×10^{-1}	5.4698×10^{-1}
500	1.1977×10^{-3}	1.3224×10^{-6}	1.3074×10^{-4}	1.4758×10^{-2}	2.1146×10^{-1}
750	3.5320×10^{-4}	1.3470×10^{-9}	1.3788×10^{-6}	1.7485×10^{-3}	9.6031×10^{-2}
1000	1.0016×10^{-6}	1.3496×10^{-12}	1.4931×10^{-8}	2.0519×10^{-4}	4.3527×10^{-2}
2000	6.8329×10^{-13}	1.5355×10^{-24}	1.9463×10^{-16}	3.94377×10^{-8}	1.8458×10^{-3}

k	r=4.5	r=5	r=5.5	r=6	r=6.5
10	1.0022	1.0051	1.0062	1.0071	1.0086
50	9.6376×10^{-1}	9.9277×10^{-1}	9.9934×10^{-1}	1	1
100	9.2138×10^{-1}	9.8365×10^{-1}	9.9800×10^{-1}	1	1
200	8.4179×10^{-1}	9.6549×10^{-1}	9.9562×10^{-1}	1	1
500	6.4121×10^{-1}	9.1297×10^{-1}	9.8866×10^{-1}	1	1
750	5.1133×10^{-1}	8.7152×10^{-1}	9.8284×10^{-1}	1	1
1000	4.0771×10^{-1}	8.3196×10^{-1}	9.7706×10^{-1}	1	1
2000	1.6445×10^{-1}	6.9046×10^{-1}	9.9545×10^{-1}	1	1

The result suggests that the convergence rate of the optimal new(weaker) bound from Theorem (1.9) behaves very similar(a little bit slower) to that of the optimal bound from Theorem (3.1). In terms of the minimum steps such that the corresponded value to be ≤ 0.01, it takes 354 steps when $r = 2$, 175 steps when $r = 2.5$, 262 steps when $r = 3$, 547 steps when $r = 3.5$, 1466 steps when $r = 4$, 5089 steps when $r = 4.5$, 24785 steps when $r = 5$, 196423 steps when $r = 5.5$, about 263170000 steps for $r = 6.5$. For $r = 6$, it requires more than 1 billion steps or even more to be ≤ 0.01.

4 Appendices

Definition 19. *σ-finite A measure space $(\chi, \sigma(\chi), \mu)$ is called finite if $\mu(\chi)$ is a finite real number (rather than ∞), and μ is thus called a σ-finite measure.*

Note that every probability measure is σ-finite.ϕ on $(\chi, \sigma(\chi))$.

Theorem 4.1. *(Radon-Nikodym Theorem) if μ and ν are two σ-finite measures on some measurable space $(\chi, \sigma(\chi))$, then $\mu \ll \nu$ if and only if μ is absolutely continuous with respect to ν.*

Besides studying papers and relavent materials about Markov chains and MCMC, I also learnt the first two chapers of *A First Look at Rigorous Probability Theory* by Rosenthal[2], which focuses on explaining the construction of probability triples. The main theorems are as the following. The proofs can be found in [2], and we omit them here.

Theorem 4.2. *(The Extension Theorem) Let τ be a semialgebra of subsets of sample space Ω. Let $P : \tau \to [0, 1]$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$, satisfying the finite superadditivity property*

$$P(\bigcup_{i=1}^{k} A_i) \geq \sum_{i=1}^{k} P(A_i), \quad \textit{whenever } A_1 \ldots, A_k \in \tau \textit{ and } \bigcup_{i=1}^{k} A_i \in \tau, \textit{ and } \{A_i\} \textit{ are disjoint}, \tag{17}$$

and also the countable monotonicity property that

$$P(A) \leq \sum_n P(A_n) \quad \textit{for } A, A_1, \cdots \in \tau \textit{ with } A \subseteq \bigcup_n A_n.$$

Then there is a σ-algebra M s.t.$\tau \subseteq M$, and a countably additive probability measure P^* on M, such that $P^*(A) = P(A)$ for all $A \in \tau$.

Although Theorem (4.2) is the main tool for proving the existence of probability triples, verying its assumptions is challenging. Thus the following two corollaries provide alternative formulas for the assumptions in Theorem (4.2), which are easier to check.

Corollary 2. *Let τ be a semialgebra of subsets of sample space Ω. Let $P : \tau \to [0, 1]$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$, satisfying (17) in Theorem (4.2), and*

$$P(A) \leq P(B) \quad \textit{whenever } A, B \in \tau \textit{ with } A \subseteq B,$$

and also the "countable subadditivity on τ", i.e.

$$P(\bigcup_n B_n) \leq \sum_n P(B_n), \quad \textit{whenever } B_1, B_2, \cdots \in \tau \textit{ and } \bigcup_n B_n \in \tau.$$

Then there is a σ-algebra M s.t. $\tau \subseteq M$, and a countably additive probability measure P^* on M, such that $P^*(A) = P(A)$ for all $A \in \tau$.

Corollary 3. *Let τ be a semialgebra of subsets of sample space Ω. Let $P : \tau \to [0, 1]$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$,satisfying the countable additivity property that*

$$P(\bigcup_n D_n) \leq \sum_n P(D_n) \quad \textit{for } D_1, D_2 \cdots \in \tau \textit{ disjoint with } \bigcup_n D_n \in \tau.$$

Then there is a σ-algebra M s.t.$\tau \subseteq M$, and a countably additive probability measure P^* on M, such that $P^*(A) = P(A)$ for all $A \in \tau$.

5 Codes

5.1 Codes for Bivariate Normal Model

total variation distance when k=10, 50, 100, 200, 500, 750, 1000, 2000

```
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^10)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^50)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^100)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^500)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^750)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^1000)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
1/2*Integrate[Abs[PDF[NormalDistribution[0,1-Sqrt[1/(4^2000)]],x]
-PDF[NormalDistribution[0,1],x]],{x,-Infinity,Infinity }]
```

optimal quantitative bound from Rosenthal(1993)

```
f[b_]=1/2-Integrate[PDF[NormalDistribution[Sqrt[b]/2,Sqrt[3/4]],x],{x,0,Sqrt[b]/2}]
+1/2-Integrate[PDF[NormalDistribution[-Sqrt[b]/2,Sqrt[3/4]],x],{x,-Sqrt[b]/2,0}]

NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-10+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=10&&j\[Element]Integers},{b,j}]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-50+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=50&&j\[Element]Integers},{b,j}]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-100+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=100&&j\[Element]Integers},{b,j}]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-200+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=200&&j\[Element]Integers},{b,j},
MaxIterations \[Rule] 10000]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-500+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=500&&j\[Element]Integers},{b,j},
MaxIterations \[Rule] 50000]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-750+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=750&&j\[Element]Integers},{b,j},
MaxIterations \[Rule] 50000]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-1000+j-1)*((5+b)/2)^(j-1)),
b>=]2&&0<j<=1000&&j\[Element]Integers},{b,j},
MaxIterations \[Rule] 50000]
NMinimize[{(1-f[b])^j+2*(((4+4*b)/(10+b))^(-2000+j-1)*((5+b)/2)^(j-1)),
b>=2&&0<j<=2000&&j\[Element]Integers},{b,j},
MaxIterations \[Rule] 60000]
```

optimal quantitative bound from Roberts and Rosenthal(2004)

```
h[x_,y_]:=1+x^2+y^2
```

```
 f[b_]:=1/2-Integrate[PDF[NormalDistribution[Sqrt[b]/2,Sqrt[3/4]],x]
 ,{x,0,Sqrt[b]/2}]+1/2-Integrate[PDF[NormalDistribution[-Sqrt[b]/2,
 Sqrt[3/4]],x],{x,-Sqrt[b]/2,0}]
 g[x_,z_]:=PDF[NormalDistribution[-Sqrt[z]/2,Sqrt[3/4]],x]*Boole
 [x\[GreaterEqual]0]+PDF[NormalDistribution[Sqrt[z]/2,Sqrt[3/4]],x]*Boole[x<0]
 Clear[K]
K[v_?NumericQ,d_?NumericQ,p_?NumericQ]:=NIntegrate[(1-f[p])^(-2)
*h[r,u]*(PDF[NormalDistribution[v/2,Sqrt[3/4]],r]-g[r,p]) *
(PDF[NormalDistribution[d/2,Sqrt[3/4]],u]-g[u,p]),{r,-100,100},{u,-100,100}]
 zz[p_]:=NMaximize[{K[v,d,p],-Sqrt[p]<=v<=Sqrt[p]
&&-Sqrt[p]<=d<=Sqrt[p]},{v,d}]
B[r_]:=Max[1,(1-f[r])*(4+4r)/(10+r)*zz[r][[1]]]
 fff[s_,t_,u_]:=(1-f[s])^t+2*((4+4*s)/(10+s))^(-u)*B[s]^(t-1)

 NMinimize[{fff[v,w,10],v>=2&&0<w<=10&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,50],v>=2&&0<w<=50&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,100],v>=2&&0<w<=100&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,200],v>=2&&0<w<=200&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,500],v>=2&&0<w<=500&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,750],v>=2&&0<w<=750&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,1000],v>=2&&0<w<=1000&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
 NMinimize[{fff[v,w,2000],v>=2&&0<w<=2000&&
w\[Element]Integers},{v,w},MaxIterations \[Rule] 60000]
```

optimal new(weaker) bound from Roberts and Rosenthal(2004)

```
 Z[r_]:=Max[1,1/(1-f[r])*(4+4r)/(10+r)*(5+r)/2]
ttt[s_,t_,u_]:=(1-f[s])^t+2*((4+4*s)/(10+s))^(-u)*Z[s]^(t-1)

 NMinimize[{ttt[v,w,10],v>=2&&0<w<=10&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,50],v>=2&&0<w<=50&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,100],v>=2&&0<w<=100&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,200],v>=2&&0<w<=200&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,500],v>=2&&0<w<=500&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,750],v>=2&&0<w<=750&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,1000],v>=2&&0<w<=1000&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
 NMinimize[{ttt[v,w,2000],v>=2&&0<w<=2000&&w\[Element]Integers},{v,w},
 MaxIterations -> 6000]
```

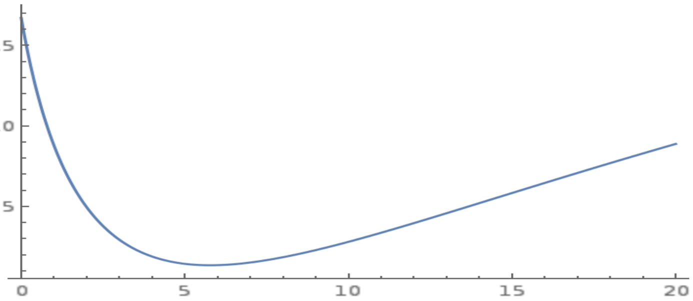
5.2 Codes for Hierarchical Poisson Model

```
s:={5,1,5,14,3,19,1,1,4,22}
t:={94.320, 15.720, 62.880, 125.760, 5.240, 31.440, 1.048, 1.048, 2.096, 10.480}
integrand[x_]:=((temp=0; For [i = 0, i < 10, i++,
temp=temp+(1.802+s[[i+1]])/(x+t[[i+1]])];
temp)-6.5)^2+(summ=0;For [i=0,i<10,i++,summ=summ+
(1.802+s[[i+1]])/((t[[i+1]]+x))^2];summ)
e[w_]:=NIntegrate[integrand[k]*(PDF[GammaDistribution
[0.01+10*1.802,1/(1+w)],k]),{k,0,Infinity}]
ek[w_]:=(1+e[w])/(1+(w-6.5)^2)
```

the command for computing ϵ when $r = 2.5$ is

```
NIntegrate[Min[PDF[GammaDistribution[0.01+10*1.802,1+(6.5-2.5)],k],
PDF[GammaDistribution[0.01+10*1.802,1+(6.5+2.5)],k]],{k,0,Infinity}]
```

For other values of ϵ, the user can just replace the r in the above command.
For computing α, A, λ and b, the plot for $e[w]$ and $ek[w]$ suggests that we only need to compare the values at $w = 6.5 - r$ and $w = 6.5 + r$. The plot for $ek[w]$ is given in the second example of section Applications, while the plot for $e[k]$ is presented here:



After acquiring the values for $\epsilon, \alpha, A, \lambda$ and b for different values of r, we get the expressions for the bounds.

optimal bound from Rosenthal(1993)

```
r=2, k=10.
NMinimize[{(1-0.249058)^j+1.12174^(10+j-1)*5.09248 ^(j-1)* 7.79584,
 0<j<=10&&j\[Element] Integers},j]
r=2.5, k=10.
NMinimize[{(1-0.144288)^j+1.54498^(-10+j-1)*5.55518^(j-1)*5.9819,
0<j<=10&&j\[Element] Integers},j]
r=3, k=10.
NMinimize[{(1-0.0747777)^j+1.81466^(-10+j-1)*6.05518^(j-1)*5.97184,
0<j<=10&&j\[Element] Integers},j]
r=3.5 k=10.
NMinimize[{(1-0.0337557)^j+1.93653^(-10+j-1)*6.84376^(j-1)*6.63933,
0<j<=10&&j\[Element] Integers},j]
r=4, k=10.
NMinimize[{(1-0.0127746)^j+2.02913^(-10+j-1)*8.54776 ^(j-1)* 8.08168,
0<j<=10&&j\[Element] Integers},j]
r=4.5, k=10.
NMinimize[{(1-0.00382688)^j+2.10017^(-10+j-1)*10.8882^(j-1)*  10.0628,
 0<j<=10&&j\[Element] Integers},j]
r=5, k=10.
NMinimize[{(1-0.000828392)^j+2.15535^(-10+j-1)*14.1012^(j-1)* 12.7823,
 0<j<=10&&j\[Element] Integers},j]
r=5.5, k=10.
NMinimize[{(1-0.000110751)^j+2.19880^(-10+j-1)*18.5608 ^(j-1)*  16.5572,
 0<j<=10&&j\[Element] Integers},j]
r=6, k=10.
NMinimize[{(1-0.00000 675804)^j+2.23347^(-10+j-1)*24.911 ^(j-1)*  21.9323,
  0<j<=10&&j\[Element] Integers},j]
r=6.5, k=10.
NMinimize[{(1-0.00000009337)^j+2.26149^(-10+j-1)*34.404^(j-1)*  29.9676,
 0<j<=10&&j\[Element] Integers},j]
```

For $k = 50, 100, 200, 500, 750, 1000, 2000$, the user can just replace the k in the above commands.

optimal new(weaker) bound from Roberts and Rosenthal(2004)

```
r=2, k=10.
NMinimize[{(1-0.249058)^j+1.12174^(-10)*Max[1,1/(1-0.249058)
*1.12174* 5.09248]^(j-1)* 7.79584, 0<j<=10&&j\[Element] Integers},j]
r=2.5, k=10.
NMinimize[{(1-0.144288)^j+1.54498^(-10)*Max[1,1/(1-0.144288
)*1.54498*5.55518]^(j-1)* 5.9819, 0<j<=10&&j\[Element] Integers},j]
r=3, k=10.
NMinimize[{(1-0.0747777)^j+1.81466^(-10)*Max[1,1/(1-0.0747777
)*1.81466* 6.05518]^(j-1)*5.97184, 0<j<=10&&j\[Element] Integers},j]
r=3.5, k=10.
NMinimize[{(1-0.0337557)^j+1.93653^(-10)*Max[1,1/(1-0.0337557
)*1.93653*6.84376]^(j-1)*6.63933,    0<j<=10&&j\[Element] Integers},j]
r=4, k=10.
NMinimize[{(1-0.0127746)^j+2.02913^(-10)* Max[1,1/(1-0.0127746
)*2.02913*8.54776]^(j-1)*8.08168,    0<j<=10&&j\[Element] Integers},j]
r=4.5, k=10.
NMinimize[{(1-0.00382688)^j+2.10017^(-10)*Max[1,1/(1-0.00382688
)*2.10017*10.8882]^(j-1)* 10.0628,    0<j<=10&&j\[Element] Integers},j]
r=5, k=10.
NMinimize[{(1-0.000828392)^j+2.15535^(-10)*Max[1,1/(1-0.000828392
)*2.15535*14.1012]^(j-1)* 12.7823,    0<j<=10&&j\[Element] Integers},j]
r=5.5, k=10.
NMinimize[{(1-0.000110751)^j+2.19880^(-10)*Max[1,1/(1-0.000110751
)*2.19880*18.5608]^(j-1)*  16.5572,    0<j<=10&&j\[Element] Integers},j]
```

```
r=6, k=10.
NMinimize[{(1-0.00000 675804)^j+2.23347^(-10)*Max[1,1/(1-0.00000 675804
)*2.23347*24.911]^(j-1)*  21.9323,   0<j<=10&&j\[Element] Integers},j]
r=6.5, k=10.
NMinimize[{(1-0.00000009337)^j+2.26149^(-10)*Max[1,1/(1-0.00000009337
)*2.26149*34.404]^(j-1)*  29.9676,   0<j<=10&&j\[Element] Integers},j]
```

For $k = 50, 100, 200, 500, 750, 1000, 2000$, the user can just replace the k in the above commands.

6 Bibliography

References

[1] N. Jain and B. Jamison (1967), Contributions to Doeblins theory of Markov processes. Z. Wahrsch. Verw. Geb. 8, 1940.

[2] J.S. Rosenthal (2000), *A first look at rigorous probability theory.* World Scientific Publishing, Singapore.

[3] S.P. Meyn and R.L. Tweedie (1993), *Markov chains and stochastic stability.* Springer-Verlag, London. Available at http://decision.csl.uiuc.edu/meyn/pages/TOC.html.

[4] Anthony Louis Almudevar, *Approximate Iterative Algorithms.* CRC Press

[5] Roberts, G.O. and J.S. Rosenthal (2004), *General state space Markov chains and MCMC algorithms.* Probability Surveys. V.1, pg 20 - 71

[6] L. Tierney (1994), *Markov chains for exploring posterior distributions* Ann. Stat. 22, 17011762.

[7] G.O. Roberts and J.S. Rosenthal (2003), *Harris Recurrence of Metropolis- Within-Gibbs and Transdimensional MCMC Algorithms.* Ann. Appl. Prob.16:2123-2139, 2006.

[8] G.L. Jones and J.P. Hobert (2001), Honest exploration of intractable prob- ability distributions via Markov chain Monte Carlo. Statistical Science 16, 312334.

[9] J.S. Rosenthal (1995), *Minorization conditions and convergence rates for Markov chain Monte Carlo.* J. Amer. Stat. Assoc. 90, 558566.

[10] P. G. Moschopoulos (1985) *The Distribution of the Sum of Independent Gamma Random Variables.* Ann. Inst. Statist. Math. 37(1985), Part A, 541-544