Exploration of Markov Chain Monte Carlo Algorithms

by

Boyi Li

Supervised by

Professor Jeffrey S. Rosenthal

A thesis submitted in conformity with the requirements
for the degree of Bachelor of Applied Science in Engineering Science
Division of Engineering Science
University of Toronto

Abstract

Exploration of Markov Chain Monte Carlo Algorithms

Boyi Li

Bachelor of Applied Science in Engineering Science

Division of Engineering Science

University of Toronto

2019

This review type of thesis summarizes the relevant core theoretical results with a slight overview the of application. The results of general space Markov chain is reviewed first, mainly Markov chain convergence theorem, theory of ergodicity and quantitative bounds. Next, the result of Adaptive MCMC is reviewed, with a focus on the theory of ergodicity. Finally, results of the boundedness of Adversarial Markov Chain are reviewed.

Acknowledgements

I sincerely thank my supervisor, Professor Jeffrey S. Rosenthal, for his patient guidance of my thesis. I can always learn interesting and useful knowledge from him. This thesis will not be completed without him. His excellent explanation for abstract proofs inspires my passion for Probability Theory. I really appreciate his effort to come up with thesis topic to me and give me intuition of abstract mathematics.

Contents

Chapter 1

General Space Markov Chain

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. The results are summarized from [3], [5], [6]. Future work includes summarizing central limit theorem on Markov chain, optimal scaling proofs, Harris recurrence, adaptive MCMC and relevant applications.

1.1 Foundations

Definition 1. General State Space

A state space $\mathcal{X}$ is general if it is equipped with a countably generated σ-algebra $\mathcal{B}(\mathcal{X})$

Definition 2. Transition Probability Kernels

If $P = \{P(x, A), x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$ is such that

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

Theorem 1.1. *For any initial measure μ on $\mathcal{B}(\mathcal{X})$ and any transition probility kernel $P = \{P(x, A), x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})\}$, there exist a stochastic process $\Phi = \{\Phi_0, \Phi_1, ...\}$ on $\Omega = \prod_{i=0}^{\infty} \mathcal{X}_i$, measurable with respect to $\mathcal{F} = \bigvee_{i=0}^{\infty} \mathcal{B}(\mathcal{X}_i)$, and a probability measure $\mathbf{P}_\mu$ on $\mathcal{F}$ such that $\mathbf{P}_\mu(B)$ is the probability of the event $\{\Phi \in B\}$, for $B \in \mathcal{F}$;and for measurable $A_i \subseteq \mathcal{X}_i$,$i = 0, ..., n$ and any n*

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

Typically, $\mathbf{P}_\mu(\Phi_n \in A|\Phi_0) = P^n(x, A)$

Theorem 1.2. *(ChapmanKolmogorov equations) For any m with $0 \leq m \leq n$,*

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

Definition 3. Let f be a bounded measurable function and μ be a σ-finite measure on $\mathcal{B}(\mathcal{X})$. Then define operator P^n such that

$$P^n f(x) = \int_{\mathcal{X}} P^n(x, dy)f(y)$$
$$\mu P^n(A) = \int_{\mathcal{X}} \mu(dx)P^n(x, A)$$

Definition 4. Stationary distribution

A probability measure $\pi(\cdot)$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a stationary measure for a Markov chain with transition probability P, if

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

Definition 5. First Hitting that and Return time

For any set $A \subseteq \mathcal{X}$, the variable τ_A is defined as $\min\{n \geq 0 : \Phi_n \in A\}$, namely, first hitting time.

For any set $A \subseteq \mathcal{X}$, the variable σ_A is defined as $\min\{n \geq 1 : \Phi_n \in A\}$, namely, first return time.

1.2 Convergence of Markov Chains

Definition 6. Total Variation Distance

The **total variation distance** between two probability measures $\nu_1(\cdot)$ and $\nu_2(\cdot)$ is:

$$\|\nu_1(\cdot) - \nu_2(\cdot)\| = \sup_A |\nu_1(A) - \nu_2(A)|$$

Proposition 1.3.

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

(b) $\|\nu_1(\cdot) - \nu_2(\cdot)\| = \frac{1}{b-a} \sup_{f:\mathcal{X}\to[0,1]} |\int f d\nu_1 - \int f d\nu_2|$ *for any* $a < b$, *and in particular*

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

(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.*

$\|P^n(x,\cdot) - \pi(\cdot)\| \leq \|P^{n-1}(x,\cdot) - \pi(\cdot)\|\ for\ n \in \mathbf{N}$

(d) More generally, letting $(\nu_i P)(A) = \int \nu_i(dx)P(x,A)$, *we always have*

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

(e) Let $t(n) = 2 \sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi(\cdot)\|$, *where* $\pi(\cdot)$ *is stationary. Then* t *is sub-multiplicative, i.e.*

$t(m+n)\ \leq t(m)t(n)\ for\ m,n \in \mathbf{N}$

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

(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* $\mathbf{P}[X = Y] = 1 - \|\mu(\cdot) - \nu(\cdot)\|$

Proof. (a): Let $\rho(\cdot)$ be any σ-*finite* measure such that $\nu_1 \ll \rho$ and $\nu_2 \ll \rho$. We could always find such measure such as $\rho = \nu_1 + \nu_2$. By Radon-Nikodym Theorem, ν_1 and ν_2 are absolute continuous with respect to ρ. Set $g = \frac{d\nu_1}{d\rho}$ and $h = \frac{d\nu_2}{d\rho}$.Then $|\int f d\nu_1 - f d\nu_2| = |\int f(g-h)d\rho| = |\int_{\{g>h\}} (g-h)d\rho + \int_{\{g<h\}} f(g-h)d\rho|$ which is maximized when $f = 1$ on $\{g > h\}$ and $f = 0$ on $\{g < h\}$(or vice-versa) .The above equation equals to $|\int_{\{g>h\}} (g-h)d\rho| = |\int_A d\nu_1 - \int_A d\nu_2| = |\nu_1(A) - \nu_2(A)|$ if let $A = \{g > h\}$ or $\{g < h\}$, corresponding to the value of $\|\nu_1(\cdot) - \nu_2(\cdot)\|$, thus prove the equivalence.

□

Proof. (b): Following part (a), $|\int f d\nu_1 - f d\nu_2| = |\int f(g-h)d\rho| = |\int_{\{g>h\}} (g-h)d\rho + \int_{\{g<h\}} f(g-h)d\rho|$ which is maximized when $f = b$ on $\{g > h\}$ and $f = a$ on $\{g < h\}$(or vice-versa) .

The equation above becomes

$|\int_{\{g>h\}} b(g-h)d\rho - \int_{\{g>h\}} a(g-h)d\rho + \int_{\{g>h\}} a(g-h)d\rho + \int_{\{g<h\}} a(g-h)d\rho|$

$= |\int_{\{g>h\}}(b(g-h) - a(g-h))d\rho + \int_{\mathcal{X}} a(g-h)d\rho| = |\int_{\{g>h\}}(b(g-h) - a(g-h))d\rho + a(\nu_1(\mathcal{X}) - \nu_2(\mathcal{X}))|$

$= |\int_{\{g>h\}}(b(g-h) - a(g-h))d\rho|$ since ν_1 and ν_2 are probability measures.

$= (b-a)|\nu_1(A) - \nu_2(A)|$ if we let $A = \{g > h\}$, thus prove the equivalence. □

Proof. (c) $|p^{n+1}(x,A) - \pi(A)| = |\int_{y\in\mathcal{X}} P^n(x,dy)p(y,A) - \int_{y\in\mathcal{X}} \pi(dy)p(y,A)|$

$= |\int_{y\in\mathcal{X}} P^n(x,dy)f(y) - \int_{y\in\mathcal{X}} \pi(dy)f(y)|$

Let $p(y,A) = f(y)$ and clearly $f \in [0,1]$

$\leq \sup_{f:\mathcal{X}\to[0,1]|\int_{y\in\mathcal{X}} P^n(x,dy)f - \int_{y\in\mathcal{X}} \pi(dy)f|} = \|p^n(x,\cdot) - \pi(\cdot)\|$

since it holds for every A

$\Rightarrow \|p^{n+1}(x,A) - \pi(A)\| \leq \|p^n(x,A) - \pi(A)\|$ for $\forall n$ □

Proof. (d) $|\int \nu_1(dx)p(x,A) - \int \nu_2(dx)p(x,A)| = |\int \nu_1(dx)f(x) - \int \nu_2(dx)f(x)|$

Let $p(y,A) = f(y)$ and clearly $f \in [0,1]$

$\leq \sup_{f:\mathcal{X}\to[0,1]}|\int \nu_1(dx)p(x,A) - \int \nu_2(dx)p(x,A)| = \|\nu_1(\cdot) - \nu_2(\cdot)\|$

Since it holds for every A

then $\|(\nu_1 P)(\cdot) - (\nu_2 P)(\cdot)\| \leq \|\nu_1(\cdot) - \nu_2(\cdot)\|$ □

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

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

Then let $f : \mathcal{X} \to [0,1]$, let $g(x) = (\hat{Q}f)(x) \equiv \int_{y\in\mathcal{X}} \hat{Q}(x,dy)f(y)$ and let $g^* = \sup_{x\in\mathcal{X}}|g(x)|$

$|g(x)| = |\int_{y\in\mathcal{X}}[P^m(x,dy) - \pi(dy)]f(y)| \leq \sup_{f:\mathcal{X}\to[0,1]}|\int_{y\in\mathcal{X}}[P^m(x,dy) - \pi(dy)]f(y)| = \|P^m(x,\cdot) - \pi(\cdot)\| \leq \sup_{x\in\mathcal{X}}\|P^m(x,\cdot) - \pi(\cdot)\| = \frac{1}{2}t(m)$ by part (a) $\Rightarrow g^* \leq \frac{1}{2}t(m)$ since it holds for $\forall x \in \mathcal{X}$. If $g^* = 0$, then clearly $\hat{P}\hat{Q}f = 0$. Otherwise, we compute that

$$2\sup_{x\in\mathcal{X}}|(\hat{P}\hat{Q}f)(x)| = 2g^* \sup_{x\in\mathcal{X}}|(\hat{P}[g/g^*)(x)| \leq t(m)\sup_{x\in\mathcal{X}}|(\hat{P}[g/g^*])(x)| \tag{1.1}$$

Since $-1 \leq g/g^* \leq 1$, we have $(\hat{P}[g/g^*])(x) \leq 2\|P^n(x,\cdot) - \pi(\cdot)\|$ by part(b).

So that $\sup_{x\in\mathcal{X}}(\hat{P}[g/g^*])(x) \le t(n)$.

$$\begin{aligned} t(n+m) &= 2\sup_{x\in\mathcal{X}}\|P^{n+m}(x,\cdot)-\pi(\cdot)\| \\ &= 2\sup_{x\in\mathcal{X}}\sup_{f:\mathcal{X}\to[0,1]}|(\hat{P}\hat{Q}f(x)| \\ &= 2\sup_{f:\mathcal{X}\to[0,1]}\sup_{x\in\mathcal{X}}|(\hat{P}\hat{Q}f(x)| \\ &\le t(m)t(n)\ by\ (1.1) \end{aligned}$$

□

Proof. (f): we first show the first equality:

$\|\mu(\cdot)-\nu(\cdot)\| = \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}|\int fd\mu(\cdot)-fd\nu(\cdot)| = \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}|\int fd\mu(\cdot)-fd\nu(\cdot)|$

$= \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}|\int f(g-h)d\rho| = \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}|\int_{\{g>h\}} f(g-h)d\rho + \int_{\{g<h\}} f(g-h)d\rho|$

$= \frac{1}{2}|\int_{\{g>h\}}(g-h)d\rho + \int_{\{g<h\}}(h-g)d\rho| = \frac{1}{2}(\int_{\{g>h\}}(g-h)d\rho + \int_{\{g<h\}}(h-g)d\rho) =\frac{1}{2}\int_{\mathcal{X}}(M-m)d\rho$

Since $M+m = g+h$, then $\int_{\mathcal{X}}(M+m)d\rho = 2$

$\frac{1}{2}\int_{\mathcal{X}}(M-m)d\rho = 1-\frac{1}{2}(2-\int_{\mathcal{X}}(M-m)d\rho = 1-\frac{1}{2}(\int_{\mathcal{X}}(M+m)d\rho - \int_{\mathcal{X}}(M-m)d\rho = 1-\int_{\mathcal{X}} md\rho$ □

Proof. (g): let $a = \int_{\mathcal{X}} md\rho$, $b = \int_{\mathcal{X}}(g-m)d\rho$ and $c = \int_{\mathcal{X}}(h-m)d\rho$. We only consider the case when they are positive, since it is trivial if they are zero. We then define random variables Z, U, V, I, jointly, such that Z has density m/a, U has density $(g-m)/b$, V has density $(h-m)/b$ and I is independent of Z, U, V with $\mathbf{P}[I=1]=a$ and $\mathbf{P}[I=0]=1-a$. Let $X=Y=Z$ if $I=1$, and $X=U$ and $Y=V$ if $I=0$. Then

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

$$\begin{aligned}\mathbf{P}(Y \in A) =&\mathbf{P}(Z \in A|I = 1)\mathbf{P}(I = 1) + \mathbf{P}(V \in A|I = 0)\mathbf{P}(I = 0)\\ =&a\int_A (m/a)d\rho + (1-a)\int_A ((h-m)/b)d\rho\\ =&\int_A md\rho + (1-a)\frac{\int_A (h-m)d\rho}{\int_{\mathcal{X}}(h-m)d\rho}\\ =&\int_A md\rho + \frac{1-\int_{\mathcal{X}} md\rho}{\int_{\mathcal{X}} hd\rho - \int_{\mathcal{X}} md\rho}\int_A (h-m)d\rho\\ =&\int_A md\rho + \int_A (h-m)d\rho\\ =&\int_A hd\rho = \int_A d\nu = \nu(A)\end{aligned}$$

we thus show that, $X \sim \mu(\cdot)$ and $Y \sim \nu(\cdot)$.Also, U supports the region $\{g > h\}$ and V supports the region $\{h > g\}$, so $\mathbf{P}[U = V] = 0$. Then $\mathbf{P}[X = Y] = \mathbf{P}[I = 1] = a = 1 - \|\mu(\cdot) - \nu(\cdot)\|$ by part(f). □

The concepts to total variance helps us to answer question like is $\lim_{n\to\infty}\|P^n(x,\cdot) - \pi(\cdot)\| = 0$? And, given $\epsilon > 0$, how large must n be so that $\|P^n(x,\cdot) - \pi(\cdot)\| < \epsilon$.
To admit convergence, we not only require that the chain have a stationary distribution, but also require the chain to be **irreducible** and **aperiodic**. We then introduce the concept of the weaker condition of $\phi-$irreducibly and aperiodic.

Definition 7. A chain is *ϕ-irreducible* if there exists a non-zero σ-finite measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$.

Definition 8. A Markov chain with stationary distribution $\pi(\cdot)$ is *aperiodic* if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, ..., \mathcal{X}_d \subseteq \mathcal{X}$ with $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i (i \leq i \leq d-1)$ and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$, such that $\pi(\mathcal{X}_1) > 0$(and hence $\pi(\mathcal{X}_i) > 0$ for all i).(Otherwise, the chain is *periodic*, with *period d*, and *periodic decomposition $\mathcal{X}_1, ..., \mathcal{X}_d$*

To better understand these two concepts, we provide an example here.

Example 1. Suppose $\pi(\cdot)$ is a probability measure with unnormalised density function π_μ with respect to d-dimensional Lebesgue measure. Consider the Metropolis-Hastings algorithm for π_μ with proposal density $q(\mathbf{x}, \cdot)$ with respect to d-dimensional Lebesgue measure. Suppose $q(\cdot, \cdot)$ is positive and continuous on $\mathbf{R}^d \times \mathbf{R}^d$, and π_μ is finite everywhere, we then show the algorithm is π-irreducible and aperiodic.

Proof. Let $\pi(A) > 0$. Then ∃ R¿0 such that $\pi(A_R) > 0$, where $A_R = A \cap B_R(\mathbf{0})$. $B_R(\mathbf{0})$ represents the ball of radius R centered at $\mathbf{0}$. Then by continuity, for any $\mathbf{x} \in \mathbf{R}^d$, $\inf_{\mathbf{y}\in A_R} \min q(\mathbf{x}, \mathbf{y}) \geq \epsilon$ for some

$\epsilon > 0$, so we have

$$P(\mathbf{x}, A) \geq P(\mathbf{x}, A_R) \geq \int_{A_R} q(\mathbf{x}, \mathbf{y}) \min\left[1, \frac{\pi_\mu(\mathbf{y})q(\mathbf{y}, \mathbf{x}}{\pi_\mu(\mathbf{x})q(\mathbf{y}, \mathbf{x}}\right] d\mathbf{y}$$

$$\geq \epsilon Leb(\{\mathbf{y} \in A_R : \pi_\mu(\mathbf{y}) \geq \pi_\mu(\mathbf{x})\}) + \frac{\epsilon}{\pi_\mu(\mathbf{x})}\pi(\{\mathbf{y} \in A_R : \pi_\mu(\mathbf{y}) \geq \pi_\mu(\mathbf{x})\})$$

Since $\pi(\cdot)$ is absolutely continuous with respect to Lebesgue measure, and since $Leb(A_R) > 0$, it follows that the terms in this final sum cannot both be 0, so that we must have $P(x, A) > 0$. Hence, the chain is π-irreducible.

For aperiodicity. Suppose that $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint subsets of $\mathcal{X}$ both of positive π measure, with $P(\mathbf{x}, \mathcal{X}_2) = 1$ for all $\mathbf{x} \in \mathcal{X}_1$. But just take any $\mathbf{x} \in \mathcal{X}_1$, then since $\mathcal{X}_1$ must have positive Lebesgue measure,

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

which is a contradiction. Therefore aperiodicity must hold. □

Then we state the main asymptotic convergence theorem, whose proof is shown in section later.

Theorem 1.4. *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. $x \in \mathcal{X}$*

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

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

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

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

Proof. Let the chain have periodic decomposition $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_d \subseteq \mathcal{X}$, and let P' be the d-step chain P^d restricted to the state space $\mathcal{X}_1$. Then P' is ϕ- irreducible and aperiodic on $\mathcal{X}_1$, with stationary distribution $\pi'(\cdot)$. We then show that for state space $\mathcal{X}_j$, the stationary distribution is $\pi' P^{j-1}(\cdot)$, for

$1 \leq j \leq d$. Let $A \subseteq \mathcal{X}_j$

$$\begin{aligned}
\int_{y\in\mathcal{X}_j} \pi' P^{j-1}(dy)P^d(y,A) &= \int_{z\in\mathcal{X}_1}\int_{y\in\mathcal{X}_j} \pi' P^{j-1}(dy)P^{d-j+1}(y,dz)P^{j-1}(z,A) \\
&= \int_{z\in\mathcal{X}_1}\int_{x\in\mathcal{X}_1}\int_{y\in\mathcal{X}_j} \pi'(dx)P^{j-1}(x,dy)P^{d-j+1}(y,dz)P^{j-1}(z,A) \\
&= \int_{z\in\mathcal{X}_1}\int_{x\in\chi_1} \pi'(dx)P^d(x,dz)P^{j-1}(z,A) \\
&= \int_{z\in\mathcal{X}_1} \pi'(dz)P^{j-1}(z,A) \\
&= \pi' P^{j-1}(A)
\end{aligned}$$

Thus we know $\pi(\cdot) = (1/d)\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)$, we then prove the Corollary when $n = md$ with $m \to \infty$. We assume without loss of generality that $x \in \mathcal{X}_1$. By Proposition (d) we have $\|P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)\| \leq \|P^m d(x,\cdot) - \pi'(\cdot)\| \; \forall j \in \mathbf{N}$

$$\begin{aligned}
\|(1/d)\sum_{i=md}^{md+d-1} P^i(x,\cdot) - \pi(\cdot)\| =& \|(1/d)\sum_{j=0}^{d-1} P^{md+j}(x,\cdot) - (1/d)\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)\| \\
\leq& (1/d)\sum_{j=0}^{d-1}\|P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot)\| \\
\leq& (1/d)\sum_{j=0}^{d-1}\|P^{md}(x,\cdot) - \pi'(\cdot)\| \\
=& \ (1/d)\sum_{j=0}^{d-1}\|P'(x,\cdot) - \pi'(\cdot)\|
\end{aligned} \tag{1.2}$$

By applying theorem to P', we have that $\lim_{m\to\infty}\|P^{md}(x,\cdot) - \pi'(\cdot)\| = 0$ for π'-a.e. $x \in \mathcal{X}_1$ Similarly, the result holds for $(\pi' P^j)(\cdot)$-a.e. $x \in \mathcal{X}_j$, for $1 < j \leq d$, the Corollary is then proved. □

1.3 Ergodicity of Markov Chain

In this section, we discuss the rate of convergence. Uniform ergodicity is one qualitative convergence rate property.

Definition 9. A Markov chain having stationary distribution $\pi(\cdot)$ is *uniformly ergodic* if

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

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

Proposition 1.5. A Markov chain with stationary distribution $\pi(\cdot)$ is uniformly ergodic if and only if $\sup_{x\in\mathcal{X}}\|P^n(x,\cdot) - \pi(\cdot)\| \leq \frac{1}{2}$ for some $n \in \mathbf{N}$.

Proof. If the chain is uniformly ergodic, then $\lim_{n\to\infty}\sup_{x\in\mathcal{X}}\|P^n(x,\cdot)-\pi(\cdot)\| \leq \lim_{n\to\infty} M\rho^n = 0$. Thus, for n to be sufficiently large, $\sup_{x\in\chi}\|P^n(x,\cdot)-\pi(\cdot)\| \leq \frac{1}{2}$. Conversely, if $\sup_{x\in\chi}\|P^n(x,\cdot)-\pi(\cdot)\| \leq \frac{1}{2}$ for some $n \in \mathbf{N}$, using notation in proposition, we have that $t(n) \equiv \beta < 1$, so that for $j \in \mathbf{N}$, $t(jn) \leq (t(n))^j = \beta^j$. Hence, from proposition,

$$\begin{aligned}\|P^m(x,\cdot)-\pi(\cdot)\| \leq & \|P^{\lfloor m/n\rfloor n}(x,\cdot)-\pi(\cdot)\| \leq \frac{1}{2}t(\lfloor m/n\rfloor n)\\ \leq & \beta^{\lfloor m/n\rfloor} \leq \beta^{-1}(\beta^{1/n})^m\end{aligned} \tag{1.3}$$

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

Remark. The Proposition continuous to hold if $\frac{1}{2}$ is replaced by θ for any $0 < \theta < \frac{1}{2}$.But not for $\theta \geq \frac{1}{2}$

To further develop the concept of uniform ergodicity, we present the concept of small sets first.

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

$$P^{n_0}(x,\cdot) \geq \epsilon\nu(\cdot) \quad x \in C, \tag{1.4}$$

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

Remark. Intuitively, the condition here means that all of the n_0-step transitions from within C, all have an "ϵ-overlap", i.e. a component of size ϵ. Small sets are widely used in *Couplings* we illustrated later. There is a notion weaker than small set, called *pseudo-small* set.

Theorem 1.6. *Consider a Markov chain with invariant probability distribution $\pi(\cdot)$. Suppose the minorisation condition is satisfied for some $n_0 \in \mathbf{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, in the special case $C = \mathcal{X}$(i.e., the entire state space is small). Then the chain is uniformly ergodic, and in fact $\|P^n(x,\cdot)-\pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0\rfloor}$ for all $x \in \mathcal{X}$, where $\lfloor r\rfloor$ is the greatest integer not exceeding r.*

Remark. The theorem helps us to find a quantitative bound on the distance to stationarity $\|P^n(x,\cdot)-\pi(\cdot)\|$, i.e. it must be $\leq (1-\epsilon)^{\lfloor n/n_0\rfloor}$. Once ϵ and n_0 are known, we can find n_*, such that $\|P^{n_*}(x,\cdot)-\pi(\cdot)\| \leq 0.001$. We can say that after n_* iterations, the Markov chain converges.

We illustrate an example and a counter-example to better understand this concept.

Example 2. Consider in dimension $d = 1$, and suppose that $\pi_\mu(x) = \mathbf{1}_{0<|x|<1}|x|^{-1/2}$, and let $q(x,y) \propto \exp-(x-y)^2/2$ we show that the any neighbourhood of 0 is not small.

Proof. let S be the set that contain 0. Then $P(x,dy) = q(x,y)dy\min\{1, \frac{\pi_\mu(y)}{\pi_\mu(x)}\}$. Let $x \in S$ and we have $x \to 0$, $P(x,dy) \to 0$. Thus the minorisation condition does not hold, so S is not small. □

Example 3. Suppose $\pi(\cdot)$ is a probability measure with unnormalised density function π_μ with respect to d-dimensional Lebesgue measure. Consider the Metropolis-Hastings algorithm for π_μ with proposal density $q(\mathbf{x}, \cdot)$ with respect to d-dimensional Lebesgue measure. Suppose $q(\cdot, \cdot)$ is positive and continuous on $\mathbf{R}^d \times \mathbf{R}^d$, and π_μ is finite everywhere, we then show that all compact sets on which π_μ is bounded are small.

Proof. Let C be a compact set on which π_μ is bounded by $k < \infty$. Let $\mathbf{x} \in C$, and D be any compact set of positive Lebesgue and π measure, such that $\inf_{\mathbf{x}\in C, \mathbf{y}\in D} q(\mathbf{x}, \mathbf{y}) = \epsilon > 0$ for all $\mathbf{y} \in D$. We then have

$$P(\mathbf{x}, d\mathbf{y}) \geq q(\mathbf{x}, \mathbf{y})\mathbf{dy} \min\{1, \frac{\pi_\mu(\mathbf{y})}{\pi_\mu(\mathbf{x})}\} \geq \epsilon d\mathbf{y} \min\{1, \frac{\pi_\mu(\mathbf{y})}{k}\}$$

which is a positive measure independent of $\mathbf{x}$. Hence, C is small. □

A weaker condition than uniform ergodicity is geometric ergodicity, which is defined as follows:

Definition 11. A Markov chain with stationary distribution $\pi(\cdot)$ is *geometrically ergodic* if

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

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

The difference bewteen geometric ergodicity and uniform ergodicity is that not the constant M may depend on the initial state x.
If the state space $\mathcal{X}$ is *finite*, then all irreducible and aperiodic Markov Chains are geometrically ergodic. However, for infinite $\mathcal{X}$ this is not the case. We then illustrate the conditions which ensure geometric ergodicity.

Definition 12. Given Markov chain transition probabilities P on a state space $\mathcal{X}$, and a measurable function $f : \mathcal{X} \to \mathbf{R}$, define the function $Pf : \mathcal{X} \to \mathbf{R}$ such that $(Pf)(x)$ is the conditional expected value of $f(X_{n+1})$, given that $X_n = x$. In symbols, $(Pf)(x) = \int_{y\in\mathcal{X}} f(y)P(x, dy)$.

Definition 13. A Markov chain satisfies a *drift condition*(or, univariate geometric drift condition) if there are constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty)$, such that

$$PV \leq \lambda V + b\mathbf{1}_C(x), \tag{1.5}$$

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

The main result guaranteeing geometric ergodicity is the following

Theorem 1.7. *Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(\cdot)$. Suppose the minorisation condition is satisfied for some $C \subseteq \mathcal{X}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose further that the drift condition is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, \infty]$ with $V(x) < \infty$ for at least one (and hence for π-a.e.) $x \in \mathcal{X}$. Then the chain is geometrically ergodic.*

The theorem is proved in the latter section and we then illustrate an example.

Example 4. Consider a simple example of geometric ergodicity of Metropolis algorithms on $\mathbf{R}$. Suppose that $\mathcal{X} = \mathbf{R}^+$ and $\pi_\mu(x) = e^{-x}$. We will use a symmetric (about x) proposal distribution $q(x, y) = q(|y - x|)$ with support contained in $[x - a, x + a]$. We then show that the algorithm is geometric ergodic.

Proof. Take drift function to be $V(x) = e^{cx}$ for some $c > 0$. For $x \geq a$, compute:

$$PV(x) = \int_{x-a}^{x} V(y)q(x,y)dy + \int_{x}^{x+a} V(y)q(x,y)dy\frac{\pi_u(y)}{\pi_u(x)}$$
$$+V(x)\int_{x}^{x+a} q(x,y)dy(1 - \pi_u(y)/\pi_u(x))$$

By the symmetry of q, this can be written as

$$\int_{x}^{x+a} I(x,y)q(x,y)dy,$$

where

$$I(x,y) = \frac{V(y)}{\pi_u(y)} + V(2x - y) + V(x)\left(1 - \frac{\pi_u(y)}{\pi_u(x)}\right) = e^{cx}\left[2 - (1 + e^{(c-1)u})(1 - e^{-cu})\right]$$

where $u = y - x$. For $c < 1$, this is $2(1 - \epsilon)V(x)$ for some positive constant ϵ.Thus in this case, we have shown that for all $x > a$

$$PV(x) \leq \int_{x}^{x+a} 2V(x)(1 - \epsilon)q(x,y)dy = (1 - \epsilon)V(x) \tag{1.6}$$

Similarly, we could show that $PV(x)$ is bounded on $[0, a]$, and that $[0, a]$ is a small set. Thus the drift condition holds and hence the algorithm is geometrically ergodic. □

1.4 Quantitative Convergence Rates

In this section, the result of quantitative bounds on convergence rates is presented.

Definition 14. The *bivariate drift condition* is satisfied if

$$\bar{P}h(x,y) \leq h(x,y)/\alpha, \quad (x,y) \notin C \times C \tag{1.7}$$

for some function $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ and some $\alpha > 1$, where

$$\bar{P}h(x, y) = \int_{\mathcal{X}} \int_{\mathcal{X}} h(z, w) P(x, dz) P(y, dw)$$

Proposition 1.8. Suppose the univariate drift condition (13) is satisfied for some $V : \mathcal{X} \to [1, \infty)$, $C \subseteq \mathcal{X}$, $\lambda < 1$, and $b < \infty$. Let $d = \inf_{x \in C^c} V(x)$. Then if $d > [b/(1 - \lambda)] - 1$, then the bivariate drift condition (1.7) is satisfied for the same C, with $h(x, y) = \frac{1}{2}[V(x) + V(y)]$ and $\alpha^{-1} = \lambda + b/(d + 1) < 1$.

Proof. If $(x, y) \notin C \times C$, either $x \notin C$ or $y \notin C$(or both), so $h(x, y) \geq (1 + d)/2$. Since univariate drift condition is satisfied, $PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b$. Then

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

Since $d > [b/(1 - \lambda)] - 1$, then $\lambda + b/(1 + d) < 1$. □

we let

$$B_{n_0} = \max \left[1, \alpha^{n_0}(1 - \epsilon) \sup_{C \times C} \bar{R}h\right] \tag{1.8}$$

where for $(x, y) \in C \times C$,
$\bar{R}h(x, y) = \int_{\mathcal{X}} \int_{\mathcal{X}} (1 - \epsilon)^{-2} h(z, w)(P^{n_0}(x, dz) - \epsilon\nu(dz))(P^{n_0}(y, dw) - \epsilon\nu(dw))$.

Theorem 1.9. *Consider a Markov chain on a state space $\mathcal{X}$, having transition kernel P. Suppose there is $C \subseteq \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, a probability distribution $\nu(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, $n_0 \in \mathbf{N}$, and $\epsilon > 0$, such that (**??**) and (1.7) hold. Define B_{n_0} by 1.8. Then for any joint distribution $\mathcal{L}(X_0, X'_0)$, and any integers $1 \leq j \leq k$, if $\{X_n\}$ and $\{X'_n\}$ are two copies of the Markov chain started in the joint initial distribution $\mathcal{L}(X_0, X'_0)$, then*

$$\|\mathcal{L}(X_k) - \mathcal{L}(X'_k)\|_{TV} \leq (1 - \epsilon)^j + \alpha^{-k}(B_{n_0})^{j-1}\mathbf{E}[h(X_0, X'_0)]. \tag{1.9}$$

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

The theorem is then proved in Section below.

1.5 Convergence Proofs using Coupling Constructions

In this section, we prove the theorems stated earlier. We focus on the method of *coupling* for the proof, which are well-suited to analyzing MCMC algorithms on general state spaces.

1.5.1 The Coupling Inequality

Suppose we have two random variables X and Y, defined jointly on some space $\mathcal{X}$. If we write $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ for their repective probability distirbutions, then we can write

$$\begin{aligned}
\|\mathcal{L}(X) - \mathcal{L}(Y)\| =& \sup_A |P(X \in A) - P(Y \in A)| \\
=& \sup_A |P(X \in A, X = Y) + P(Y \in A, X \neq Y) \\
& -P(Y \in A, Y = X) - P(Y \in A, Y \neq X)| \\
=& \sup_A |P(X \in A, X \neq Y) - P(Y \in A, Y \neq X)| \qquad (1.10) \\
=& \sup_A (P(X \in A, X \neq Y) - P(Y \in A, Y \neq X)) \\
& \textit{or} \quad \sup_A (P(Y \in A, X \neq Y) - P(X \in A, Y \neq X)) \\
\leq& P(X \neq Y)
\end{aligned}$$

The last inequality holds because both $\sup_A(P(X \in A, X \neq Y))$ and $\sup_A(P(Y \in A, Y \neq X))$ are non-negative and are smaller than $P(X \neq Y)$.

The coupling equality shows that the variation distance between the laws of two random variables is bounded by the probability that they are equal.

1.5.2 Small Sets and Coupling

Suppose C denotes the small set. The idea of coupling is to run two copies $\{X_n\}$ and $\{X_n^{'}\}$ of the Markov chain, each of which marginally follows the updating rules $P(x, \cdot)$, but whose joint construction(using C) gives them as high a probability as possible of becoming equal to each other.

THE COUPLING CONSTRUCTION:

Start with $X_0 = x$ and $X_0^{'} \sim \pi(\cdot)$, and $n = 0$, and repeat the following loop forever.

Beginning of Loop. Given X_n and $X_0^{'}$:

1. If $X_n = X_0^{'}$, choose $X_n = X_0^{'} \sim P(X_n, \cdot)$, and replace n by $n + 1$.

2. Else, if $(X_n, X_0^{'}) \in C \times C$, then:

(a) w.p. ϵ, choose $X_{n+n_0} = X^{'}_{n+n_0} \sim \nu(\cdot)$;

(b) else, w.p. $1-\epsilon$, conditionally independently choose

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

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

In the case $n_0 > 1$, for completeness go back and construct $X_{n+1}, ..., X_{n+n_0-1}$ from their correct conditional distributions given X_n and X_{n+n_0}, and conditionally and independently construct $X'_{n+1}, ..., X'_{n+n_0-1}$ from their correct conditional distributions given $X'_n, ..., X'_{n+n_0}$. In any case, replace n by $n+n_0$.

3. Else, conditionally independently choose $X_{n+1} \sim P(X_n,\cdot)$ and $X'_n \sim P(X'_{n+1},\cdot)$, and replace n and $n+1$.

Then return to Beginning of Loop

We then check that $\mathbf{P}[X_n \in A] = P^n(x, A)$ and $\mathbf{P}[X'_n \in A] = \pi(A)$ for all n.

Proof. It trivial that the equality holds for condition 1 and 3, since these two variables are independently updated based on transition kernel P.

For conditional 2, when $(X_n, X^{'}_0) \in C \times C$

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

$$X'_{n+n_0} \sim \epsilon\nu(\cdot) + \frac{1-\epsilon}{1-\epsilon}[P^{n_0}(X'_n,\cdot) - \epsilon\nu(\cdot)] = P^{n_0}(X'_n,\cdot)$$

It then follows that $\mathbf{P}[X_{n+n_0} \in A] = P^{n+n_0}(x, A)$ and $\mathbf{P}[X'_{n+n_0} \in A] = \pi(A)$ for all n.

For $1 \leq a \leq n_0 - 1$ and given $X_{n+a-1} = b$,$X'_{n+a-1} = b'$,we update X_{n+a} by

$$\mathbf{P}[X_{n+a} \in A | X_{n+a-1} = b, X_{n+n_0} = c] = \int_A P(b, dx) P^{n_0-a}(x, c)$$

, continuing this pattern and same thing holds for X'_{n+a}. Since they are updated by transition kernel P, the result is proved.

The **Coupling equality** then shows that

$$\|P^n(x,\cdot) - \pi(\cdot)\| = \sup_A(\mathbf{P}[X_n \in A]) - \mathbf{P}[X'_n \in A]) \leq \mathbf{P}[X_n \neq X'_n]$$

We then use this inequality to prove the theorem we state in the previous section.

1.5.3 Proof of Theorem 1.6

In this case, $C = \mathcal{X}$, so every n_0 steps we have probability at least ϵ of making X_n and X'_n equal. Then if $n = n_0 m$, then $\mathbf{P}[X_n \neq X'_n] \leq (1-\epsilon)^m$. Hence from coupling inequality, $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^m = (1-\epsilon)^{n/n_0}$ in this case. It then follows from Proposition that $\|P^n(x,\cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for any n.

□

1.5.4 Proof of Theorem 1.4

Theorem 1.10. *Every ϕ-irreducible Markov chain, on a state space with countably generated σ-algebra, contains a small set $C \in \mathcal{X}$ with $\phi(C) > 0$.(In fact, each $B \subseteq \mathcal{X}$ with $\phi(B) > 0$ in turn contains a small set $C \subseteq B$ with $\phi(C) > 0$) Furthermore, the minorisation measure $\nu(\cdot)$ may be taken to satisfy $\nu(C) > 0$.*

The idea behind this proof is that, if one can show that the pair (X_n, X'_n) will hit $C \times C$ infinitely often, then they will have infinitely many opportunity to couple, with probability $\geq \epsilon > 0$ of coupling each time. Hence, they will eventually couple with probability 1, thus proving Theorem .

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

Proof. Suppose to the contrary that the conclusion does not hold,

$$\pi\{x \in \mathcal{X} : \mathbf{P}_x(\tau_A = \infty) > 0\} > 0 \tag{1.11}$$

Then the following claims are made:

Claim 1. Condition (1.11) implies that there are constant $l, l_0 \in \mathbf{N}$, $\delta > 0$, and $B \subseteq \mathcal{X}$ with $\pi(B) > 0$, such that

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

Claim 2. Let B, l, l_0, and δ be as in Claim 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 $1 \leq r \leq j$,

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

Assuming the claim. We have by stationarity that for any $j \in \mathbf{N}$,

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

For $j > 1/\pi(B)\delta$, this gives $\pi(A^C) > 1$, which is impossible. This gives a contradiction, and hence completes the proof Lemma20, subject to the proofs of Claim 1 and 2 below.

Proof of Claim 1. By (1.11), we can find δ_1 and a subset $B_1 \subseteq \mathcal{X}$ with $\pi(B_1) > 0$, such that $\mathbf{P}_x(\tau_A < \infty) \leq 1 - \delta_1$ for all $x \in B_1$. On the other hand, since $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$, 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 $x \in B_2$.

Set $\eta = \#\{k \geq 1; X_{kl_0} \in B_2\}$.Then for any $r \in \mathbf{N}$ and $x \in \mathcal{X}$, we have $\mathcal{P}_x(\tau_A = \infty, \eta = r) \leq (1 - \delta_2)^r$. In particular, $\mathcal{P}_x(\tau_A = \infty, \eta = r) = 0$. Hence for $x \in B_2$, we have

$$\begin{aligned}\mathbf{P}_x(\tau_A =&\infty, \eta < \infty) = 1 - \mathbf{P}_x(\tau_A = \infty, \eta = \infty) - \mathbf{P}_x(\tau_A < \infty) \\ &\geq 1 - 0 + (1 - \delta_1) = \delta_1\end{aligned}$$

Hence, there is $l \in \mathbf{N}$, $\delta > 0$, and $B \subseteq B_2$ with $\pi(B) > 0$, such that

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

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

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

Proof of Claim 2. Compute using staionarity and then Claim 1 that

$$\begin{aligned}&\int_{x\in\mathcal{X}} \pi(dx)\mathbf{P}_x[S = r, X_{jL} \notin A] \\ &= \int_{x\in\mathcal{X}} \pi(dx) \int_{y\in B} P^{rL}(x, dy)\mathbf{P}_x[S = -\infty, X_{jL} \notin A] \\ &= \int_{y\in B} \int_{x\in\mathcal{X}} \pi(dx) P^{rL}(x, dy)\mathbf{P}_y[S = -\infty, X_{(j-r)L} \notin A] \\ &= \int_{y\in B} \pi(dy)\mathbf{P}_y[S = -\infty, X_{(j-r)L} \notin A] \\ &\geq \int_{y\in B} \pi(dy)\delta = \pi(B)\delta\end{aligned}$$

□

Let C be a small set as in Theorem 1.10. Consider the coupling construction $\{(X_n, Y_n\}$. Let $G \subseteq \mathcal{X} \times \mathcal{X}$ be the set of (x, y) such that $\mathbf{P}_{(x,y)}(\exists n \geq 1; X_n = Y_n) = 1$. From the coupling constuction, if $(X_0, X_0') \equiv (x, X_0') \in G$, then $\lim_{n\to\infty} \mathbf{P}[X_n = X_n'] = 1$, so that $\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\| = 0$, proving the theorem. Hence, it suffices to show that for π-a.e. $x \in \mathcal{X}$, we have $\mathbf{P}([(x, X_0') \in G]) = 1$.

Let G be as above, let $G_x = \{y \in \mathcal{X}; (x, y) \in G\}$ for $x \in \mathcal{X}$, and let $\bar{G} = \{x \in \mathcal{X}; \pi(G_x) = 1\}$. Then Theorem follows from $\pi(\bar{G}) = 1$. To achieve this, we first prove a lemma, below.

Lemma 2. *Consider an aperiodic Markov chain on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Let $\nu(\cdot)$ be any probability measure on $\mathcal{X}$. Assume that $\nu \ll \pi(\cdot)$, and that for all $x \in \mathcal{X}$, there is $n = n(x) \in \mathbf{N}$ and $\delta = \delta(x) > 0$ (for example, this always holds if ν is a minorisation measure for a small or petite set which is reachable from all states). 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 $n_* \in \mathbf{N}$ with $T \supseteq \{n_*, n_* + 1, n_* + 2, ...\}$.*

Proof. Since $P^{(n(x))} \geq \delta(x)\nu(\cdot)$ for all $x \in \chi$, it follows that T is non-empty.

If $n, m \in T$, then we have

$$\begin{aligned} \int_{x \in \mathcal{X}} \nu(dx) P^{n+m}(x, \cdot) &= \int_{x \in \mathcal{X}} \int_{y \in \mathcal{X}} \nu(dx) P^n(x, dy) P^m(y, \cdot) \\ &\geq \int_{y \in \mathcal{X}} \delta_n \nu(dy) P^m(y, \cdot) \geq \delta_n \delta_m \nu(\cdot) \end{aligned} \tag{1.12}$$

thus, T is additive, i.e. if $n, m \in T$, then $n + m \in T$. We then prove $gcd(T) = 1$. By leema, if T is non-empty and additive, and $gcd(T) = 1$, then there is $n_* \in \mathbf{N}$ such that $T \supseteq n_*, n_* + 1, n_* + 2, ...$, as claimed.

Suppose that $gcd(T) = d > 1$. A contradiction will be derived.

For $1 \leq d \leq j$, let

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

Then $\bigcup_i^d \mathcal{X}_i = \mathcal{X}$ by assumption above. Now, let

$$S = \bigcup_{i \neq j} (\mathcal{X}_i \cup \mathcal{X}_j) \tag{1.13}$$

let

$$\bar{S} = S \cup \{x \in \mathcal{X}; \exists m \in \mathbf{N} \textit{ s.t. } P^m(x, S) > 0\} \tag{1.14}$$

and let

$$\mathcal{X}_i' = \mathcal{X}_i \setminus \bar{S} \tag{1.15}$$

Then $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_d$ are disjoint by construction since S is removed. Also if $x \in \mathcal{X}_i'$, then $P(x, \bar{S}) = 0$, so that $P(x, \mathcal{X} \setminus \bar{S}) = P(x, \bigcup_{j=1}^d \mathcal{X}_j') = 1$. If $P(x, \mathcal{X}_{i+1}') > 0$ and $P(x, \mathcal{X}_{i+k}') => 0$, then x have the positive probability to be in state $\mathcal{X}'_{i+1}$ or $\mathcal{X}'_{i+k}$ after one step. Then we have

$$\begin{aligned}\int_{z\in A}\int_{y\in\mathcal{X}_{i+1}} P(x,dy)P^{l^*(y)(d-i-1)}(y,dz) &= P^{l^*(y)d-i}(A)\\ \geq \int_{z\in A}\int_{y\in\mathcal{X}_{i+1}} P(x,dy)\delta(y)\nu(dz) &\geq \delta^*(y)\nu(A)\end{aligned} \tag{1.16}$$

$$\begin{aligned}\int_{z\in A}\int_{y\in\mathcal{X}_{i+k}} P(x,dy)P^{l^{'}(y)(d-i-k)}(y,dz) &= P^{l^{'}(y)d-i-k+1}(A)\\ \geq \int_{z\in A}\int_{y\in\mathcal{X}_{i+k}} P(x,dy)\delta(y)\nu(dz) &\geq \delta^{'}(y)\nu(A)\end{aligned} \tag{1.17}$$

Thus, $x \in \mathcal{X}'_i \cap \mathcal{X}'_{i+k-1}$ then x would be in two different $\mathcal{X}'_j$ at once, contradicting their jointedness.

We claim that for all $m \geq 0$, $\nu P^m(\mathcal{X}_i \cap \mathcal{X}_j) = 0$ whenever $i \neq j$. If $\nu P^m(\mathcal{X}_i \cap \mathcal{X}_j) > 0$, then there would $S' \subseteq \mathcal{X}$, $l_1, l_2 \in \mathbf{N}$ and $\delta > 0$ such that for all $x \in S^{'}$, $P^{l_1 d-i}(x,\cdot) \geq \delta\nu(\cdot)$ and $P^{l_2 d-i}(x,\cdot) \geq \delta\nu(\cdot)$ implying that $l_1 d - i + m \in T$ and $l_2 d - j + m \in T$, contradicting the fact that $gcd(T) = d$. Then $m = 0$, we have $\nu(\mathcal{X}_i \cap \mathcal{X}_j) = 0$ for $i \neq j$. If $m > 0$, then $\nu(\{x \in \mathcal{X}; m \in \mathbf{N}\ s.t.\ P^m(x,S) \geq \delta\nu(\cdot)\}) = 0$. Then $\nu(\bar{S}) \leq \nu(\bigcup_{i\neq j}(\mathcal{X}_i \cap \mathcal{X}_j) + \nu(\{x \in \mathcal{X}; m \in \mathbf{N}\ s.t.\ P^m(x,S) \geq \delta\nu(\cdot)\}) \leq 0$, which implies $\nu(\bar{S}) = 0$. Therefore $\nu(\bigcup_{i=1}^d \mathcal{X}'_i) = \nu(\bigcup_{i=1}^d \mathcal{X}_i) - \nu(\bar{S}) = \nu(\mathcal{X}) = 1$. Since $\nu \ll \pi$, we must have $\pi(\bigcup_{i=1}^d \mathcal{X}'_i) > 0$.

Thus we conclude that from all of this that $\mathcal{X}'_1, ..., \mathcal{X}'_d$ are subsets of positive π-measure, which respect to which the chain is periodic, contradicting the assumption of aperiodicity □

Lemma 3. $\pi(\bar{G}) = 1$

Proof. First prove that $(\pi \times \pi)(G) = 1$. Indeed, since $\nu(C) > 0$, $\phi(C) > 0$, by Theorem (1.10) and since Markov chain is ϕ-irreducible, from lemma 2, we know for any $(x,y) \in \mathcal{X} \times \mathcal{X}$, the joint chain has positive probability of eventually hitting $C \times C$. By applying lemma 1,the joint chain will return to $C \times C$ with probability 1 from $\pi \times \pi$-a.e. $(x,y) \in C \times C$. Once the joint chain reaches $C \times C$, the joint chain 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 is absolutely continous with respect to $\pi \times \pi$ and hence by lemma 1, the joint chain will repeatedly 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 as $X_n = X'_n$.And by coupling construction, each time the joint chain is in $C \times C$, it has probability $\geq \epsilon$ of forcing $X_n = X'_n$. Hence, eventually, we will have $X_n = X'_n$, thus proving that $(\pi \times \pi)(G) = 1$
If $\pi(\bar{(}G) < 1$, contradicting with the fact that $(\pi \times \pi)(G) = 1$. □

1.5.5 Proof of Theorem 1.9

First assume that $n_0 = 1$ in the minorisation condition for the small set for the small C, then we write B_{n_0} as B, then we consider the case when $n_0 > 1$.

Let

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

and let τ_1, τ_2,... be the times of the successive visits of $\{(X_n, X'_n)\}$ to $C \times C$. Then for any integer j with $1 \leq j \leq k$,

$$\mathbf{P}[X_k \neq X'_k] = \mathbf{P}[X_k \neq X'_k, N_{k-1} \geq j] + \mathbf{P}[X_k \neq X'_k, N_{k-1} < j]$$

The event $X_k \neq X'_k, N_{k-1} \geq j\}$ is contained in the event that the first j coin flips all camp up tails. Hence, $\mathbf{P}[X_k \neq X'_k, N_{k-1} \geq j] \leq (1-\epsilon)^j$, which bounds the first term in (1.9).

To bound the second term in (1.9), let

$$M_k = \alpha^k B^{-N_{k-1}} h(X_k, X'_k) \mathbf{1}(X_k \neq X'_k), \quad k = 0, 1, 2, ...$$

where $(N_{-1} = 0)$.

Lemma 4. *We have*

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

i.e. $\{M_K\}$ *is a supermartingale.*

Proof. If $(X_k, X'_k) \notin C \times C$, then $N_k = N_{k-1}$, so

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

by (1.7). If $(X_k, X'_k) \in C \times C$, then $N_k = N_{k-1} + 1$, assuming $X_k \neq X'_k$(since if $X_k = X'k$, then the result is trivial, then we have

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

by (1.8). Hence, $\{M_k\}$ is a supermartingale. □

Since $B\geq 1$, we have

$$\begin{aligned}
\mathbf{P}[X_k\neq X'_k,&N_{k-1}<j]=\mathbf{P}[X_k\neq X'_k,N_{k-1}\leq j-1]\leq\mathbf{P}[X_k\neq X'_k,B^{-N_{k-1}}\geq B^{-(j-1)}]\\
&=\mathbf{P}[\mathbf{1}(X_k\neq X'_k)B^{-N_{k-1}}\geq B^{-(j-1)}]\leq B^{j-1}\mathbf{E}[\mathbf{1}(X_k\neq X'_k)B^{-N_{k-1}}] \quad (by\ Markov's\ inequality)\\
&\leq B^{j-1}\mathbf{E}[\mathbf{1}(X_k\neq X'_k)B^{-N_{k-1}}h(X_k,X'_k)] \quad (since\ h\geq 1)\\
&=\alpha^{-k}B^{j-1}\mathbf{E}[M_k] \quad \leq\alpha^{-k}B^{j-1}\mathbf{E}[M_0] \quad (since\ \{M_k\}\ is\ supermartingale)\\
&=\alpha^{-k}B^{j-1}\mathbf{E}[h(X_0,X'_0)]
\end{aligned}$$

Theorem 1.9 then follows by combining these two bounds.

Finally, we consider the changes required if $n_0>1$. In this case, the visits to $C\times C$ corresponding to the "filling in" times for going back and constructing $X_{n+1},...,X_{n+n_0}$(also X') in step 2 of the coupling contruction should not be counted. Thus let N_k count the number of visits to $C\times C$, and $\{\tau_i\}$ the actual visit times, avoiding all such "filling in" times. Thus, we consider,

$$\mathbf{P}[X_k\neq X'_k]=\mathbf{P}[X_k\neq X'_k,N_{k-n_0}\geq j]+\mathbf{P}[X_k\neq X'_k,N_{k-n_0}<j]$$

Same as above, the first part is bounded by $(1-\epsilon)^j$. Considering the second part, define $t(k)$ as the latest time $\leq k$ which does not correspond to a "filling in" time. Then $\{M_{t(k)}\}$ is a martingale, where $M_k=\alpha^kB^{-N_{k-n_0}}h(X_k,X'_k)\mathbf{1}(X_k\neq X'_k), \quad k=0,1,2,...$.

Proof. If $(X_{t(k)},X'_{t(k)})\notin C\times C$, then $N_k=N_{k-n_0}$ and $t(k)=k$. Then the proof follows the same as first part of the proof above.

If $(X_{t(k)},X'_{t(k)})\in C\times C$, then we only consider the case such that $(X_{t(k+1)},X'_{t(k+1)})=(X_{s+n_0},X'_{s+n_0})$, where $s=k-n_0+1$ is the latest time $\leq k$ which does not correspond to a "filling time", otherwise the result is trivial. Then $M_{t(k+1)} = M_{s+n_0}$ and $N_s = N_{s-n_0} + 1$

$$\begin{aligned}
&\mathbf{E}[M_{s+n_0}|X_0, ..., X_s, X'_0, ..., X'_s] \\
&\quad = \alpha^{s+n_0} {B_{n_0}}^{-N_{s-n_0}-1}\mathbf{E}[h(X_{s+n_0}, X'_{s+n_0})\mathbf{1}(X_{s+n_0} \neq X'_{s+n_0})|X_s, X'_s] \\
&\quad = \alpha^{s+n_0} {B_{n_0}}^{-N_{s-n_0}-1}(1-\epsilon)(\bar{R}h)(X_s, X'_s) \\
&\quad = M_s \alpha^{n_0} {B_{n_0}}^{-1}(1-\epsilon)(\bar{R}h)(X_s, X'_s)/h(X_s, X'_s) \\
&\quad \leq M_s \quad ,
\end{aligned}$$

Thus, $M_{t(k)}$ is a martingale. □

Since $B_{n_0} \geq 1$, and only consider k that is not correspondent to the filling time. we have

$$\begin{aligned}
\mathbf{P}[X_k \neq X'_k, N_{k-n_0} < j] &\leq \mathbf{P}[X_{t(k)} \neq X'_{t_{(k)}}, N_{t(k)-n_0} < j] \\
&= \mathbf{P}[X_{t(k)} \neq X'_{t(k)}, N_{t(k)-n_0} \leq j-1] \leq \mathbf{P}[X_{t(k)} \neq X'_{t(k)}, {B_{n_0}}^{-N_{t(k)-n_0}} \geq {B_{n_0}}^{-(j-1)}] \\
&= \mathbf{P}[\mathbf{1}(X_{t(k)} \neq X'_{t(k)}){B_{n_0}}^{-N_{t(k)-n_0}} \geq {B_{n_0}}^{-(j-1)}] \\
&\leq {B_{n_0}}^{j-1}\mathbf{E}[\mathbf{1}(X_{t(k)} \neq X'_{t(k)}){B_{n_0}}^{-N_{t(k)-n_0}}] \quad (by\ Markov's\ inequality) \\
&\leq {B_{n_0}}^{j-1}\mathbf{E}[\mathbf{1}(X_{t(k)} \neq X'_{t(k)}){B_{n_0}}^{-N_{t(k)-n_0}} h(X_{t(k)}, X'_{t(k)})] \quad (since\ h \geq 1) \\
&= \alpha^{-t(k)}{B_{n_0}}^{j-1}\mathbf{E}[M_{t(k)}] \quad \leq \alpha^{-t(k)}{B_{n_0}}^{j-1}\mathbf{E}[M_0] \quad (since\ \{M_{tk}\}\ is\ supermartingale) \\
&= \alpha^{-k}{B_{n_0}}^{j-1}\mathbf{E}[h(X_0, X'_0)]
\end{aligned}$$

Theorem 1.9 then follows by combining these two bounds and two cases.

1.5.6 Proof of Theorem 1.7

The proof of this theorem makes use of Theorem 1.9. To begin, set $h(x, y) = \frac{1}{2}[V(x) + V(y)]$. The proof will use the following technical result.

Lemma 5. *We may assume without loss of generality that*

$$\sup_{x \in C} V(x) < \infty \tag{1.18}$$

Specifically, given a small set C and drift condition V satisfying (1.4) and (1.5), we can find a small set $C_0 \subseteq C$ such that (1.4) and (1.5) still hold (with the same n_0 and ϵ and b, but with λ replaced by some $\lambda_0 < 1$), and such that (1.18) also holds.

Proof. Let λ and b be as in (1.5). Choose δ with $0 < \delta < 1 - \lambda$, let $K = b/(1 - \lambda - \delta)$, and set

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

Since $C_0 \subseteq C$, (1.4) continues to hold on C_0. It then remains to verify that (1.5) holds with C replaced by C_0, and λ replaced by λ_0. clearly, (1.5) holds for $x \in C_0$ and $x \notin C$. Finally for $x \in C \setminus C_0$, we have $V(x) \geq K$, and so using the original drift condition 1.5, we have

$$(PV)(x) \leq \lambda V(x) + b\mathbf{1}_C(x) = (1-\delta) - (1-\lambda-\delta)V(x) + b$$

$$\leq (1-\delta)V(x) - (1-\lambda-\delta)K + b = (1-\delta)V(x) = \lambda_0 V(x),$$

showing that (10) still holds, with C replaced by C_0 and λ replaced by λ_0. □

Thus, for the reminder of the proof, it is valid to assume that (1.18) holds. Together with (1.6), implies that

$$\sup_{(x,y)\in C\times C} \bar{R}h(x,y) < \infty, \tag{1.19}$$

which ensures that the B_{n_0} in (1.8) is finite.

Let $d = \inf_{C^c} V$. Then we see from Proposition 1.8 that the bivariate drift condition (1.7) will hold, provided that $d > b/(1-\lambda) - 1$. In that case, Theorem 1.7 follows (in a quantitative version) by combining Proposition 1.8 with Theorem 1.9.

However, if $d \leq b/(1-\lambda) - 1$, then this argument does not go through. The plan is to enlarge the small set C so that the new value of d satisfies $d > b/(1-\lambda) - 1$ and to use aperiodicity to show that C remains a small set. Theorem 9 will then follow from Proposition 1.8 and Theorem 1.9 as above. Because there is no direct control over the new values of n_0 and ϵ, this approach does not provide a quantitative convergence rate bound.

To proceed, choose any $d' > b/(1-\lambda) - 1$, let $S = \{x \in \mathcal{X}; V(x) \leq d'\}$, and set $C' = C \cup S$. Thus $\inf_{x\in C'^c} V(x) \geq d' > b/(1-\lambda) - 1$. Furthermore, since V is bounded on S by construction, then (1.18) will still hold with C replaced by C'. It then follows from (1.19) that it is still ture that $B_{n_0} < \infty$. Now, it remains to show that C' is a small set.

Before continuing, the notion of "petite set" is introduced.

Definition 15. A subset $C \subseteq \mathcal{X}$ is *petite* (or, (n_0, ϵ, ν)-petite), relative to a Markov chain P, if there exists a positive integer n_0, $\epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that

$$\sum_{i=1}^{n_0} P^i(x,\cdot) \geq \epsilon\nu(\cdot) \quad x \in C. \tag{1.20}$$

Intuitively, the definition of petite set is like that of small set, except that it allows different states in C to cover the minorisation measure $\epsilon\nu(\cdot)$ at different times i (for each $x \in C$, $\exists i$ and $\delta(x) > 0$ such that $P^i(x,\cdot) \geq \delta(x)\epsilon\nu(\cdot)$) . It is obvious that every small set is petite set but the converse is false in general, as the petite set condition does not itself rule out *periodic* behaviour of the chain (for example, some of the states $x \in C$ cover $\epsilon\nu(\cdot)$ only at odd times, and others only at even times). However, for an aperiodic, ϕ-irreducible Markov chain, all petite sets are small sets.

Lemma 6. *For an aperiodic, ϕ-irreducibel Markov chain, all petite sets are small sets.*

Proof. Let R be $(n_0, \epsilon, \nu(\cdot))$-petite, so that $\sum_{i=1}^{n_0} P^i(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in R$. Let T be as in Lemma 2. Then $\int_{x\in\mathcal{X}} \sum_{i=1}^{n_0} \nu(dx)P^i(x,\cdot) = \sum_{i=1}^{n_0} \int_{x\in\mathcal{X}} \nu(dx)P^i(x,\cdot) \geq \epsilon\nu(\cdot)$, so we must have $i \in T$ for some $1 \leq i \leq n_0$, so that T is non-empty. Hence, from Lemma 2, we can find n_* and $\delta_n > 0$ such that $\int \nu(dx)P^n(x,\cdot) > \delta_n\nu(\cdot)$ for all $n \geq n_*$. Let $r = \min\Big\{\delta_n; n_* \leq n \leq n_* + n_0 - 1\Big\}$, and set $N = n_* + n_0$. Then for $x \in R$,

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

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

To make use of Lemma 6, the following lemma is stated.

Lemma 7. *Let $C' = C \cup S$ where $S = \{x \in \mathcal{X}; V(x) \leq d'\}$ for some $d' < \infty$, as above. Then C' is petite.*

Proof. 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 drift condition (1.5) implies that W_n is a supermartingale. Indeed, if $\tau_C \leq n$, then

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

while if $\tau_C > n$, then $X_n \notin C$, so using (1.5),

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

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

$$\begin{aligned}\mathbf{P}[\tau_C \geq N | X_0 = x] &= \mathbf{P}[\lambda^{-\tau_C} \geq \lambda^{-N} | X_0 = x] \\ &\leq \lambda^N \mathbf{E}[\lambda^{-\tau_C} | X_0 = x] \leq \lambda^N \mathbf{E}[W_{\tau_C} | X_0 = x] \\ &\leq \lambda^N \mathbf{E}[W_0 | X_0 = x] = \lambda^N V(x) \leq \lambda^N d',\end{aligned}$$

so that $\mathbf{P}[\tau_C < N | X_0 = x] \geq r$.

On the other hand, recall that C is $(n_0, \epsilon, \nu(\cdot))$-small, so that $P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot)$ for $x \in C$. Then for $x \in S$

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

So $C' = S \cup C$ is petite. Then we have,

Lemma 8. *C' is a small set*

Proof. Combine Lemma 6 and Lemma 7, then we prove this Lemma 8. □

Hence, Theorem 9 is proved. □

Chapter 2

Fundamentals of Adaptive MCMC

[4]

2.1 Preliminaries

Let $\pi(\cdot)$ be a fixed "target" probability distribution, on a state space $\mathcal{X}$ with σ-algebra $\mathcal{F}$. The goal of MCMC is to approximately sample from $\pi(\cdot)$ through the use of Markov chains.

Let $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain kernels on $\mathcal{X}$, each of which has $\pi(\cdot)$ as a stationary distribution: $(\pi P_\gamma)(\cdot) = \pi(\cdot)$.

Assuming P_γ is ϕ-irreducible and aperiodic, this implies that P_γ be ergodic for $\pi(\cdot)$, i.e. that for all x, $\lim_{n\to\infty}\|P_\gamma^n(x,\cdot) - \pi(\cdot)\|$. That is, P_γ represents a "valid" MCMC algorithm. So, if γ is fixed, then the Markov chain algorithm described by P_γ will eventually converge to $\pi(\cdot)$.

Some choices of γ may lead to far less efficient algorithms than others and to know in advance which choices of γ are preferable might be difficult. To make the algorithm as efficient as possible, adaptive MCMC proposes that at each time n, the choice of γ is given by a $\mathcal{Y}$-valued random variable Γ_n, updated according to specified rules.

Formally, for $n = 0, 1, 2, ...$, we propose a $\mathcal{X}$-value random variable X_n representing the state of the algorithm at time n, and a $\mathcal{Y}$-valued random variable Γ_n representing the choice of kernel to be used when updating from X_n to X_{n+1}. We let

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

be the filtration generated by $\{(X_n, \Gamma_n\}$. Thus,

$$\mathbf{P}[X_{n+1} \in B | X_n = x, \Gamma_n = \gamma, \mathcal{G}_{n-1}] = P_\gamma(x, B), \quad x \in \mathcal{X}, \gamma \in \mathcal{Y}, B \in \mathcal{F}, \tag{2.1}$$

while the conditional distribution of Γ_{n+1} given $\mathcal{G}_n$ is to be specified by the particular adaptive algorithm being used. We let

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

record the conditional probabilities for X_n for the adaptive algorithm, given the initial conditions $X_0 = x$ and $\Gamma_0 = \gamma$.

Finally, we let

$$T(x, \gamma, n) = \|A^{(n)}((x,\gamma), \cdot) - \pi(\cdot)\| \equiv \sup_{B\in\mathcal{F}} |A^{(n)}((x,\gamma), B) - \pi(B)|$$

denote the total variation distance bewteen the distribution of our adaptive algorithm at time n, and the targe distribution $\pi(\cdot)$. The adaptive algorithm is ergodic if $\lim_{n\to\infty} T(x, \gamma, n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$. In this section below, we will try to answer the question "Will the adaptive chain necessarily be ergodic?"

2.2 Some special cases

First introduce some special cases of Adaptive MCMC schemes,

- Traditional MCMC: $\Gamma_n \equiv 1$ for all n.
- Systematic-scan hybrid algorithm: $(\Gamma_n) = (1, 2, ..., d, 1, 2, ..., d, 1, 2, ...)$, where P_i moves only the i^{th} coordinate.
- Random-scan hybrid algorithm: $\{\Gamma_n\}$ are i.i.d. $\sim$ Uniform{1,2,...,d}.

All those algorithms above are *independent adaptation* such that for all n, Γ_n is independent of X_n. For independent adaptations, stationary of $\pi(\cdot)$ is guaranteed:

Proposition 2.1. Consider an independent adaptation algorithm $A^{(n)}((x,\gamma),\cdot)$, where $\pi(\cdot)$ is stationary for each $P_\gamma(x,\cdot)$. Then $\pi(\cdot)$ is also stationary for each $P_\gamma(x,\cdot)$. Then $\pi(\cdot)$ is also stationary for $A^{(n)}((x,\gamma),\cdot)$, i.e.

$$\int_{x\in\mathcal{X}} \mathbf{P}[X_{n+1}\in B|X_n = x, \mathcal{G}_{n-1}]\pi(dx) = \pi(B), \quad B \in \mathcal{F}$$

Proof. Using (1), and the independence of Γ_n and X_n, and the stationary of $\pi(\cdot)$ for P_γ, we have:

$$\begin{aligned}
&\int_{x\in\mathcal{X}} \mathbf{P}[X_{n+1}\in B|X_n = x, \mathcal{G}_{n-1}]\pi(dx)\\
=&\int_{x\in\mathcal{X}}\int_{\gamma\in\mathcal{Y}} \mathbf{P}[X_{n+1}\in B|X_n = x, \Gamma_n=\gamma, \mathcal{G}_{n-1}]\mathbf{P}[\Gamma_n \in d\gamma|X_n = x, \mathcal{G}_{n-1}]\pi(dx)\\
=&\int_{x\in\mathcal{X}}\int_{\gamma\in\mathcal{Y}} P_\gamma(x,B)\mathbf{P}[\Gamma_n\in d\gamma|\mathcal{G}_{n-1}]\pi(dx)\\
=&\int_{\gamma\in\mathcal{X}} \mathbf{P}[\Gamma_n\in d\gamma|\mathcal{G}_{n-1}]\int_{x\in\mathcal{X}} P_\gamma(x,B)\pi(dx)\\
=&1\cdot\pi(B) = \pi(B)
\end{aligned}$$

□

However, for independent adaptions, irreducibility might be destroyed

Example 5. Let $\mathcal{X} = \{1,2,3,4\}$, with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 2/7$, and $\pi\{4\} = 1/7$. Let $P_1(1,\{2\}) = P_1(3,\{1\}) = P_1(4,\{3\}) = 1$, and $P_1(2,\{3\}) = P_1(2,\{4\}) = 1/2$. Similarly, $P_2(2,\{1\}) = P_2(3,\{2\}) = P_2(4,\{3\}) = 1$, and $P_2(1,\{3\}) = P_2(1,\{4\}) = 1/2$. We could check that two chains P_1 and P_2 are irreducible and aperiodic, with stationary distribution $\pi(\cdot)$. On the other hand, $(P_1P_2)(1,\{1\}) = 1$, so when beginning in state 1, the systematic-scan adaptive chain P_1P_2 alternates bewteen states 1 and 2 but never reaches the state 3. Hence, this adaptive algorithm fails to be irreducible, and also $T(x,\gamma,n)$ does not converge to as $n\to\infty$, even though each individual P_i is ergodic.

2.2.1 Examples

To illustrate the limitations of adaptive MCMC, and the application of our theorems, the following running example is presented.

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

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

i.e. uniform on all the integers within γ of x, aside from x itself. The kernel P_γ then proceeds, given X_n, Γ_n, by first choosing a proposal state $Y_{n+1} \sim Q_{\Gamma_n}(X_n, \cdot)$. With probability $\min[1, \pi(Y_{n+1})/\pi(X_n)]$ it then accepts this proposal by setting $X_{n+1} = Y_{n+1}$. Otherwise, with probability 1-$\min[1, \pi(Y_{n+1})/\pi(X_n)]$, it rejects this proposal by setting $X_{n+1} = X_n$ (If $Y_{n+1} \notin \mathcal{X}$, then the proposal is always rejected, which corresponds to setting $\pi(y) = 0$ for $y \notin \mathcal{X}$.)

We define the adaptive scheme as follows. Begin with $\Gamma_0 = 1$(say). Let $M \in \mathbf{N} \cup \{\infty\}$ and let $p : \mathbf{N} \to [0, 1]$. For $n = 0, 1, 2, ...,$ given X_n and Γ_n, if the next proposal is accepted (i.e., if $X_{n+1} \neq X_n$) and $\Gamma_n < M$, then with probability $p(n)$ let $\Gamma_{n+1} = \Gamma_n + 1$, otherwise let $\Gamma_{n+1} = \Gamma_n$. Otherwise, if the next proposal is rejected (i.e., if $X_{n+1} = X_n$) and $\Gamma_n > 1$, then with probability $p(n)$ let $\Gamma_{n+1} = \Gamma_n - 1$, otherwise let $\Gamma_{n+1} = \Gamma_n$. In words, with probability $p(n)$, we increase γ (to a maximum of M) each time a proposal is accepted, and decrease γ (to minimum of 1) each time a proposal is rejected.

The specific versions of this scheme is recorded below:

- The "original running example" has $M = \infty$ and $p(n) \equiv 1$, i.e. it modifies Γ_n in every iteration except when $\Gamma_n = 1$ and the next proposal is rejected

- The "singly-modified running example" has $M = \infty$ but arbitrary $p(n)$.

- The "doubly-modified running example "has $M < \infty$ and arbitrary $p(n)$.

- The "One-Two" version has $M = 2$ and $p(n) \equiv 1$.

We now provide an example that such adaptive scheme can completely destroy convergence to $\pi(\cdot)$:

Example 6. Let $\epsilon > 0$, and consider One- Two version with $K = 4$, $a = \epsilon$, and $b = \epsilon^3$. Then it is easily verified that there is $c > 0$ such that $\mathbf{P}[X_3 = \Gamma_3 = 1 | X_0 = x, \Gamma_0 = \gamma] \geq c\epsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, i.e. the algorithm has $O(\epsilon)$ probability of reaching the configuration $\{x = \gamma = 1\}$. On hte other hand, $\mathbf{P}[X_1, \Gamma_1 = 1 | X_0 = \Gamma_0 = 1] = 1 - \epsilon^2/2$. On the other hand, $\mathbf{P}[X_1, \Gamma_1 = 1 | X_0 = \Gamma_0 = 1] = 1 - \epsilon^2/2$, i.e.

the algorithm has just $O(\epsilon^2)$ probability of leaving the configuration $\{x = \gamma = 1\}$ once it is there. This probabilistic asymmetry implies that $\lim_{\epsilon\to 0}\lim_{n\to\infty}\mathbf{P}[X_n = \Gamma_n = 1] = 1$. Hence,

$$\lim_{\epsilon\to 0}\lim_{n\to\infty} T(x,\gamma,n) \geq \lim_{\epsilon\to 0}(1-\pi\{1\}) = \lim_{\epsilon\to 0}(1-\epsilon) = 1. \tag{2.2}$$

In particular, for any $\delta > 0$, there is $\epsilon > 0$ with $\lim_{n\to\infty} T(x,\gamma,n) \geq 1-\delta$, so the algorithm does not converge at all.

2.2.2 Uniformly Converging Case.

Theorem 2.2. *Consider an adaptive MCMC algorithm, on a state space $\mathcal{X}$, with adaption index $\mathcal{Y}$, so $\pi(\cdot)$ is stationary for each kernel P_γ for $\gamma \in \mathcal{Y}$. Assume that:*

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

(b) [Diminishing Adaptation] $\lim_{n\to\infty} D_n = 0$ in probability, where $D_n = \sup_{x\in\mathcal{X}}\|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\|$ is a $\mathcal{G}_{n+1}$-measurable random variable (depending on random values Γ_n and Γ_{n+1}).

Proof. Let $\epsilon > 0$, by (a), we choose $N = N(\epsilon)$. Let $H_n = \{D_n \geq \epsilon/N^2\}$ and use condition (b) to choose $n^* = n^*(\epsilon) \in \mathbf{N}$ large enough so that

$$\mathbf{P}(H_n) \leq \epsilon/N, \quad n \geq n^* \tag{2.3}$$

Fix a "target time" $K \geq n^* + N$. The idea of the proof is to construct a coupling which depends on the target time K and to prove $\mathcal{L}(X_k) \approx \pi(\cdot)$. Define the event $E = \cap_{i=n}^{n+N-1} H_i^c$. It follows from 2.3 that for $n \geq n^*$, we have $\mathbf{P}(E) \geq 1-\epsilon$. By the triangle inequality and induction, on event E we have $\sup_{x\in\mathcal{X}}\|P_{\Gamma_{n+k}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \leq \epsilon/N$ for all $k \leq N$, and in particular

$$\|\mathbf{P}_{\Gamma_{K-N}}(x,\cdot) - \mathbf{P}_{\Gamma_m}(x,\cdot)\| < \epsilon/N \ on\ E, \quad x \in \mathcal{X}, K-N \leq m \leq K \tag{2.4}$$

we first construct the original adaptive chain $\{X_n\}$ together with the adaptation sequence $\{\Gamma_n\}$, starting with $X_0 = x$ and $\Gamma_0 = \gamma$. Claim that on E, we could construct a second chain $\{X_n'\}_{n=K-N}^K$ such that $X_{K-N}' = X_{K-N}$, and $X_n' \sim P_{\Gamma_{K-N}}(X_{n-1}',\cdot)$ for $K-N+1 \leq n \leq K$, and $\mathbf{P}[X_i' = X_i\ for\ K-N \leq i \leq m] \geq 1 - [m-(K-N)]\epsilon/K$.

The claim is trivial true for $m = K-N$. Assume the inequality holds for some value m. Then conditional on $\mathcal{G}_m$ and the event that $X_i' = X_i$ for $K-N \leq i \leq m$, we have $X_{m+1} \sim P_{\Gamma_m}(X_m,\cdot)$ and $X_{m+1}' \sim P_{\Gamma_{K-N}}(X_m',\cdot) = P_{\Gamma_{K-N}}(X_m,\cdot)$. It follow from 2.4 that the conditional distributions of X_{m+1} and X_{m+1}' within ϵ/N of each other. Then by Proposition (1.3)(g), we know that the conditional probability that $X_{m+1} = X'_{m+1}$ is great than or equal to $1-\epsilon/N$. It follows that

$$\begin{aligned}\mathbf{P}[X_i = X'_i \ for\ K-N \le i \le m+1] &\ge \mathbf{P}[X_i = X'_i \ for\ K-N \le i \le m](1-\epsilon/N)\\ &= (1-[m-(K-N)]\epsilon/N)(1-\epsilon/N)\\ &= 1-[m-(K-N)]\epsilon/N - \epsilon/N\\ &= 1-[(m+1)-(K-N)]\epsilon/N + [m-(K-N)](\epsilon/N)^2\\ &\ge 1-[(m+1)-(K-N)]\epsilon/N\end{aligned}$$

Thus, the claim follows from the induction.

This shows that on E, $\mathbf{P}[X'_K = X_K] \ge 1-(K-(K-N))\epsilon/N = 1-\epsilon$. That is, $\mathbf{P}[X'_K \ne X_K, E] < \epsilon$.

Using the condition (a) that conditioning on X_{K-N}, we have $\|P^N_{\Gamma_{K-N}}(X_{K-N},\cdot) - \pi(\cdot)\| < \epsilon$. Then $\|\int P^N_{\Gamma_{K-N}}(x,\cdot)P^{K-N}_\gamma(y,dx) - \int \pi(\cdot)P^{K-N}_\gamma(y,dx)\| = \|\mathcal{L}(X'_K) - \pi(\cdot)\| < \epsilon \int P^{K-N}_\gamma(y,dx) = \epsilon$. It again follows from Proposition 1.3 (g) that we can construct $Z \sim \pi(\cdot)$ such that $\mathbf{P}[X'_K \ne Z] < \epsilon$. Furthermore, we can construct all of $\{X_n\}$, $\{X'_n\}$ and Z jointly on a common probability space, by first constructing $\{X_n\}$ and $\{X'_n\}$ as above, and then constructing Z conditional on $\{X_n\}$ and $\{X'_n\}$ from any conditional distribution satisfying that $Z \sim \pi(\cdot)$ and $\mathbf{P}[X'_K \ne Z] < \epsilon$.

Then on event E, we have $\|\mathcal{L}(X_K) - \pi(\cdot)\| = \|\mathcal{L}(X_K) - \mathcal{L}(X^{'}_K) + \mathcal{L}(X^{'}_K) - \pi(\cdot)\| \le \|\mathcal{L}(X_K) - \mathcal{L}(X^{'}_K)\| + \|\mathcal{L}(X^{'}_K) - \pi(\cdot)\|$. By coupling inequality, on event E, we have $\|\mathcal{L}(X_K) - \pi(\cdot)\| \le \mathbf{P}[X_K \ne X'_K] + \mathbf{P}[X'_K \ne Z]$. Thus we have

$$\mathbf{P}[X_K \ne Z] \le \mathbf{P}[X_K \ne Z, E] + \mathbf{P}[E^c] \le \mathbf{P}[X_k \ne X'_K, E] + \mathbf{P}[X'_K \ne Z, E] + \mathbf{P}[E^c] < \epsilon + \epsilon + \epsilon = 3\epsilon$$

Hence, $\|\mathcal{L}(X_K) - \pi(\cdot)\| < 3\epsilon$, i.e. $T(x,\gamma,K) < 3\epsilon$. Since $K \ge n^* + N$ was arbitrary, this means that $T(x,\gamma,K) \le 3\epsilon$ for all sufficiently large K. Hence, $\lim_{K\to\infty} T(x,\gamma,K) = 0$. □

We then have following corollaries from Theorem 2.2.

Corollary 2. Suppose an adaptive MCMC algorithm satisfies Diminishing Adaptation, and also that each P_γ is ergodic for $\pi(\cdot)$(i.e., $\lim_{n\to\infty}\|P^n_\gamma(x,\dot{)} - \pi(\cdot)\| = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$). Suppose further that $\mathcal{X}$ and $\mathcal{Y}$ are finite. Then the adaptive algorithm is ergodic.

Proof. Let $\epsilon > 0$. By assumption for each $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, there is $N(x,\gamma,\epsilon)$ such that $\|P^{N(x,\gamma,\epsilon)}_\gamma - \pi(\cdot)\| < \epsilon$. Letting $N(\epsilon) = \max_{x\in\mathcal{X},\gamma\in\mathcal{Y}} N(x,\gamma,\epsilon)$, we see that condition (a) of Theorem 5 is satisfied. So we could apply Theorem 5 and the result follows. □

Corollary 3. The doubly-modified running example (presented above) is ergodic provided that the adaptation probabilities $p(n)$ satisfy $\lim_{n\to\infty} p(n) = 0$.

Proof. In that example, for each γ, since Metropolis-Hasting Algorithm is used, P_γ is π-irreducible and aperiodic, and hence ergodic for $\pi(\cdot)$. Furthermore, both $\mathcal{X}$ and $\mathcal{Y}$ are finite. Also, for this scheme we have for each $n \in \mathbf{N}$ $D_n = a \leq 1$ with probability $p(n)$ and $D_n = 0$ with probability $1 - p(n)$, thus given $\epsilon > 0$, $\mathbf{P}(D_n > \epsilon) \leq p(n)$. Because $\lim_{n\to\infty} p(n) = 0$, as $n \to \infty$, $\mathbf{P}(D_n > \epsilon) \leq \epsilon$. Thus thee Diminishing Adaptation holds. Hence the result follows from Corollary 6. □

Corollary 4. Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaption property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that $\mathcal{X} \times \mathcal{Y}$ is compact in some topology, with respect to which the mapping $(x, \gamma) \to d(x, \gamma, n)$ is continuous for each fixed $n \in \mathbf{N}$. Then the adaptive algorithm is ergodic.

Proof. Fix $\epsilon > 0$. For $n \in \mathbf{N}$, let $\mathcal{W}_n \subseteq \mathcal{X} \times \mathcal{Y}$ be the set of all pairs of (x, γ) such that $\|P_\gamma^n(x, \cdot) - \pi(\cdot)\| < \epsilon$. Since each P_γ is ergodic, this means that every pair (x, γ) is in $\mathcal{W}_n$ for all sufficiently large n. Hence $\cup_n \mathcal{W}_n = \mathcal{X} \times \mathcal{Y}$.

By continuity, the pre-image of an open set is open, so $\mathcal{W}_n$ is an open set. By compactness of $\mathcal{X} \times \mathcal{Y}$, there is a finite set $\{n_1, ..., n_r\}$ such that $\mathcal{W}_{n_1} \cup ... \cup \mathcal{W}_{n_r} = \mathcal{X} \times \mathcal{Y}$. Letting $N = N(\epsilon) = \max[n_1, ..., n_r]$, we find that the condition of Theorem 5 is satisfied. The result follows. □

The following lemma is sometimes useful in applying Corollary 8.

Lemma 9. *Suppose the mapping $(x, \gamma) \to P_\gamma(x, \cdot)$ is continuous with respect to a product metric space topology, meaning that for each $x \in \mathcal{X}, \gamma \in \mathcal{Y}$, and $\epsilon > 0$, there is $\delta = \delta(x, \gamma, \epsilon) > 0$, such that $\|P_{\gamma'}(x', \cdot) - P_\gamma(x, \cdot)\| < \epsilon$ for all $x' \in \mathcal{X}$ and $\gamma' \in \mathcal{Y}$ satisfying $d(x, x') + d(\gamma, \gamma') < \delta$(for some distance metrics on $\mathcal{X}$ and $\mathcal{Y}$). Then for each $n \in \mathbf{N}$, the mapping $(x, \gamma) \to d(x, \gamma, n)$ is continuous.*

Proof. Given $x \in \mathcal{X}, \gamma \in \mathcal{Y}, n \in \mathbf{N}$ and $\epsilon > 0$, find $\delta > 0$ with $\|P_{\gamma'}(x', \cdot) - P_\gamma(x, \cdot)\| < \epsilon/n$ whenever $d(x, x') + d(\gamma, \gamma') < \delta$. Then given x' and γ' with $d(x, x') + d(\gamma, \gamma') < \delta$, as in proof of Theorem 5 we can construct X'_n and X_n with $X'_n \sim P^n_{\gamma'}(x', \cdot)$, $X_n \sim P^n_\gamma(x, \cdot)$. Specifically, starting with $X_0 = x$ and $\Gamma_0 = \gamma$, $X'_0 = x'$ and $\Gamma'_0 = \gamma'$, we construct the un-adaptive chain recursively as follows. First, we have $X'_1 \sim P_{\gamma'}(x', \cdot)$ and $X_1 \sim P_\gamma(x, \cdot)$, then given X'_m and X_m, we have $X'_{m+1} \sim P_{\gamma'}(X'_m, \cdot)$ and $X_{m+1} \sim P_\gamma(X_m, \cdot)$. We claim that $\mathbf{P}[X'_i = X_i \ for \ 1 \leq i \leq m] \geq 1 - m\epsilon/n$ for $1 \leq m \leq n$.

We show it by induction. When $m = 1$, by Proposition (1.3)(g), we can ensure that $\mathbf{P}[X'_1 = X_1] \geq 1 - \epsilon/n$. Then conditional on the event $X'_i = X_i$ for $1 \leq i \leq m$, then the conditional distribution of X_{m+1} and X'_{m+1} are within ϵ/n of each other. Hence by Proposition (1.3)(g) we can ensure that X_{m+1} and X'_{m+1} with probability $\geq 1 - \epsilon/n$. It follows that $\mathbf{P}[X'_i = X_i, \ for \ 1 \leq i \leq m+1] \geq \mathbf{P}[X'_i = X_i \ for \ 1 \leq i \leq m](1 - \epsilon/n) \geq [1 - m\epsilon/n][1 - \epsilon/n] \geq 1 - \frac{(m+1)}{n}\epsilon$. The claim then follows by induction.

Thus $\mathbf{P}[X_n = X'_n] \geq 1 - \epsilon$ by construction. So $\mathbf{P}[X_n \neq X'_n] < \epsilon$. Thus, by coupling inequality, we have $\|\mathcal{L}(X'_n) - \mathcal{L}(X_n)\| < \epsilon$. By triangle inequality, this implies that $\|\mathcal{L}(X'_n) - \pi(\cdot)\|$ and $\|\mathcal{L}(X_n) - \pi(\cdot)\|$ are within ϵ of each other. It is also clear that $X_n \sim P^n_\gamma(x,\cdot), X'_n \sim P^n_{\gamma'}(x,\cdot)$. Hence $\|P^n_\gamma(x,\cdot) - \pi(\cdot)\|$ and $\|P^n_{\gamma'}(x,\cdot) - \pi(\cdot)\|$ are within ϵ of each other, thus giving the result. □

The continuity conditions in Lemma 9 will be satisfied if the transition kernels have bounded densities with continuous dependencies.

Corollary 5. Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, $P_\gamma(x, dz) = f_\gamma(x,z)\lambda(dz)$ has a density $f_\gamma(x,\cdot)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$. Finally, suppose the $f_\gamma(x,z)$ are uniformly bounded, and that for each fixed $z \in \mathcal{X}$, the mapping $(x,\gamma) \to f_\gamma(x,z)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then the adaptive algorithm is ergodic.

Proof. We have that

$$\|P_{\gamma'}(x',\cdot) - P_\gamma(x,\cdot)\| = \frac{1}{2}\int_{\mathcal{X}} [M(y) - m(y)]\lambda(dy), \tag{2.5}$$

By continuity of the mapping $(x,\gamma) \to f_\gamma(x,y)$, and the finiteness of $\lambda(\cdot)$, it follows from the Bounded Convergence Theorem that the mapping $(x,\gamma) \to P_\gamma(x,\cdot)$ is continuous. The result then follows by applying Lemma 9 to Corollary 2.2. □

Metropolis-Hastings algorithms do not have densities because they have positive probability of rejecting the proposal. However, if the proposal kernels have densities, then asimilar result srill holds:

Corollary 6. Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each P_γ is ergodic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, P_γ represents a Metropolis-Hasting algorithm with proposal kernel $Q_\gamma(x,dy) = f_\gamma(x,y)\lambda(dy)$ having a density $f_\gamma(x,\cdot)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$, with corresponding density g for $\pi(\cdot)$ so that $\pi(dy) = g(y)\lambda(dy)$. Finally, suppose that the $f_\gamma(x,y)$ are uniformly bounded, and for each fixed $y \in \mathcal{X}$, the mapping $(x,\gamma) \to f_\gamma(x,y)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then the adaptive algorithm is ergodic.

Proof. In this case, the probability of accepting a proposal from x is given by:

$$a_\gamma(x) = \int_{\mathcal{X}} \min\left[1, \frac{g(y)f_\gamma(y,x)}{g(x)f_\gamma(x,y)}\right] f_\gamma(x,y)\lambda(dy),$$

which is a jointly continuous function of $(x, \gamma) \in \mathcal{X} \times \mathcal{Y}$ by the Bounded convergence theorem. The probability measure $P_\gamma(x, \cdot)$ as:

$$P_\gamma(x, dz) = [1 - a_\gamma(x)]\delta_x(dz) + p_\gamma(x, z)\lambda(dz) \tag{2.6}$$

hwere $p_\gamma(x, z)$ is jointly continuous in x and γ. Iterating this, we can write the $n-$step transition law as:

$$P^n_\gamma(x, dz) = [1 - a_\gamma(x)]^n\delta_x(dz) + p^n_\gamma(x, z)\lambda(dz)$$

for appropriate jointly continuous $p^n_\gamma(x, z)$.

We can assume without loss of generality that $a_\gamma(x) = 1$ whenever $\lambda\{x\} > 0$, i.e. that $\delta_x(\cdot)$ and $\pi(\cdot)$ are orthogonal measures. (Indeed, if $\lambda\{x\} > 0$, then we can modify the proposal densities so as to include $[1 - a_\gamma(x)]\delta_x(dz)$ as part of $p_\gamma(x, z)\lambda(dz)$.) It then follows

$$\|P^n_\gamma(x, \cdot) - \pi(\cdot)\| = [1 - a_\gamma]^n + \frac{1}{2}\int_{\mathcal{X}} |p^n(x, z) - g(z)|\lambda(dz).$$

This quantity is jointly continuous in x and γ, again by the Bounded Convergence Theorem. Moreover, by ergodicity, it converges to zero as $n \to \infty$ for each fixed x and γ. Hence, the results follows by Corollary 5. □

2.2.3 Non-uniformly converging Case

In this section the uniform convergence rate condition (a) of Theorem 2.2 is relaxed. To proceed, for $\epsilon > 0$, define the "ϵ convergence time function" $M_\epsilon : \mathcal{X} \times \mathcal{Y} \to \mathbf{N}$ by

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

If each individual P_γ is ergodic, then $M_\epsilon(x, \gamma) < \infty$.

Theorem 2.3. *Consider an adaptive MCMC algorithm with Diminishing Adaptation (i.e., $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| = 0$ in probability). Let $x_* \in \mathcal{X}$ and $\gamma_* \in \mathcal{X}$. Then $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$ provided that for all $\epsilon > 0$, the sequence $\{M_\epsilon(X_n, \Gamma_n)\}^\infty_{n=0}$ is bounded in probability given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, i.e. for all $\delta > 0$, there is $N \in \mathbf{N}$ such that $\mathbf{P}[M_\epsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$ for all $n \in N$.*

Proof. From the condition above, we could find $N \in \mathbf{N}$ such that (writing that $\mathcal{G}_0$ for $\{X_0 = x_*, \mathcal{Y}_0 = \gamma_*\}$),

$$\mathbf{P}[M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N | \mathcal{G}_0] \leq \epsilon$$

Since we start with $X_0 = x_*, \Gamma_0 = \gamma_*$, we have $\mathbf{P}[\mathcal{G}_0] = 1$. Thus, $\mathbf{P}[M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N] \leq \epsilon$.

From the proof of Theorem 2.2, we have for all $\epsilon > 0$ there is $n_* \in \mathbf{N}$ such that for all $N \in \mathbf{N}$ and all $K \geq n_* + N$, we can construct the chain $\{X_n\}$, $\{X'_n\}$, and $Z \sim \pi(\cdot)$ such that

$$\begin{aligned}
&\mathbf{P}[X_K \neq Z] \leq \mathbf{P}[X_K \neq X'_K, E] + \mathbf{P}[X'_K \neq Z, E] + \mathbf{P}[E^c] \\
&= \mathbf{P}[X_K \neq X'_K, E] + \mathbf{P}[X'_K \neq Z, E, M_\epsilon(X_{K-N}, \Gamma_{K-N}) \leq N] \\
&+ \mathbf{P}[X'_K \neq Z, E, M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N] + \mathbf{P}[E^c] \\
&= \mathbf{P}[X_K \neq X'_K, E] + \mathbf{P}[X'_K \neq Z, E | M_\epsilon(X_{K-N}, \Gamma_{K-N}) \leq N]\mathbf{P}[M_\epsilon(X_{K-N}, \Gamma_{K-N}) \leq N] \\
&+ \mathbf{P}[X'_K \neq Z, E | M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N]\mathbf{P}[M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N] + \mathbf{P}[E^c] \\
&< \epsilon + \epsilon + \epsilon + \mathbf{P}[M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N] = 3\epsilon + \mathbf{P}[M_\epsilon(X_{K-N}, \Gamma_{K-N}) > N] \leq 4\epsilon
\end{aligned}$$

By coupling inequality, $\|\mathcal{L}(X_K) - \pi(\cdot)\| < 4\epsilon$, i.e. $T(x_*, \gamma_*, K) < 4\epsilon$. Since $K \geq n^* + N$ is arbitrary, this means that $T(x_*, \gamma_*, K) < 4\epsilon$ for arbitrary large K. Thus $\lim_{n\to\infty} T(x_*, \gamma_*, n) = 0$.

□

Lemma 10. *Let $\{e_n\}_{n=0}^{\infty}$ be a sequence of real numbers. Suppose $e_{n+1} \leq \lambda_n + b$ for some $0 \leq \lambda < 1$ and $0 \leq b < \infty$, for all $n = 0, 1, 2, 3, ...$ Then* $\sup_n e_n \leq \max[e_0, b/(1-\lambda)]$

Proof. Prove this lemma by induction. For $n = 1$, we know $e_1 \leq \lambda e_0 + b$. Then either $e_1 \leq e_0$ or $e_1 \geq e_0$. If $e_1 \geq e_0$, then $\lambda e_1 + b \geq \lambda e_0 + b \geq e_1 \Rightarrow e_1 \leq b/(1-\lambda)$. Thus $\sup e_1 \leq \max[e_0, b/(1-\lambda)$.

Suppose $\sup_n e_n \leq \max[e_0, b/(1-\lambda)]$. We have two cases $e_{n+1} \leq e_n$ or $e_{n+1} \geq e_n$. If $e_{n+1} \leq e_n$, then $e_{n+1} < \max[e_0, b/(1-\lambda)]$. If $e_{n+1} \geq e_n$, $e_{n+1} \leq \lambda e_{n+1} + b \Rightarrow e_{n+1} \leq b/(1-\lambda)$, which implies $e_{n+1} \leq \max[e_0, b/(1-\lambda)]$. In either case, we have $e_{n+1} \leq \max[e_0, b/(1-\lambda)]$. The result then follows. □

Lemma 11. *Let $\{W_n\}_{n=0}^{\infty}$ be a sequence of non-negative random variables. If $\sup_n \mathbf{E}(W_n) < \infty$, then $\{W_n\}$ is bounded in probability.*

Proof. Since $\sup_n \mathbf{E}(W_n) < \infty$, let $K = \sup_n \mathbf{E}(W_n)$. Given $\epsilon > 0$, there exist $M = K/\epsilon$, such that, by Markov inequality, $\mathbf{P}[W_n > M] \leq \mathbf{P}[W_n \geq M] \leq K/M = \epsilon \Rightarrow \sup_n \mathbf{P}[W_n > M] \leq \epsilon$. Thus, $\{W_n\}$ is bounded in probability. □

2.2.4 Laws of large numbers

The sequences of random variables $X_1, X_2, ..., X_n$ generated by adaptive MCMC are usually combined together to form averages of the form $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$ to estimate the mean $\pi(g) = \int g(x)\pi(dx)$ of a function $g : \mathcal{X} \to R$. To justify such approximations, we require laws of large numbers for ergodic averages if the form:

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

either in probability or almost surely, for suitably regular functions g.

In this section, the laws of large numbers which hold for adaptive MCMC under weaker assumptions are considered.

Theorem 2.4 (Weak Law of Large Number)**.** *Consider an adaptive MCMC algorithm. Suppose ethat conditions* (*a*) *and* (*b*) *of Theorem 5 hold. Let* $g : \mathcal{X} \to \mathbf{R}$ *be a bounded measurable function. Then for any starting values* $x \in \mathcal{X}$ *and* $\gamma \in \mathcal{Y}$*. Then for any starting values* $x \in \mathcal{X}$ *and* $\gamma \in \mathcal{Y}$*, conditional on* $X_0 = x$ *and* $\Gamma_0 = \gamma$ *we have*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g) \tag{2.7}$$

in probability as $n \to \infty$*.*

Proof. Assume without loss of generality that $\pi(g) = 0$. Let $a = \sup_{x\in\mathcal{X}}|g(x)| < \infty$. Denote $\mathbf{E}_{\gamma,x}$ for expectations with respect to the Markov chain kernel P_γ when started from $X_0 = x$, and write $\mathbf{P}_{\gamma,x}$ for the corresponding probabilities. Denote $\mathbf{E}$ and $\mathbf{P}$ for expectations and probabilities with respect to the adaptive chain.

The usual law of large numbers for Markov chains implies that for each fixed $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, $\lim_{n\to\infty} \mathbf{E}_{\gamma,x}|\frac{1}{n}\sum_{i=1}^{n} g(X_i)| \to \pi(g) = 0$. Condition (a) implies that this convergence can be bounded uniformly over choices of x and γ, i.e. given $\epsilon > 0$ we can find an integer N such that

$$\mathbf{E}_{\gamma,x}\left(|\frac{1}{n}\sum_{i=1}^{n} g(X_i)|\right) < \epsilon, \quad x \in \mathcal{X}, \gamma \in \mathcal{Y}.$$

In terms of this N, condition b is used to find $n* \in \mathbf{N}$ satisfying (2). The coupling argument in the proof of Theorem 5 then implies that on the event E(which has probability $\geq 1 - \epsilon$, for all $n \geq n_*$, the adaptive chain sequence $X_{n+1}, ..., X_{n+N}$ can be coupled with probability $\geq 1 - \epsilon$ with a corresponding sequence arising from the fixed Markov chain P_{Γ_n}. In other words, since $|g| \leq a$, we have

$$\left|\mathbf{E}\left(\frac{1}{N}\left|\sum_{i=n+1}^{n+N} g(X_i)\right|\Big|\mathcal{G}_n\right) - \mathbf{E}_{\Gamma_n,X_n}\left(\frac{1}{N}\left|\sum_{i=n+1}^{n+N} g(X_i)\right|\right)\right|$$
$$\leq a\epsilon + a\mathbf{P}(E^c) \leq 2a\epsilon$$

The first term $a\epsilon$ comes to the fact the if the sequences X_n and X'_n are not coupled (with probability $\leq \epsilon$), the differences is less than $a\epsilon$ because g is bounded by a. The second term $a\mathbf{P}(E^c)$ means that if it is not on event E(with probability $< \epsilon$), the difference is less than $a\mathbf{P}(E^c)$ because g is bounded by a.

By triangle inequality, we have

$$\mathbf{E}\left(\frac{1}{N}\left|\sum_{i=n+1}^{n+N} g(X_i)\right|\Big|\mathcal{G}_n\right) < (1 + 2a)\epsilon \tag{2.8}$$

Now consider any integer T sufficiently large that

$$\max\left[\frac{an^*}{T}, \frac{aN}{T}\right] \le \epsilon \tag{2.9}$$

Then ($\lfloor r \rfloor$ denotes the greatest integer not exceeding r) we have:

$$\begin{aligned}\Big|\frac{1}{T}\sum_{i=1}^{T} g(X_i)\Big| &\le \Big|\frac{1}{T}\sum_{i=1}^{n^*} g(X_i)\Big| + \Bigg|\frac{1}{\lfloor\frac{T-n^*}{N}\rfloor}\sum_{j=1}^{\lfloor\frac{T-n^*}{N}\rfloor}\frac{1}{N}\sum_{k=1}^{N} g(X_{N(j-1)+k+n^*})\Bigg| \\ &\quad + \Big|\frac{1}{T}\sum_{n^*+\lfloor\frac{T-n^*}{N}\rfloor N+1}^{T} g(X_i)\Big| \\ &\le \Big|\frac{1}{T}\sum_{i=1}^{n^*} g(X_i)\Big| + \frac{1}{\lfloor\frac{T-n^*}{N}\rfloor}\sum_{j=1}^{\lfloor\frac{T-n^*}{N}\rfloor}\Bigg|\frac{1}{N}\sum_{k=1}^{N} g(X_{N(j-1)+k+n^*})\Bigg| \\ &\quad + \Big|\frac{1}{T}\sum_{n^*+\lfloor\frac{T-n^*}{N}\rfloor N+1}^{T} g(X_i)\Big|\end{aligned} \tag{2.10}$$

By 2.9, the first and last terms on the right-hand side of 2.10 are each $\le \epsilon$. By 2.8, the middle term is an average of terms each of which has absolute expectation $\le (1+2a)\epsilon$, thus we have that

$$\mathbf{E}\left(\Big|\frac{1}{T}\sum_{i=1}^{T} g(X_i)\Big|\right) \le \epsilon + (1+2a)\epsilon + \epsilon = \epsilon(3+2a).$$

Markov's inequality then gives that

$$\mathbf{P}\left(\Big|\frac{1}{T}\sum_{i=1}^{T} g(X_i)\Big| \ge \epsilon^{1/2}\right) \le \epsilon^{1/2}(3+2a).$$

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

Chapter 3

Stability of Adversarial Markov Chains

[1]

3.1 Set up and assumptions

Let $\mathcal{X}$ be a nonempty general state space, on which is defined a metric η, giving rise to a corresponding Borel σ algebra $\mathcal{F}$. Assume that $\mathcal{X}$ contains some specified "origin" point $\mathbf{0} \in \mathcal{X}$. Let P be the transition probability kernel for a fixed time-homogeneous Markov chain on $\mathcal{X}$. Assume that P is Harris ergodic with stationary probability distribution π, so that

$$\lim_{n \to \infty} \|P^n(x, \cdot) - \pi\| = 0, \quad x \in \mathcal{X} \tag{3.1}$$

To relate the Markov chain to the geometry of $\mathcal{X}$, assume that there is a constant $D < \infty$ such that P never moves more than a distance D, that is such that

$$P(x, \{y \in \mathcal{X} : \eta(x, y) \leq D\}) = 1, \quad x \in \mathcal{X} \tag{3.2}$$

Let $K \in \mathcal{F}$ be a fixed bounded nonempty subset of $\mathcal{X}$, and for $r > 0$ let K_r be the set of all states within a distance r of K, so that each K_r is also bounded.

Based on the above assumptions, the "adversarial Markov chain" $\{X_n\}$ is defined as follows. It begins with $X_0 = x_0$ for some specific initial state x_0; For simplicity, assume that $x_0 \in K$. When ever the process is outside of K, it moves according to the Markov transition probabilities P, that is,

$$\mathbf{P}(X_{n+1} \in A | X_0, X_1, ..., X_n) = \mathbf{P}(X_n, A), \quad n \leq 0, A \in \mathcal{F}, X_n \notin K. \tag{3.3}$$

When the process is inside of K, it can move arbitrarily, according to an adversary's wishes, depending on the time n, or the chain's history in a nonanticipatory manner (i.e., adapted to $\{X_n\}$), subject only to measurability [$i.e.\mathbf{P}(X_{n+1} \in A | X_0, X_1, ..., X_n)$ must be well defined for all $n \geq 0$ and $A \in \mathcal{F}$], and to the restriction that it can't move more than a distance D at each iteration - or more specifically that from K, it can only move to points within K_D. In summary, $\{X_n\}$ is a stochastic process which is "mostly" a Markov chain following the transition probabilities P, except that it is modified by an adversary when it is within the bounded subset K.

In this paper tries to find conditions that guarantees that such process $\{X_n\}$ will be bounded in probability, that is, will satisfy that

$$\lim_{L \to \infty} \sup_{n \in \mathbf{N}} \mathbf{P}(\eta(X_n, \mathbf{0}) > L | X_0 = x_0) = 0 \tag{3.4}$$

3.1.1 Results

Consider two new assumptions. The first provides an upper bound on the Markov vhain transitions out of K_D:

(A1) There is $M < \infty$, and a probability measure μ_* concenetrated on $K_{2D} \backslash K_D$, such that $P(x, dz) \leq M\mu_*(dz)$ for all $x \in K_D \setminus K$ and $z \in K_{2D} \setminus K_D$

The second assumption bounds an expected hitting time:

(A2) The expected time for a Markov chain following the transitions P to reach the subset K_D, when started from the distribution μ_* in (A1), is finite.

In terms of the above two assumptions, the following theorem is presented:

Theorem 3.1. *In the set up of 3.1, if (A1) and (A2) hold for the same μ_*, then (3.4) holds, that is, $\{X_n\}$ is bounded in probability.*

To prove Theorem 3.1, we begin by letting $\{Y_n\}$ be a "cemetery process" which begins in the distribution μ_* at time 0, and then follows the fixed transition kernel P, and then dies as soon as it hits K_D. Assumption (A2) then says that this cemetery process $\{Y_n\}$ has finite expected lifetime. For $L > l_0 := \sup\{\eta(x, \mathbf{0}) : x \in K_D\}$, let $B_L = \{x \in \mathcal{X} : \eta(x, \mathbf{0}) \geq L\}$, and let N_L denote the cemetery process's total occupation time of B_L (i.e., the number of iterations that $\{Y_n\}$ spends in B_L before it dies). Before proving Theorem 3.1, a Lemma is stated and proved.

Lemma 12. *Let $\{X_n\}$ be the adversarial process as defined previously. Then assuming* (A1)*, for any $n \in \mathbf{N}$, and any $L > l_0$, and any $x \in K$, we have*

$$\mathbf{P}(X_n \in B_L | X_0 = x) \leq M\mathbf{E}(N_L),$$

where N_L is the occupation time of B_L fot the cemetery process $\{Y_n\}$ defined above.

Proof. Let σ be the last return time of $\{X_n\}$ to K_D by time n, this exists because $X_0 \in K_D$. Let μ_k be the law of X_k when starting from $X_0 = x_0$. Then letting $I = K_D \setminus K$ ("inside") and $O = K_{2D} \setminus K_D$

("outside"), then

$$\begin{aligned}
&\mathbf{P}(X_n \in B_L | X_0 = x_0) \\
&\quad = \sum_{k=0}^{n-1} \mathbf{P}(X_n \in B_L, \sigma = k | X_0 = x_0) \\
&\quad = \sum_{k=0}^{n-1} \int_{y \in I} \int_{z \in O} \mathbf{P}(X_k = dy, X_{k+1} = dz, X_n \in B_L, \sigma = k | X_0 = x_0) \\
&\quad = \sum_{k=0}^{n-1} \int_{y \in I} \int_{z \in O} \mu_k(dy) P(y, dz) \times \mathbf{P}(X_n \in B_L, \sigma = k | X_0 = x_0, X_k = y, X_{k+1} = z) \\
&\quad = \sum_{k=0}^{n-1} \int_{y \in I} \int_{z \in O} \mu_k(dy) M\mu_*(dz) \times \mathbf{P}(X_n \in B_L, \sigma = k | X_0 = x_0, X_k = y, X_{k+1} = z) \\
&\quad \leq \sum_{k=0}^{n-1} \int_{y \in I} \int_{z \in O} \mu_k(dy) M\mu_*(dz) \mathbf{P}(Y_{n-k-1} \in B_L | Y_0 = z) \\
&\quad \leq M \sum_{k=0}^{n-1} \int_{z \in O} \mathbf{P}(Y_{n-k-1} \in B_L | Y_0 = z) \mu_*(dz) \\
&\quad \leq M \sum_{j=0}^{\infty} \int_{z \in O} \mathbf{P}(Y_j \in B_L | Y_0 = z) \mu_*(dz) \\
&\quad = M \sum_{j=0}^{\infty} \int_{z \in O} \mathbf{E}[\mathbf{1}_{\{Y_j \in B_L\}} | Y_0 = z] \mu_*(dz) = M \int_{z \in O} \mathbf{E}[\sum_{j=0}^{\infty} \mathbf{1}_{\{Y_j \in B_L\}} | Y_0 = z] \mu_*(dz) \\
&\quad = M \int_{z \in O} \mathbf{E}[N_L | Y = z] \mu_*(dz) = M\mathbf{E}[N_L]
\end{aligned}$$

The result then follows. □

We now use this lemma to prove Theorem 3.1.

Proof. For each $A \in \mathcal{F}$, let $\nu(A)$ be the above cemetery process's expected occupation measure, which is the expected number of iterations that the cemetary process $\{Y_n\}$ spends in the subset A. Then the total measure $\nu(\mathcal{X})$ is the expected liftetime of the cemetery process, and is thus finite by (A2). Hence, by the continuity of measures,

$$\lim_{L \to \infty} \nu(B_L) = \nu(\cap B_L) = \nu(\varnothing) = 0.$$

This shows that $\mathbf{E}(N_L) \to 0$ as $L \to \infty$. Hence, by Lemma 12,

$$\lim_{L \to \infty} \sup_{n \in \mathbf{N}} \mathbf{P}(X_n \in B_L | X_0 = x_0) \leq M \lim_{L \to \infty} \mathbf{E}(N_L) = 0,$$

so $\{X_n\}$ is bounded in probability. □

Consider another assumption:

(A3) The set $K_{2D} \setminus K_D$ is small for P; that is, there is some probability measure ν_* on $\mathcal{X}$, and some $\epsilon > 0$, and some $n_0 \in \mathbf{N}$, such that $p^{n_0}(x, A) \geq \epsilon \nu_*(A)$ for all states $x \in K_{2D} \setminus K_D$ and all subsets $A \in \mathcal{F}$.

The another theorem is as follows:

Theorem 3.2. *In the setup of Section 2, if* (A1) *and* (A3) *hold where either* (a)$\nu_* = \mu_*$ *or* (b) P *is reversible and* $\mu_* = \pi|_{K_{2D}\setminus K_D}$, *then 3.4 holds; that is,* $\{X_n\}$ *is bounded in probability.*

Before prove this theorem, we first prove two additional probability lemmas:

Lemma 13. *Consider a Markov chain with stationary probability distribution* π, *and let* $A \in \mathcal{F}$ *with* $\pi(A) > 0$. *Then:*

1. $\mathbf{E}_{\pi|A}(\tau_A) = 1/\pi(A) < \infty$, *where* τ_A *is the first return time to* A.

2. *For all* $k \in \mathbf{N}$, $\mathbf{E}_{\pi|A}(\tau_A^k) = k/\pi(A) < \infty$, *where* $\tau_A^{(k)}$ *is the* k^{th} *return time to* A.

Proof. For Part 1, using Theorem 10.0.1 of [2] with $B = \mathcal{X}$, we obtain

$$1 = \pi(\mathcal{X}) = \int_{x\in A} \pi(dx)\mathbf{E}_x\Big[\sum_{n=1}^{\tau_A} \mathbf{1}_{X_n \in \mathcal{X}}\Big] = \int_{x\in A} \pi(dx)\mathbf{E}_x[\tau_A] = \pi(A)\mathbf{E}_{\pi|A}[\tau_A],$$

giving the result.

For Part 2. Expand the original Markov chain to a new Markov chain on $\mathcal{X} \times \{0, 1, ..., k-1\}$, where the first variable is the original chain, and the second variable is the count (mod k) of the number of times the chain has returned to A. So each time the original chain visits A, the second variable increases by 1(mod k). Then the expanded chain has stationary distribution $\pi \times Uniform\{0, 1, ..., k-1\}$. Hence, by part 1, if we begin in $(\pi|A) \times \delta_0$, then the expected return time of the expanded chain to $A \times \{0\}$ equals $1/[\pi(A) \times (1/k)] = k/\pi(A)$. But the first return time of the expanded chain to $A \times \{0\}$ corresponds precisely to the k^{th} return time of the original chain to A. □

Lemma 14. *Let* $\{W_n\}$ *be a sequence of nonnegative random variables each with finite mean* $m < \infty$, *and let* $\{I_n\}$ *be a sequence of indicator variables each with* $\mathbf{P}(I_n = 1) = p > 0$. *Assume that the sequence of pairs* $\{(W_n, I_n)\}$ *is i.i.d. [i.e., the sequence* $\{Z_n\}$ *is i.i.d where* $Z_n = (W_n, I_n)$*]. Let* $\tau = \inf\{n : I_n = 1\}$, *and let* $S = \sum_{i=1}^{\tau} W_i$. *Then* $\mathbf{E}(S) = \frac{m}{p} < \infty$.

Proof. We can write $S = \sum_{i=1}^{\infty} W_i \mathbf{1}_{\tau \geq i}$. Now, the event $\{\tau \geq i\}$ is equivalent to the event that $I_1 = I_2 = \cdots = I_{i-1} = 0$. Hence it is contained in $\sigma(Z_1, ..., Z_{i-1})$ and is thus independent of W_i by assumption.

Also, τ is distributed as Geometric(p) and hence has mean $\frac{1}{p}$. We then compute that

$$\begin{aligned}\mathbf{E}(S) = \mathbf{E}\left(\sum_{i}^{\infty} W_i \mathbf{1}_{\tau\geq i}\right) &= \sum_{i=1}^{\infty} \mathbf{E}(W_i \mathbf{1}_{\tau \geq i}) \\ &= \sum_{i=1}^{\infty} \mathbf{E}(W_i)\mathbf{E}\mathbf{1}_{\tau\geq i} = \sum_{i=1}^{\infty} m\mathbf{P}(\tau \geq i) = m\mathbf{E}(\tau) = m/p,\end{aligned}$$

The result then as follows. □

Now consider two more Lemmas which helps us prove Theorem.

Lemma 15. *Consider a ϕ-irreducible Markov chain on a state space $(\mathcal{X}, \mathcal{F})$ with transitional kernel P and stationary probability distribution π. Let $B, C \in \mathcal{F}$. Suppose C is a small set for P with minorizing measure μ; that is, there is $\epsilon > 0$ and $n_0 \in \mathbf{N}$ such that $P^{n_0}(x, A) \geq \epsilon\mu(A)$ for all states $x \in C$ and all subsets $A \in \mathcal{F}$. Let τ_B be the first hitting time of B. Then $\mathbf{E}_\mu(\tau_B) < \infty$.*

Proof. It suffices to consider the case where $n_0 = 1$, since if not we can replace P by P^{n_0} and note that the hitting time of B by P is at most n_0 times the hitting time of B by P^{n_0}.

The proof uses the Nummelin [2] splitting technique. Consider the Markov chain on state space $\mathcal{X} \times \{0, 1\}$, where the second variable is an indicator of whether or not to regenerate according to μ.

Let $\alpha = \mathcal{X} \times \{1\}$. Then α is a Markov chain atom (i.e., the chain has identical transition probabilities from every state in α), and it has stationary measure $\pi(\alpha) = \epsilon\pi(C) > 0$. So, starting in μ^*(corresponding to the original chain starting in μ). If the chain arrives in α, then it will return to α in finite expected time $1/\pi(\alpha) < \infty$ by Lemma 13.

Now, let W_n be the number of iterations between the $(n-1)^{th}$ and n^{th} returns to α, and let $I_n = 1$ if the chain visits B during the $(n-1)^{th}$ and n^{th} visit to *alpha* , otherwise $I_n = 0$. Then $\mathbf{P}[I_n = 1] > 0$ by the ϕ-irreducibility of P. Hence, $\{W_n, I_n\}$ satisfies the conditions of Lemma 14.

Therefore, by Lemma 14, the expected number of iterations until we complete a tour which includes a visit to B is finite. Hence, the expected hitting time of B is finite. □

Lemma 16. *Let P be a Markov chain transition kernel on $(\mathcal{X}, \mathcal{F})$, with stationary probability measure π. Let $C \in \mathcal{F}$ such that $\pi(C) > 0$. Assume that C is a small set for P; that is for some $n_0 \in \mathbf{N}$ and $\beta > 0$ and probability measure ν,*

$$P^{n_0}(x, A) \geq \beta\nu(A), \quad A \in \mathcal{F}, x \in C. \tag{3.5}$$

Then,

$$P^{n_0}(P^*)^{n_0} \geq \frac{1}{4}\beta^2\pi(A \cap C), \quad A \in \mathcal{F}, x \in C \tag{3.6}$$

where P^* is the $L^2(\pi)$ adjoint of P. In particular, if P is reversible with respect to π, so that $P^* = P$, then

$$P^{2n_0}(x, A) \geq \frac{1}{4}\beta^2\pi(A \cap C), A \in \mathcal{F}, x \in C.$$

Proof.

$$D(\epsilon) := \Big\{x \in \mathcal{X} : \frac{d\nu}{d\pi}(x) > \epsilon\Big\} \tag{3.7}$$

By replacing P by P^{n_0} and P^* by $(P^*)^{n_0}$. It suffices to assume that $n_0 = 1$. Now, the Radon-Nikodym derivative $\frac{d\nu}{d\pi}$ of ν with respect to π satisfies that $\int_{\mathcal{X}} \frac{d\nu}{d\pi}(x)\pi(dx) = \nu(\mathcal{X}) = 1$. For every $\epsilon \in [0, 1]$, let Since ν is absolutely continuous with respect to ν, $\nu(A) = 0$ whenever $\pi(A) = 0$. If $\pi(D(\epsilon)) = 0$, then $\nu(D(\epsilon)) = \int_{D(\epsilon)} \epsilon\pi(dx) > \epsilon\pi(D(\epsilon)) = 0$. Contradiction happens. So $\pi(D(\epsilon)) > 0$. Then compute

$$\nu(D(\epsilon)^c) = \int_{D(\epsilon)^c} \frac{\nu}{\pi}(x)\pi(dx) \leq \epsilon \int_{\mathcal{X}} \pi(dx) = \epsilon$$

hence

$$\nu(D(\epsilon)) \geq 1 - \epsilon. \tag{3.8}$$

The adjoint P^* satisfies

$$\pi(dx)P(x, dy) = \pi(dy)P^*(y, dx). \tag{3.9}$$

Now let $x \in C$, and $A \in \mathcal{F}$ with $A \cap C \neq \varnothing$. Using first 3.5 and then 3.7,

$$\begin{aligned} PP^*(x, A) = \int_{z\in\mathcal{X}} P(x, dz)P^*(z, A) &\geq \beta \int_{z\in\mathcal{X}} P^*(z, A \cap C)\nu(dz) \\ &\geq \beta \int_{z\in D(\epsilon)} \int_{y\in A\cap C} P^*(z, dy)\epsilon\pi(dz) \end{aligned}$$

To continue, use 3.9, then 3.5 again and finally 3.8 to obtain

$$\begin{aligned} PP^*(x, A) &\geq \beta\epsilon \int_{z\in D(\epsilon)} \int_{y\in A\cap C} \pi(dy)P(y, dz) \\ &\geq \beta^2\epsilon\nu(D(\epsilon))\pi(A \cap C) \geq \beta^2\epsilon(1 - \epsilon)\pi(A \cap C) \end{aligned}$$

.

Setting $\epsilon = \frac{1}{2}$, we have

$$PP^*(x, A) \geq \beta * \pi(A \cap C)/\pi(C).$$

where $\beta* = \frac{1}{4}\beta^2\pi(C)$. □

Then, we have following corollary.

Corollary 7. (A3) with $\nu_* = \mu_*$ implies (A2).

Proof. This follows immediately by applying Lemma 15 with $C = K_{2D} \setminus K_D$, and $B = K_D$, and $\mu = \mu_* = \nu_*$. □

Proof of Theorem 5

Proof. Under assumption (a), the result follows by combining Corollary 7 and Theorem 3.1. Under assumption (b) such that P is reversible and $\mu_* = \pi_{K_{2D} \setminus K_D}$. It follow from Lemma 16 that (A3) also holds with $\nu_* = \pi_{K_{2D} \setminus K_D} = \mu_*$. Hence assumption (a) still applies, so 3.4 again follows. □

Bibliography

[1] Radu Craiu, Lawrence Gray, Krzysztof Latuszynski, Neal Madras, Gareth O. Roberts, and Jeffrey Rosenthal. Stability of adversarial markov chains, with an application to adaptive mcmc algorithms. *The Annals of Applied Probability*, 25, 03 2014.

[2] Sean Meyn and R L. Tweedie. *Markov Chains and Stochastic Stability.* 01 2009.

[3] Gareth O. Roberts and Jeffrey Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1, 04 2004.

[4] Gareth O. Roberts and S Rosenthal. Coupling and ergodicity of adaptive mcmc. 04 2019.

[5] Gareth O. Roberts and Jerey S. Rosenthal. Small and pseudo-small sets for markov chains. *Stochastic Models*, 17, 01 2001.

[6] Jeffrey Seth Rosenthal. *A First Look at Rigorous Probability Theory.* 01 2006.