Markov Chains Convergence Theorems

Supervised Project

University of Toronto

Spring 2006

Supervisor: Professor Jeffrey S. Rosenthal [1]

Student: Olga Chilina [2]

[1]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S3G3. Email: jeff@math.toronto.edy. Web: http://probability.ca/jeff/

[2]Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S3G3. Email: olgac@utstat.toronto.edu

Contents

1 Introduction

In this project we shall discuss conditions on the convergence of Markov chains and present some convergence rate results. We shall introduce the coupling construction and use it to prove convergence theorems for Markov chains. Mostly this project is based on the article *General State Space Markov Chains and MCMC Algorithms* by Gareth O. Roberts and Jeffrey S. Rosenthal (see reference [1]). We'll discuss conditions on the convergence of Markov chains, and consider the proofs of convergence theorems in details. We will modify some of the proofs, and try to improve some parts of them. But in total we'll be repeating the main ideas from the indicated above article.

First of all, let us give a general idea about Markov chains and introduce a few notations that we shall use from now on. For more detailed definitions see [3] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in practice.*

Suppose we generate a sequence of random variables, $\{X_0, X_1, X_2, ...\}$, such that at each time $n \geq 0$, the next state X_{n+1} is sampled from a distribution $P(X_{n+1}|X_n)$ which depends only on the current state of the chain, $\{X_n\}$. In other words, given X_n, the next state X_{n+1} doesn't depend on the other states of the chain $\{X_0, X_1, ..., X_{n-1}\}$. This sequence is called a *Markov chain,* and $P(.|.)$ is called the *transition kernel* of the chain. So we can define:

Definition. A *Markov chain* is characterized by three ingredients: a *state space* $\mathcal{X}$, an *initial distribution,* and *transition kernel.* The transition kernel is a function $P(x, A)$ that takes values between 0 and 1, and such that for any $n \geq 0$

$$P\{X_{n+1} \in A | X_n = x\} = P(x, A)$$

for all $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$. That is, $P(x, \cdot)$ is the distribution of the Markov chain after one step given that it starts at x.

For a probability distribution ν on $\mathcal{X}$, we define $\forall A \subseteq \mathcal{X}$

$$(\nu P)(A) = \int P(x, A)\nu(dx)$$

to be the distribution of the position of a Markov chain with transition kernel P and initial distribution ν after one step. A Markov chain has a *stationary* distribution π if

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

Note: $P(x, dy)$ is the probability of moving to a small measurable subset $dy \in \mathcal{X}$ given that the move starts at x.

Also for a real-valued function h on $\mathcal{X}$ define

$$(Ph)(x) = \int P(x, dy)h(y) = E[h(X_1)|X_0 = x].$$

The product of two transition kernels P and Q is the transition kernel defined by

$$(PQ)(x, A) = \int P(x, dy)Q(y, A)$$

for all $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$. The n^{th} iterate $P^n = PP^{n-1}$ for $n \geq 2$, and we say that P^0 is the identity kernel that puts probability one on the staying at the initial value.

Using this notations, we can write $P\{X_n \in A|X_0 = x\} = P^n(x, A)$ for any $n \geq 0$.

The first return time of a Markov chain to a set $A \subseteq \mathcal{X}$ we denote by τ_A, that is,

$$\tau_A = \inf\{n \geq 1 : X_n \in A\},$$

and $\tau_A = \infty$ if the chain never returns to A.

The *indicator function* of a set C is

$$I_C(x) = \begin{cases} 1 & \text{if } x \in C \\ 0 & \text{if } x \notin C \end{cases}$$

For a probability distribution ν on $\mathcal{X}$ a statement holds for 'ν-a.e. x' if ν gives probability zero to the set of points in $\mathcal{X}$ where the statement fails.

Consider one more useful notation. Let $P_{ij}(n) = P[X_n = j|X_0 = i]$. Then for a countable state space $\mathcal{X}$ we can define:

Definition. A Markov chain $\{X_n\}$ is called *irreducible* if for all i, j, there exists an $n > 0$ such that $P_{ij}(n) > 0$.

Let $\tau_{ii} = min\{n > 0 : X_n = i|X_0 = i\}$. Then we say that an irreducible chain $\{X_n\}$ is *recurrent* if $P[\tau_{ii} < \infty] = 1$ for some (and hence for all) i. Otherwise, $\{X_n\}$ is *transient*. Another equivalent condition for recurrence is

$$\sum_n P_{ij}(n) = \infty$$

for all i, j.

An irreducible recurrent chain $\{X_n\}$ is called *positive recurrent* if $E[\tau_{ii}] < \infty$ for some (and hence for all) i. Otherwise, it is called *null-recurrent*. The equivalent condition for positive recurrence is the existence of a stationary distribution for the Markov chain, that is there exists $\pi(\cdot)$ such that

$$\sum_i \pi(i)P_{ij}(n) = \pi(j)$$

for all j and $n \geq 0$.

An irreducible chain $\{X_n\}$ is called *aperiodic* if for some (hence for all) i, $gcd\{n > 0 : P_{ii}(n) > 0\} = 1$.

Note that the above definitions are for a discrete state space. We shall consider a general state space, so we'll give corresponding definitions for this case in the next following sections.

2 Total Variation Norm

Suppose we have a Markov chain $\{X_n\}$ with n-step transition law, i.e.

$$P^n(x, A) = P[X_n \in A | X_0 = x], \ \forall A \subseteq \mathcal{X}.$$

We want to know how close $P^n(x, A)$ is to the stationary distribution $\pi(A)$ for large n, and how large this n should be, if we want $P^n(x, A)$ to be very close to $\pi(A)$. To answer this question we define *total variation norm.*

Definition. The *total variation norm* between two probabilitiry measures $\mu_1(\cdot)$ and $\mu_2(\cdot)$ is

$$||\mu_1(\cdot) - \mu_2(\cdot)|| = \sup_A |\mu_1(A) - \mu_2(A)| \quad (1)$$

Now we can restate our question in terms of the total variation norm: Is $\lim_{n\to\infty} ||P^n(x, \cdot) - \pi(\cdot)|| = 0$? To proceed let's discuss some properties of the total variation norm. But before we shall give a few useful definitions from measure theory.

Definition. A measure space $(\Omega, \mathcal{B}, \rho)$ is a *finite* measure space if $\rho(\Omega) < \infty$; it is σ - *finite* if the total space is the union of countable family of sets of finite measure, i.e. there is a countable set $\mathcal{F} \subset \mathcal{B}$ such that $\rho(A) < \infty$ $\forall A \in \mathcal{F}$, and $\Omega = \bigcup_{A \subset \mathcal{F}} A$. In this case we also say that ρ is a σ - *finite.* Any finite measure space is σ - finite. But the converse is not always true, e.g. Lebesgue measure λ in R^n: it is σ-finite, but not finite, in fact, $R^n = \bigcup_{k \in N} [-k, k]^n$, but $\lambda(R^n) = \infty$.

Definition. A measure μ is *absolutely continuous* w.r.t. measure ρ ($\mu \ll \rho$) if $\mu(E) = 0$ $\forall E$ such that $\rho(E) = 0$.

By the Radon-Nikodym Theorem, this is equivalent to saying that

$$\mu(E) = \int_E f d\rho,$$

for some integrable function f. The function f is like a derivative, and is called the *Radon-Nikodym derivative,* denoted by $\frac{d\mu}{d\rho}$.

The following proposition was proven in [1] for a particular case. We shall prove it in general:

Proposition 2.1. $||\mu_1(\cdot) - \mu_2(\cdot)|| = \frac{1}{b-a} \sup_{f:\mathcal{X}\to[a,b]} |\int f d\mu_1 - \int f d\mu_2|$ $\forall a < b$

Proof: Let $\rho = \mu_1 + \mu_2$. Then $\mu_1 \ll \rho$ and $\mu_2 \ll \rho$. Define functions $g = \frac{d\mu_1}{d\rho}$ and $h = \frac{d\mu_2}{d\rho}$. Then for any $f : \mathcal{X} \to [a, b]$ we have

$$\left| \int f d\mu_1 - \int f d\mu_2 \right| = \left| \int f(g-h)d\rho \right| = \left| \int\limits_{\{g>h\}} f(g-h)d\rho + \int\limits_{\{g<h\}} f(g-h)d\rho \right|$$

Since $g - h \geq 0$ on the set $\{g > h\}$, then from $a \leq f(x) \leq b$ it follows that

$$a \int\limits_{\{g>h\}} (g-h)d\rho \leq \int\limits_{\{g>h\}} f(g-h)d\rho \leq b \int\limits_{\{g>h\}} (g-h)d\rho \tag{2}$$

Since $g - h \leq 0$ on the set $\{g < h\}$, then

$$b \int\limits_{\{g<h\}} (g-h)d\rho \leq \int\limits_{\{g<h\}} f(g-h)d\rho \leq a \int\limits_{\{g<h\}} (g-h)d\rho \tag{3}$$

From (2) and (3), using the equality $\int\limits_{\mathcal{X}} (g-h)d\rho = \mu_1(\mathcal{X}) - \mu_2(\mathcal{X}) = 1 - 1 = 0$, we have that

$$\begin{aligned}
(b-a) \int\limits_{\{g<h\}} (g-h)d\rho + a \int\limits_{\mathcal{X}} (g-h)d\rho &= b \int\limits_{\{g<h\}} (g-h)d\rho + a \int\limits_{\{g>h\}} (g-h)d\rho \\
&\leq \int\limits_{\mathcal{X}} f(g-h)d\rho \\
&\leq b \int\limits_{\{g>h\}} (g-h)d\rho + a \int\limits_{\{g<h\}} (g-h)d\rho \\
&= (b-a) \int\limits_{\{g>h\}} (g-h)d\rho + a \int\limits_{\mathcal{X}} (g-h)d\rho \\
&= (b-a) \int\limits_{\{g>h\}} (g-h)d\rho
\end{aligned}$$

So,

$$(b-a) \int\limits_{\{g<h\}} (g-h)d\rho \leq \int\limits_{\mathcal{X}} f(g-h)d\rho \leq (b-a) \int\limits_{\{g>h\}} (g-h)d\rho$$

Thus,

$$\sup_{f:\mathcal{X}\to[a,b]}\left|\int\limits_{\mathcal{X}} f(g-h)d\rho\right| = \max\left\{(b-a)\int\limits_{\{g>h\}}(g-h)d\rho, (b-a)\left|\int\limits_{\{g<h\}}(g-h)d\rho\right|\right\}$$

$$\text{so, } \frac{1}{b-a}\sup_{f:\mathcal{X}\to[a,b]}\left|\int\limits_{\mathcal{X}} f(g-h)d\rho\right| = \max\left\{\int\limits_{\{g>h\}}(g-h)d\rho, \left|\int\limits_{\{g<h\}}(g-h)d\rho\right|\right\} \quad (4)$$

Now, $\forall A \subseteq \mathcal{X}$ we have

$$||\mu_1(A)-\mu_2(A)|| = \left|\int\limits_{A}(g-h)d\rho\right| = \left|\int\limits_{A\bigcap\{g>h\}}(g-h)d\rho + \int\limits_{A\bigcap\{g<h\}}(g-h)d\rho\right|$$

Hence, the biggest value is obtained if either $A=\{g>h\}$, or $A=\{g<h\}$. So,

$$\begin{aligned}
||\mu_1(\cdot)-\mu_2(\cdot)|| &= \max\left\{\int\limits_{\{g>h\}}(g-h)d\rho, \left|\int\limits_{\{g<h\}}(g-h)d\rho\right|\right\} \\
&= \frac{1}{b-a}\sup_{f:\mathcal{X}\to[a,b]}\left|\int\limits_{\mathcal{X}} f(g-h)d\rho\right| \text{ (see (4))}
\end{aligned}$$

□

So, in particular, we have $||\mu_1(\cdot)-\mu_2(\cdot)|| = \sup_{f:\mathcal{X}\to[0,1]}|\int fd\mu_1 - \int fd\mu_2|$ and $||\mu_1(\cdot)-\mu_2(\cdot)|| = \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}|\int fd\mu_1 - \int fd\mu_2|$

The next few propositions are not changed much, but we added several details, so to make proofs more clear.

Proposition 2.2. If $\pi(\cdot)$ is stationary for our MC, then

$$||P^{n+1}(x,\cdot)-\pi(\cdot)|| \le ||P^n(x,\cdot)-\pi(\cdot)||,$$

$n \in N$, i.e. $||P^n(x,\cdot)-\pi(\cdot)||$ is non-decreasing in n.

Proof:

$$\begin{aligned}
\left|P^{n+1}(x,A)-\pi(A)\right| &= \left|\int\limits_{y\in\mathcal{X}} P^n(x,dy)P(y,A) - \int\limits_{y\in\mathcal{X}}\pi(dy)P(y,A)\right| \\
&= \left|\int\limits_{y\in\mathcal{X}} P^n(x,dy)f(y) - \int\limits_{y\in\mathcal{X}}\pi(dy)f(y)\right| \text{ (here } f(y)=P(y,A)) \\
&\le ||P^n(x,\cdot)-\pi(\cdot)|| \text{(by Proposition 2.1 for a=0, b=1)}
\end{aligned}$$

$\square$

Proposition 2.3. Let $t(n) = 2\sup_{x\in\mathcal{X}} ||P^n(x,\cdot) - \pi(\cdot)||$, where $\pi(\cdot)$ is a stationary distribution. Then t is *submultiplicative*, i.e.

$$t(n+m) \le t(n)t(m)$$

for $n, m \in N$.

Proof: Let $P^*(x,\cdot) = P^n(x,\cdot) - \pi(\cdot)$ and $Q^*(x,\cdot) = P^m(x,\cdot) - \pi(\cdot)$. Then we'll have that

$$\begin{aligned}
(P^*Q^*f)(x) &\equiv \int\limits_{y\in\mathcal{X}} f(y) \int\limits_{z\in\mathcal{X}} [P^n(x,dz) - \pi(dz)][P^m(z,dy) - \pi(dy)] \\
&= \int\limits_{y\in\mathcal{X}} f(y)\Big[\int\limits_{z\in\mathcal{X}} P^n(x,dz)P^m(z,dy) - \int\limits_{z\in\mathcal{X}} P^n(x,dz)\pi(dy) \\
&- \int\limits_{z\in\mathcal{X}} P^m(z,dy)\pi(dz) + \int\limits_{z\in\mathcal{X}} \pi(dy)\pi(dz)\Big] \\
&= \int\limits_{y\in\mathcal{X}} f(y)[P^{n+m}(x,dy) - \pi(dy) - \pi(dy) + \pi(dy)] \text{ since } \pi \text{ is stationary} \\
&= \int\limits_{y\in\mathcal{X}} f(y)[P^{n+m}(x,dy) - \pi(dy)]
\end{aligned}$$

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

$$\begin{aligned}
\bar{g} &= \sup_{x\in\mathcal{X}} \Big|\int\limits_{y\in\mathcal{X}} (P^m(x,dy) - \pi(dy))f(y)\Big| \\
&\le \sup_{x\in\mathcal{X}} \Big[\sup_{f:\mathcal{X}\to[0,1]} \Big|\int\limits_{y\in\mathcal{X}} f dP^m - \int\limits_{y\in\mathcal{X}} f d\pi\Big|\Big] \\
&= \sup_{x\in\mathcal{X}} ||P^m(x,\cdot) - \pi(\cdot)|| \text{ by Proposition 2.1} \\
&= \frac{1}{2}t(m)
\end{aligned}$$

So, $2\bar{g} \le t(m)$. If $\bar{g} = 0$, then $\sup_{x\in\mathcal{X}} |(Q^*f)(x)| = 0 => P^*Q^*f = 0$. If $\bar{g} \ne 0$, then

$$\begin{aligned}
2\sup_{x\in\mathcal{X}} |(P^*Q^*f)(x)| &= 2\frac{\bar{g}}{\bar{g}} \sup_{x\in\mathcal{X}} |(P^*g)(x)| \\
&= 2\bar{g} \sup_{x\in\mathcal{X}} |(P^*\frac{g}{\bar{g}})(x)| \\
&\le t(m) \sup_{x\in\mathcal{X}} |(P^*\frac{g}{\bar{g}})(x)|
\end{aligned}$$

Since $-1 \leq \frac{g}{\bar{g}} \leq 1$, we have

$$\begin{aligned}
(P^*\frac{g}{\bar{g}})(x) &\equiv \int\limits_{y\in\mathcal{X}} P^*(x,dy)\frac{g}{\bar{g}}(y) \\
&= \int\limits_{y\in\mathcal{X}} P^n(x,dy)\frac{g}{\bar{g}}(y) - \int\limits_{y\in\mathcal{X}} \pi(dy)\frac{g}{\bar{g}}(y) \\
&\leq \sup_{\frac{g}{\bar{g}}:\mathcal{X}\to[-1,1]} \left| \int\limits_{y\in\mathcal{X}} \frac{g}{\bar{g}} dP^n - \int\limits_{y\in\mathcal{X}} \frac{g}{\bar{g}} d\pi \right| \\
&= 2||P^n(x,\cdot) - \pi(\cdot)|| \text{ by Proposition 2.1}
\end{aligned}$$

So, $\sup_{x\in\mathcal{X}}(P^*\frac{g}{\bar{g}})(x) \leq t(n)$. Hence

$$\begin{aligned}
& t(n+m) = 2\sup_{x\in\mathcal{X}} ||P^{n+m}(x,\cdot) - \pi(\cdot)|| \\
= \; & 2\sup_{x\in\mathcal{X}} \sup_{f:\mathcal{X}\to[0,1]} \left| \int f dP^{n+m} - \int f d\pi \right| \text{ (by Proposition 2.1)} \\
= \; & 2\sup_{x\in\mathcal{X}} \sup_{f:\mathcal{X}\to[0,1]} |(P^*Q^*f)(x)| = 2\bar{g}\sup_{x\in\mathcal{X}} \sup_{\frac{g}{\bar{g}}:\mathcal{X}\to[-1,1]} |(P^*\frac{g}{\bar{g}})(x)| \\
\leq \; & t(m)\sup_{x\in\mathcal{X}} \sup_{\frac{g}{\bar{g}}:\mathcal{X}\to[-1,1]} |(P^*\frac{g}{\bar{g}})(x)| \leq t(n)t(m)
\end{aligned}$$

□

Example. Let $\mathcal{X} = \{1,2\}$, $P(1,\{1\}) = 0.3$, $P(1,\{2\}) = 0.7$, $P(2,\{1\}) = 0.4$, $P(2,\{2\}) = 0.6$. Let $\pi(1) = \frac{4}{11}$, $\pi(2) = \frac{7}{11}$, then π is a stationary distribution. Indeed, $\sum_{i=1}^{2} P_{i1}\pi(i) = 0.3 \cdot \frac{4}{11} + 0.4 \cdot \frac{7}{11} = \frac{1.2+2.8}{11} = \frac{4}{11} = \pi(1)$, and $\sum_{i=1}^{2} P_{i2}\pi(i) = 0.7 \cdot \frac{4}{11} + 0.6 \cdot \frac{7}{11} = \frac{2.8+4.2}{11} = \frac{7}{11} = \pi(2)$.

Let's check if the above proposition is true, if we calculate

$$t(n) = \sup_{x\in\mathcal{X}} ||P^n(x,\cdot) - \pi(\cdot)||$$

without factor of 2. When $x = 1$ we have $|P(1,\{1\}) - \pi(\{1\})| = |0.3 - \frac{4}{11}| = |0.3 - 0.3636363| = 0.0636$, and $|P(1,\{2\}) - \pi(\{2\})| = |0.7 - \frac{7}{11}| = |0.7 - 0.6363636| = 0.0636$. When $x = 2$ we have $|P(2,\{1\}) - \pi(\{1\})| = 0.0363$, and $|P(2,\{2\}) - \pi(\{2\})| = 0.0363$. So, $\sup_{x\in\mathcal{X}} ||P(x,\cdot) - \pi(\cdot)|| = 0.0636$. Similarly, we can calculate that $\sup_{x\in\mathcal{X}} ||P^2(x,\cdot) - \pi(\cdot)|| = 0.00636$. So, $t(1+1) = t(2) = 0.00636 > 0.004045 = (0.0636)^2 = t(1)t(1)$. This example shows that we, indeed, need factor of 2 in the property we just proved.

Proposition 2.4. If $\mu_1(\cdot)$ and $\mu_2(\cdot)$ have densities g and h, respectively, w.r.t. some σ-finite measure $\rho(\cdot)$, and $M = max(g,h)$ and $m = min(g,h)$, then

$$||\mu_1(\cdot) - \mu_2(\cdot)|| = \frac{1}{2}\int_{\mathcal{X}} (M-m)d\rho = 1 - \int_{\mathcal{X}} m d\rho$$

Proof: In the proof of Proposition 2.1 let $a = -1$ and $b = 1$. Then

$$||\mu_1(\cdot) - \mu_2(\cdot)|| = \frac{1}{2}\left(\int_{g>h} (g-h)d\rho + \int_{g<h} (h-g)d\rho\right) = \frac{1}{2}\int_{\mathcal{X}} (M-m)d\rho$$

Now, since $M + m = g + h$, it follows that

$$\int_{\mathcal{X}} (M+m)d\rho = \int_{\mathcal{X}} g d\rho + \int_{\mathcal{X}} h d\rho = 1 + 1 = 2$$

So,

$$\begin{aligned}\frac{1}{2}\int_{\mathcal{X}} (M-m)d\rho &= 1 - \frac{1}{2}\left(2 - \int_{\mathcal{X}} (M-m)d\rho\right) \\ &= 1 - \frac{1}{2}\left(\int_{\mathcal{X}} [(M+m) - (M-m)]d\rho\right) = 1 - \int_{\mathcal{X}} m d\rho\end{aligned}$$

□

Proposition 2.5. For any probability measures $\mu_1(\cdot)$ and $\mu_2(\cdot)$ with densities g and h, respectively, w.r.t. some σ-finite measure $\rho(\cdot)$, there exist jointly defined random variables X and Y such that $X \sim \mu_1(\cdot)$, $Y \sim \mu_2(\cdot)$, and $P(X = Y) = 1 - ||\mu_1(\cdot) - \mu_2(\cdot)||$.

Proof: Let $a = \int_{\mathcal{X}} m d\rho$, $b = \int_{\mathcal{X}} (g-m)d\rho = \int_{\mathcal{X}} g d\rho - \int_{\mathcal{X}} m d\rho = 1 - a$, $c = \int_{\mathcal{X}} (h-m)d\rho$, where $m = min(g,h)$. When any of a, b, c are zero, the proof is trivial. So, consider the case when they are all positive. We jointly construct random variables U, V, W and I such that U has density $\frac{m}{a}$, V has density $\frac{g-m}{b}$, W has density $\frac{h-m}{c}$, and I is independent of U, V, and W with $P[I = 1] = a$ and $P[I = 0] = 1 - a$. Let $X = Y = U$ is $I = 1$, and $X = V$, $Y = W$ if $I = 0$. Since $X \sim \frac{m}{a} \cdot (a) + \frac{g-m}{b} \cdot (1-a) = m + g - m = g$ $=> X \sim \mu_1(\cdot)$ and, similarly, $Y \sim \mu_2(\cdot)$. Also, $P[X = Y] = P[I = 1] = a = \int_{\mathcal{X}} m d\rho = 1 - ||\mu_1(\cdot) - \mu_2(\cdot)||$ (by the previous property).

□

3 When Does a Markov Chain Converge?

Our goal is to find out when and how fast our Markov chain converges to the stationary distribution. Note that even if a Markov chain has a stationary distribution $\pi(\cdot)$, it still may fail to converge to it. Consider the following example:

Example. Let $\mathcal{X} = \{1,2,3\}$, $\pi(\{1\}) = \pi(\{2\}) = \pi(\{3\}) = \frac{1}{3}$. Let $P(1,\{1\}) = P(1,\{2\}) = P(2,\{1\}) = P(2,\{2\}) = \frac{1}{2}$ and $P(3,\{3\}) = 1$. Then it's easy to check that π is stationary, but if we start at $\{1\}$, that is, if $X_0 = 1$, then $X_n \in \{1,2\}$ $\forall n$. Thus, $P(X_n = 3) = 0$ $\forall n$, and so, $P(X_n = 3) \not\to \pi(\{3\})$, and therefore the distribution of X_n doesn't coverge to $\pi(\cdot)$. (This example, and examples below can be also found in [1].)

Note that here we have a reducible Markov chain, and countable state space. For uncountable state space we'll need a general definition for irreducibility.

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

Note: If a chain is irreducible, then it has many different irreducibility distributions. However, it is possible to show that any irreducible chain has a maximal irreducible distribution in the sense that all other irreducibility distributions are absolutely continuous with respect to it. Maximal irreducibility distributions are not unique but are equivalent, in sense that they have the same null sets. From now on when we say *ϕ-irreducible,* we mean that ϕ is a maximal irreducibility distribution.

But even if our chain is ϕ-irreducible it still might not converge to its stationary distribution.

Example. Let $\mathcal{X} = \{1,2,3\}$ with $\pi(\{1\}) = \pi(\{2\}) = \pi(\{3\}) = \frac{1}{3}$. Let $P(1,\{2\}) = P(2,\{3\}) = P(3,\{1\}) = 1$. Then $\pi(\cdot)$ is stationary and the chain is ϕ-irreducible (take $\phi(\cdot) = \delta_1(\cdot)$). But if we start at $\{1\}$, that is, $X_0 = 1$, then $X_n = 1 \; \forall n = 3k$, $k = 1, 2, ...$ So, $P(X_n = 1)$ oscillates between 0 and 1, and again we have that $P(X_n = 1) \not\to \pi(\{3\})$, and there is no convergence.

In this case we have a periodic chain, so to have convergence we need to get rid of the periodicity of our Markov chain.

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

3.1 The Asymptotic Convergence Theorem

Now we can formulate the main theorem of the current section.

Theorem 1. If a Markov chain with a stationary distribution $\pi(\cdot)$ on a state space $\mathcal{X}$ with countable generated σ-algebra is ϕ-irreducible and aperiodic, 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 any measurable $A \subseteq \mathcal{X}$.

Note that the 'π-a.e.' part is very important in this theorem, because the chain may fail to converge on a set of π-measure zero.

Example. Let $\mathcal{X} = \{1, 2, ...\}$. Let $P(1, \{1\}) = 1$, and for $x \geq 2$ $P(x, \{1\}) = \frac{1}{x^2}$ and $P(x, \{x+1\}) = 1 - \left(\frac{1}{x^2}\right)$. Then the chain has a stationary distribution $\pi(\cdot) = \delta_1(\cdot)$, and it is π-irreducible and aperiodic. But if $X_0 = x \geq 2$, then $P[X_n = x + n \; \forall n] = \prod_{j=x}^{\infty} \left(1 - \frac{1}{j^2}\right) > 0$. So, $||P_n(x, \cdot) - \pi(\cdot)|| \not\to 0 \; \forall x \geq 2$.

But what if we want our statement be true for all x, not just π-a.e.? It turns out that it's enough for a chain to be *Harris recurrent.*

Definition. A ϕ-irreducible Markov chain is *Harris recurrent* if for $\forall A \subseteq \mathcal{X}$ with $\phi(A) > 0$, and $\forall x \in \mathcal{X}$, $P[\exists n : X_n \in A | X_0 = x] = 1$, in words, the chain will eventually reach A from x with probability 1.

Now we know *when* our Markov chain converges, but we still don't know *how fast*.The following section will partially answer this question.

4 Quantative Convergence Rates

We want is to find quantative bounds on convergence rates, i.e. to find some explicit function $g(x,n)$ such that $||P^n(x,\cdot)-\pi(\cdot)|| \leq g(x,n)$ and which is getting small when n grows up.

4.1 Geometric and Uniform Ergodicity

Consider two types of Markov chains that allow us to define bounds for the total variation form. For this task we need a stronger condition:

Definition. A Markov chain is *ergodic* if it's irreducible, aperiodic and positive Harris recurrent.

One convergence rate condition that is often considered is geometric ergodicity

Definition. An ergodic Markov chain with stationary distribution π is *geometrically ergodic* if there exist a non-negative extended real valued function M which is finite for π-a.e. $x \in \mathcal{X}$, and a positive constant $\rho < 1$ such that for $n = 1, 2, 3...$

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

A stronger condition is uniform ergodicity:

Definition. An ergodic Markov chain with stationary distribution π is *uniformly ergodic* if there exist a positive, finite constant M and a positive constant $\rho < 1$ such that

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

for all x and all n.

Proposition 4.2.1. A Markov chain with stationary distribution π is uniformly ergodic if and only if $\sup_{x\in\mathcal{X}} ||P^n(x,\cdot)-\pi(\cdot)|| < \frac{1}{2}$ for some $n \in N$.

Proof: (=>) Let the Markov chain be uniformly ergodic, then

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

So, for n large enough we have $\sup_{x\in\mathcal{X}} ||P^n(x,\cdot) - \pi(\cdot)|| < \frac{1}{2}$.

(<=) Let $\sup_{x\in\mathcal{X}} ||P^n(x,\cdot)-\pi(\cdot)|| < \frac{1}{2}$ for some $n \in N$. Then by Proposition 2.3 of the total variation norm we have that $2\sup_{x\in\mathcal{X}} ||P^n(x,\cdot)-\pi(\cdot)|| = t(n) \equiv \beta < 1$, so using submultiplicativity we get that $\forall j \in N$ $t(jn) = t(\underbrace{n + ... + n}_{j}) \le \underbrace{t(n) \cdot ... \cdot t(n)}_{j} = t(n)^j = \beta^j$. Therefore from Proposition 2.2 of the total variation norm it follows that $||P^m(x,\cdot)-\pi(\cdot)|| \le ||P^{\lfloor m/n \rfloor n}(x,\cdot)-\pi(\cdot)|| \le \frac{1}{2}t(\lfloor m/n \rfloor n) \le \frac{1}{2}\beta^{\lfloor m/n \rfloor} \le \beta^{\lfloor m/n \rfloor} \le \beta^{-1}(\beta^{1/n})^m = M\rho^m$, where $M = \beta^{-1} < \infty$ and $\rho = \beta^{1/n} < 1$. So the chain is uniformly ergodic.

□

Before we state and prove the convergence theorems for geometrically and uniformly ergodic Markov chains let us recall that if we are given a Markov chain with transition probability P on a state space $\mathcal{X}$, and a measurable function $f : \mathcal{X} \to R$, we define the function $Pf : \mathcal{X} \to R$ so that $(Pf)(x)$ is the conditional expected value of $f(X_{n+1})$ given $X_n = x$, i.e. $(Pf)(x) = E[f(X_{n+1})|X_n = x]$ or,

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

We also define minorisation and drift conditions for a Markov chain:

Definition. A subset $C \subseteq \mathcal{X}$ is *small (or* (n_0,ϵ,ν)*-small)* if there exist $n_0 \in \mathbf{N}$, $\epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that

$$P^{n_0}(x,A) \ge \epsilon\nu(A) \tag{5}$$

$\forall x \in C$ and $\forall$ measurable $A \subseteq \mathcal{X}$. The inequality (5) is called a *minorisation condition.*

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

$$PV(x) \le \lambda V(x) + bI_C(x) \tag{6}$$

$\forall x \in \mathcal{X}$ and some small set $C \subseteq \mathcal{X}$.

Definition. A Subset $C \subseteq \mathcal{X}$ is *petite* (*(n_0, ϵ, ν)-petite*), relative to a Markov chain kernel P, if $\exists n_0 \in \mathbf{N}$, $\epsilon > 0$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that

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

for all $x \in C$.

It's easy to check that any small set is petite. The converse is false in general, but

Lemma 4.2.2. For an aperiodic, ϕ-ireducible Markov chain, all petite sets are small sets. (The proof of this lemma can be found in Chapter 5 of [2].)

To prove convergence theorems we shall use the coupling method, so let's introduce it first.

4.2 Method of Coupling

Here we shall repeat the main idea of coupling from [1]. Suppose we have two random variables X and Y, defined jointly on some space $\mathcal{X}$. Let $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ be their probability distributions. Then

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

Thus,

$$||\mathcal{L}(X) - \mathcal{L}(Y)|| \leq P(X \neq Y). \tag{7}$$

In words, the total variation norm between the laws of two random variables is bounded by the probability that they are not equal. The inequality (7) is called *the coupling inequality.*

Now, suppose C is a small set. The coupling idea is to run two copies of the Markov chain, $\{X_n\}$ and $\{X'_n\}$, each of which is being marginally updated from $P(x, \cdot)$, but whose joint construction gives a high probability to become equal to each other, i.e. to couple.

THE COUPLING CONSTRUCTION:

We start with $X_0 = x$ and $X'_0 \sim \pi(\cdot)$, where π is a stationary distribution of our Markov chain $\{X_n\}$, and $n = 0$, and repeat the following loop forever.

Beginning of Loop. Given X_n and X'_n:

1. If $X_n = X'_n$, then choose $X_{n+1} = X'_{n+1} \sim P(X_n, \cdot)$, and replace n by $n + 1$.

2. Else, if $(X_n, X'_n) \in C \times C$, then:

   (a) with probability ϵ, choose $X_{n+n_0} = X'_{n+n_0} \sim \nu(\cdot)$;

   (b) else, with probability $1 - \epsilon$, conditionally independently choose

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

   If $n_0 > 1$, then we go back to construct $X_{n+1}, ..., X_{n+n_0-1}$ from their correct conditional distributions given X_n and X_{n+n_0}, and also conditionally independently construct $X'_{n+1}, ..., X'_{n+n_0-1}$ from their correct conditional distributions given X'_n and 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 $X'_{n+1} \sim P(X'_n, \cdot)$, and replace n by $n + 1$.

Return to Beginning of Loop.

Under this construction, X_n and X'_n are each marginally updated according to the transition kernel P, and $P[X_n \in A] = P^n(x, A)$ and $P[X'_n \in A] =$

$\pi(A)$ for all n and $A \subseteq \mathcal{X}$. Note that the two chains are run independently until they both in C at which time the minorisation splitting construction is utilised. Such construction helps us to ensure successful coupling of the two chains.

We shall show that we can use it to obtain bounds on $||P^n(x,\cdot) - \pi(\cdot)||$. In fact, we shall use the coupling construction to prove important theorems that we'll state below.

4.3 Statements and Proofs of Convergence Theorems

We shall need a *bivariate drift condition* of the form:

$$\overline{P}h(x,y) \leq \frac{h(x,y)}{\alpha} \tag{8}$$

$\forall (x,y) \notin C \times C$, for some function $h : \mathcal{X} \times \mathcal{X} \to [1,\infty)$ and some $\alpha > 1$, where

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

Here $\overline{P}$ represents running *two* independent copies of the Markov chain.

Then we can rewrite the coupling inequality as

$$||P^n(x,\cdot) - \pi(\cdot)|| \leq P[X_n \neq X'_n].$$

The following proposition will be very useful in our proofs. So let's state and prove it in a little more details than it is in [1].

Proposition 4.3.1. Let the univariate drift condition $PV \leq \lambda V + bI_C$ be 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 $d > \frac{b}{1-\lambda} - 1$, then the bivariate drift condition holds for the same set C, with $h(x,y) = \frac{1}{2}[V(x) + V(y)]$ and $\alpha = \left(\lambda + \frac{b}{d+1}\right)^{-1} > 1$.

Proof: When $(x,y) \notin C \times C$, then either $x \notin C$ or $y \notin C$, or both. Without loss of generality assume that $x \notin C \Rightarrow x \in C^c \Rightarrow V(x) \geq d = \inf_{x \in C^c} V(x)$, and we have that $V(y) \geq 1 \; \forall y$. Thus,

$$h(x,y) = \frac{1}{2}[V(x) + V(y)] \geq \frac{1}{2}[d+1] = \frac{1+d}{2}.$$

So, $\frac{2h(x,y)}{1+d} \geq 1$.

Also we have that

$$\begin{aligned} PV(x) &\leq \lambda V(x) + bI_C = \lambda V(x) \\ PV(y) &\leq \lambda V(y) + bI_C \leq \lambda V(y) + b. \end{aligned}$$

Therefore,

$$PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b.$$

Then

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

And if $d > \frac{b}{1+\lambda} - 1 => 1 - \lambda > \frac{b}{1+d} => \lambda + \frac{b}{1+d} < 1$, then $\alpha = \left(\lambda + \frac{b}{1+d}\right)^{-1} > 1$. So, the bivariate drift condition is satisfied.

□

Now, let $B = max[1, \alpha(1-\epsilon)\sup_{C\times C} \overline{R}h]$, where

$$\overline{R}h(x,y) = \int_{\mathcal{X}} \int_{\mathcal{X}} (1-\epsilon)^{-2} h(z,w)[P(x,dz) - \epsilon\nu(dz)][P(y,dw) - \epsilon\nu(dw)] \quad (9)$$

for all $(x,y) \in C \times C$.

Using all above we can now give quantative bounds on total variation norm.

Theorem 2. Consider a Markov chain on a state space $\mathcal{X}$ with transition kernel P. Suppose $\exists C \subseteq \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, a probability distribution $\nu(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, $n_0 \in \mathbf{N}$, $\epsilon > 0$ such that $P^{n_0}(x,\cdot) \geq \epsilon\nu(\cdot)$ and $\overline{P}h(x,y) \leq \frac{h(x,y)}{\alpha}$, $\forall (x,y) \notin C \times C$. Define B as above. Then for any joint initial distribution $\mathcal{L}(X_0, X_0')$, and any integers $1 \le j \le k$, 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

$$||\mathcal{L}(X_k) - \mathcal{L}(X_k')|| \le (1-\epsilon)^j + \alpha^{-k} B^{j-1} E[h(X_0, X_0')].$$

Proof: Let $N_k = \#\{m : 0 \le m \le k, (X_m, X_m') \in C \times C\}$. Then for $1 \le j \le k$ we have that

$$\begin{aligned} ||\mathcal{L}(X_k) - \mathcal{L}(X_k')|| &\le P[X_k \ne X_k'] \\ &= P[X_k \ne X_k', Nk-1 \ge j] + P[X_k \ne X_k', N_{k-1} < j] \end{aligned}$$

We can right away estimate the first term of the above sum, since

$$\{X_k \ne X_k', Nk-1 \ge j\} \subseteq \{\text{the first j coin flips all come up tails}\}.$$

Thus,

$$P[X_k \ne X_k', N_{k-1} \ge j] \le (1-\epsilon)^j.$$

So, it's left to estimate the second term. To do that, define

$$M_k = \alpha^k B^{-N_{k-1}} h(X_k, X_k') I_{\{X_k \ne X_k'\}},$$

$k = 0, 1, 2, ...; N_{-1} = 0$.

Claim: $\{M_k\}$ is a supermartingale, i.e.

$$E[M_{k+1} | X_0, ... X_k, X_0', ..., X_k'] \le M_k.$$

Proof of the Claim: Consider two cases:

<u>Case 1</u>: $(X_k, X_k') \notin C \times C$, then $N_k = N_{k-1}$. Thus,

$$\begin{aligned} E[M_{k+1} | X_0, ..., X_k, X_0', ..., X_k'] &= E[M_{k+1} | X_k, X_k'] \text{ (since } \{X_n\} \text{ is a Markov chain)} \\ &= \alpha^{k+1} B^{-N_{k-1}} E[h(X_{k+1}, X_{k+1}') I_{\{X_{k+1} \ne X_{k+1}'\}} | X_k, X_k'] \\ &\le \alpha^{k+1} B^{-N_{k-1}} E[h(X_{k+1}, X_{k+1}') | X_k, X_k'] I_{\{X_k \ne X_k'\}} \\ &= M_k \alpha \frac{E[h(X_{k+1}, X_{k+1}') | X_k, X_k']}{h(X_k, X_k')} \\ &= M_k \alpha \frac{\overline{P} h(X_k, X_k')}{h(X_k, X_k')} \le M_k \alpha \frac{h(X_k, X_k')}{\alpha h(X_k, X_k')} \\ &\quad \text{(since } \overline{P} h(x,y) \le \tfrac{h(x,y)}{\alpha} \text{ for } (x,y) \notin C \times C) \\ &= M_k \end{aligned}$$

So, in this case $\{M_k\}$ is a supermartingale.

Case 2: $(X_k, X'_k) \in C \times C$, then $N_k = N_{k-1} + 1$. Assume that $X_k \neq X'_k$ (otherwise case is trivial). Then

$$\begin{aligned}
E[M_{k+1}|X_0, ..., X_k, X'_0, ..., X'_k] &= E[M_{k+1}|X_k, X'_k] \text{ (since } \{X_n\} \text{ is a Markov chain)} \\
&= \alpha^{k+1} B^{-N_{k-1}-1} E[h(X_{k+1}, X'_{k+1}) I_{\{X_{k+1} \neq X'_{k+1}\}} | X_k, X'_k] \\
&= \alpha^{k+1} B^{-N_{k-1}-1} (1-\epsilon) \overline{R} h(X_k, X'_k) \\
&= M_k \alpha B_{-1} (1-\epsilon) \frac{\overline{R} h(X_k, X'_k)}{h(X_k, X'_k)} \\
&\leq M_k \text{ (since } B = max(1, \alpha(1-\epsilon) \sup_{C \times C} \overline{R} h) \text{ and } h \geq 1)
\end{aligned}$$

Thus,$\{M_k\}$ is a supermartingale.

□

So, since $\{M_k\}$ is a supermartingale, we can show, by taking expectation and using induction, that $E[M_k] \leq E[M_0]$ $\forall k = 0, 1, 2, ...$

Now, since $B \geq 1$,

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

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

□

Now we shall state two important results that give us more information about the rate of convergence of the Markov chain to its stationary distribution.

Theorem 3. Let $\{X_n\}$ be a Markov chain with stationary distribution $\pi(\cdot)$. Suppose the minorisation condition holds for all $x \in \mathcal{X}$. More formally, let for all $x \in \mathcal{X}$ $P^{n_0}(x,\cdot) \geq \epsilon\nu(\cdot)$ be satisfied for some $n_0 \in \mathbf{N}$ and $\epsilon > 0$ and probability measure $\nu(\cdot)$, i.e. $\mathcal{X}$ is a small set itself. 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 $x \in \mathcal{X}$.

Proof: Since $C = \mathcal{X}$, every n_0 iteration X_n and X'_n might be equal with probability at least ϵ. Then, if $n = n_0 m$, $P[X_n \neq X'_n] \leq (1-\epsilon)^m$. Therefore,

$$||P^n)x,\cdot) - \pi(\cdot)|| \leq P[X_n \neq X'_n] \leq (1-\epsilon)^m = (1-\epsilon)^{\frac{n}{n_0}}.$$

Then, by Proposition 2.2, $||P^n(x,\cdot) - \pi(\cdot)|| \leq (1-\epsilon)^{\lfloor \frac{n}{n_0} \rfloor}$ $\forall n$.

□

Before we state the next theorem, let's prove the follwing lemma, that appears to be a useful tool in our proof.

Lemma 4.3.2. Given a small set C satisfying the minorisation condition, and a drift fucntion V satisfying the drift condition, there exists a small set $C_0 \subseteq C$ such that these conditions still hold (with the same n_0, ϵ and b, but with λ replaced by some $\lambda_0 < 1$), and such that

$$\sup_{x \in C} V(x) < \infty \tag{10}$$

Proof: Let λ and b be as in the drift condition. Choose δ such that $0 < \delta < 1 - \lambda$, and let $\lambda_0 = 1 - \delta$. Thus, $\lambda < \lambda_0 < 1$. Let $K = \frac{b}{1-\lambda-\delta}$, and let $C_0 = C \bigcap \{x \in \mathcal{X} : V(x) \leq K\}$. Since $C_0 \subseteq C$, then the minorisation condition hold for C_0. So, we need to show that the drift condition also holds on C_0 with λ_0, i.e. for all $x \in \mathcal{X}$

$$PV(x) \leq \lambda_0 V(x) + bI_{C_0}(x).$$

We have that for all $x \in \mathcal{X}$

$$PV(x) \leq \lambda V(x) + bI_C(x).$$

Thus, for all $x \in C_0$ we have

$$PV(x) \leq \lambda V(x) + bI_C(x) = \lambda V(x) + b \leq \lambda_0 V(x) + b = \lambda_0 V(x) + bI_{C_0}(x).$$

For all $x \in C^c$ we have

$$PV(x) \leq \lambda V(x) + bI_C(x) = \lambda V(x) \leq \lambda_0 V(x) = \lambda_0 V(x) + bI_{C_0}(x).$$

For all $x \in C \setminus C_0$ we have that $V(x) > K$, and therefore

$$\begin{aligned} PV(x) \leq \lambda V(x) + bI_C(x) &= (1-\delta)V(x) - (1-\lambda-\delta)V(x) + b \\ &\leq (1-\delta)V(x) - (1-\lambda-\delta)K + b \\ &= \lambda_0 V(x) - b + b = \lambda_0 V(x) \\ &= \lambda_0 V(x) + bI_{C_0}(x) \end{aligned}$$

□

Theorem 4. Let $\{X_n\}$ be a Markov chain with stationary distribution $\pi(\cdot)$. Suppose that the minorization condition is satisfied for some $C \subset \mathcal{X}$, $n_0 \in \mathbf{N}$, $\epsilon > 0$ and probability measure $\nu(\cdot)$. Suppose that also 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 (hence for π-a.e.) $x \in \mathcal{X}$. Then the chain is geometrically ergodic.

Proof: Set $h(x,y) = \frac{1}{2}[V(x) + V(y)]$. According the above lemma we can assume that (10) holds, which is together with the drift condition gives us that $\sup_{C \times C} \overline{R}h(x,y) < \infty$. Thus,

$$B = max[1, \alpha(1-\epsilon) \sup_{C \times C} \overline{R}h] < \infty.$$

Let $d = \inf_{x \in C^c} V(x)$. Then if $d > \frac{b}{1-\lambda} - 1$, by Proposition 4.3.1, the bivariate drift condition will hold. In this case our theorem follows from Theorem 2. But if $d \leq \frac{b}{1-\lambda} - 1$, then we cannot prove the theorem in that way. The condition $d > \frac{b}{1-\lambda} - 1$ ensures that the chain is aperiodic, and when we don't have this condition, we have to use the aperiodicity more directly. If we could enlarge C so that the new value of d would satisfy $d > \frac{b}{1-\lambda} - 1$, and to use apperiodicity to show that this enlarged C is still a small set, then we would finish the proof. (*Note:* We then shall have no direct control over the new values of n_0 and ϵ, and, thus, this approach does not provide a quantitative convergence rate bound.)

So, let's continue. Choose any $d' > \frac{b}{(1-\lambda)-1}$, and let $S = \{x \in \mathcal{X} : V(x) \leq d'\}$. Set $C' = C \bigcup S$. Then

$$d = \inf_{x \in C^c} V(x) \geq d' > \frac{b}{1-\lambda} - 1.$$

Since (10) holds on C and V is bounded on S, then (10) will still hold on $C' = C \bigcup S$. Then $B < \infty$ even upon replacing C by C'. Therefore our theorem will follow from Proposition 4.3.1. and Theorem 2, if we show that C' is a small set. Since our chain is aperiodic and ϕ-irreducible, then, by Lemma 4.2.2, it's enough to show that C' is a pitite set. (*Note:* This is where we use aperiodicity.)

Claim: C' is a petite set.

Proof of the Claim: Since $\lambda < 1$, it is possible to find N large enough so that $r \equiv 1 - \lambda^N d > 0$. Let's denote the first return time to C by $\tau_C = \inf\{n \geq 1, X_n \in C\}$. Let $Z_n = \lambda^{-n} V(X_n)$, and let $W_n = Z_{min(n,\tau_C)}$. Then if $\tau_C \leq n$, we have that

$$\begin{aligned} E[W_{n+1}|X_0, X_1, ..., X_n] &= E[Z_{min(n+1,\tau_C)}|X_0, X_1, ..., X_n] \\ &= E[Z_{\tau_C}|X_0, X_1, ..., X_n] \\ &= Z_{\tau_C} = W_n. \end{aligned}$$

And if $\tau_C > n$, then $X_n \notin C$, and, thus,

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

$$\begin{aligned} &\leq \lambda^{-(n+1)}\lambda V(X_n) \text{ (since } PV \leq \lambda V + bI_C = \lambda V) \\ &= \lambda^{-n}V(X_n) = W_n. \end{aligned}$$

Hence we showed that $\{W_n\}$ is a supermartingale. Now, for $x \in S$ we have:

$$\begin{aligned} P[\tau_C \geq N|X_0 = x] &= P[\lambda^{-\tau_C} \geq \lambda^{-N}|X_0 = x] \text{ (since } \lambda^{-1} > 1) \\ &\leq \frac{E[\lambda^{-\tau_C}|X_0 = x]}{\lambda^{-N}} \text{ (by Markov)} \\ &\leq \lambda^N E[\lambda^{-\tau_C} V(X_{\tau_C})|X_0 = x] \text{ (since } V \geq 1) \\ &= \lambda^N E[Z_{\tau_C}|X_0 = x] \\ &\leq \lambda^N E[Z_0|X_0 = x] \text{ (since } \{Z_{min(n,\tau_C)}\} = \{W_n\} \text{ is a supermartingale)} \\ &= \lambda^N E[V(X_0)|X_0 = x] = \lambda^N V(x) \leq \lambda^N d \text{ (since } x \in S). \end{aligned}$$

Therefore,

$$P[\tau_C < N|X_0 = x] = 1 - P[\tau_C \geq N|X_0 = x] \geq 1 - \lambda^N d \equiv r.$$

Since $\forall x \in C$ $P^{n_0}(x, \cdot) \geq r\epsilon\nu(\cdot)$, then $\forall x \in S$ we have that

$$\begin{aligned} \sum_{i=1+n_0}^{N+n_0} P^i(x, \cdot) &= \sum_{i=1}^{N} P^{i+n_0}(x, \cdot) \geq \int_C \sum_{i=1}^{N} P^i(x, dy) P^{n_0}(y, \cdot) \\ &\geq \sum_{i=1}^{N} P^i(x, C)\epsilon\nu(\cdot) = \sum_{i=1}^{N} P_x(X_i \in C)\epsilon\nu(\cdot) \\ &\geq P_x\Big[\bigcup_{i=1}^{N}\{X_i \in C\}\Big]\epsilon\nu(\cdot) = P[\tau_C < N|X_0 = x]\epsilon\nu(\cdot) \geq r\epsilon\nu(\cdot). \end{aligned}$$

Hence, $\forall x \in S \bigcup C$, $\sum_{i=n_0}^{N+n_0} P^i(x, \cdot) \geq r\epsilon\nu(\cdot)$. Thus, $C' = C \bigcup S$ is petite, which proves the claim and, hence, the theorem 4.

□

5 Proof of the Asymptotic Convergence Theorem

In this section we shall prove the Theorem 1 that we stated in the Section 3, i.e. if we have a ϕ-irreducible and aperiodic Markov chain on a state space $\mathcal{X}$, and it has a stationary distribution $\pi(\cdot)$, then $\lim_{n\to\infty} ||P^n(x,\cdot)-\pi(\cdot)|| = 0$, for π-a.e. $x \in \mathcal{X}$. The main details of this proof are taken from [1].

As we can notice, this theorem does not assume the existence of a small set C, which we need in order to use the coupling construction. Therefore, let us give an important result about the existence of a small set.

Theorem 5.1. Every ϕ-irreducible Markov chain, on a state space $\mathcal{X}$ with countably generated σ-algebra, contains a small set $C \subseteq \mathcal{X}$ with $\phi(C) > 0$. Furthermore, the minorisation measure $\nu(\cdot)$ may be taken to satisfy $\nu(C) > 0$.

The proof of this theorem can be found in Chapter 5 of [2].

Now, knowing that a small set C exists, we can show that the pair (X_n, X'_n) in the coupling construction will hit $C \times C$ infinitely often. Thus, they will have infinitely many opportunities to couple with probability greater than $\epsilon > 0$. Therefore they will couple eventually with probability 1, proving our asymptotic theorem.

We'll state two useful lemmas that we are going to use, but for now we leave them without proofs, although the proofs can be found in [1].

Lemma 5.2. Let $\{X_n\}$ be a Markov chain on a state space $\mathcal{X}$ with a stationary distribution $\pi(\cdot)$. Suppose that for some $A \subseteq \mathcal{X}$, we have $P_x(\tau_A < \infty) > 0\ \forall x \in \mathcal{X}$. Then for π-a.e. $x \in \mathcal{X}$, $P_x(\tau_A < \infty) = 1$, where $\tau_A = \inf\{n \geq 1 : X_n \in A\}$.

Lemma 5.3. Let $\{X_n\}$ be an aperiodic Markov chain on a state space $\mathcal{X}$ with a stationary distribution $\pi(\cdot)$. Let $\nu(\cdot)$ be any probability measure on $\mathcal{X}$. Assume that $\nu(\cdot) \ll \pi(\cdot)$, and that $\forall x \in \mathcal{X}$ there exist $n = n(x) \in \mathbf{N}$ and $\delta = \delta(x) > 0$ such that $P^n(x,\cdot) \geq \delta\nu(\cdot)$. Also, 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 $\exists n_* \in N$ with $T \supseteq \{n_*, n_*+1, n_*+2, ...\}$.

Now, let C be a small set as in Theorem 5.1. Let $\{(X_n, X'_n)\}$ be our coupling construction. Let $G = \{(x,x') \in \mathcal{X} \times \mathcal{X} : P_{(x,x')}(\exists n \geq 1 : X_n = X'_n) = 1\}$. If $(X_0, X'_0) \equiv (x, X'_0) \in G$, then from the coupling construction it follows that $\lim_{n\to\infty} P[X_n = X'_n] = 1$. Hence, $\lim_{n\to\infty} ||P^n(x,\cdot) - \pi(\cdot)|| = 0$, which proves Theorem 1. So, it's enough for us to show that $P[(x, X'_0) \in G] = 1$ for π-a.e. $x \in \mathcal{X}$.

To show this, let $G_x = \{x' \in \mathcal{X} : (x, x,') \in G\}$ for $x \in \mathcal{X}$, and let $\overline{G} = \{x \in \mathcal{X} : \pi(G_x) = 1\}$. Then if we prove that $\pi(\overline{G}) = 1$, we'll finish our proof.

Let's first show that $(\pi \times \pi)(G) = 1$. Since, by Theorem 5.1, $\nu(C) > 0$, applying Lemma 5.3 we have that from any $(x,x') \in \mathcal{X} \times \mathcal{X}$, the joint chain will eventually hit $C \times C$ with positive probability. Then by applying Lemma 5.2 to the joint chain, it follows that the joint chain will return to $C \times C$ with probability 1 from $(\pi \times \pi)$-a.e. $(x,x') \notin C \times C$. When the joint chain hits $C \times C$, then conditional on not coupling, it will update from $\overline{R}$ which must be absolutely continuous with respect to $\pi \times \pi$, and therefore, applying Lemma 5.2 one more time, the chain will return to $C \times C$ with probability 1. Thus, the joint chain will repeatedly return to $C \times C$ with probability 1, until we get $X_n = X'_n$. By the coupling construction, whenever the chain hits $C \times C$, the event that $X_n = X'_n$ has probability greater than $\epsilon > 0$. So, eventually we shall have $X_n = X'_n$ and, thus, $(\pi \times \pi)(G) = 1$.

Now, assume by contradiction that $\pi(\overline{G}) < 1$. Then

$$(\pi \times \pi)(G^c) = \int_{\mathcal{X}} \pi(dx)\pi(G_x^c) = \int_{\mathcal{X}} \pi(dx)[1 - \pi(G_x)] = \int_{\overline{G}^c} \pi(dx)[1 - \pi(G_x)] > 0,$$

which contradicts the fact that $(\pi \times \pi)(G) = 1$. Hence, $\pi(\overline{G}) = 1$.

This finishes the proof of the asymptotic theorem.

□

6 V-Uniform Ergodicity

Let P_1 and P_2 be Markov transition kernels, and for a positive function $1 \leq V < \infty$ define the *V-norm distance* between P_1 and P_2 as

$$|||P_1(\cdot) - P_2(\cdot)|||_V := \sup_{x \in \mathcal{X}} \frac{||P_1(x, \cdot) - P_2(x, \cdot)||}{V(x)} \tag{11}$$

Note: We usually consider the distance $|||P^n - \pi|||_V$, which is actually is not defined by (11), because π is a probability measure, not a kernel. But if we consider π as a kernel by defining $\pi(x, A) := \pi(A)$, $A \subseteq \mathcal{X}$, $x \in \mathcal{X}$, then $|||P^n - \pi|||_V$ is well-defined.

Let's first show that $|||\cdot|||_V$ is an operator norm.

Lemma 6.1. Let L_V^∞ be the vector space of all functions $f : \mathcal{X} \to R_+$ such that

$$|f|_V := \sup_{x \in \mathcal{X}} \frac{|f(x)|}{V(x)} < \infty.$$

If $|||P^n - \pi|||_V$ is finite, then $P^n - \pi$ is a bounded operator from L_V^∞ to itself, and $|||P^n - \pi|||_V$ is its operator norm.

Proof: We can rewrite $|||\cdot|||$ as

$$\begin{aligned}
|||P^n - \pi|||_V &= \sup_{x \in \mathcal{X}} \left\{ \frac{\sup_{|g| \leq V} |P^n(x, g) - \pi(g)|}{V(x)} \right\} \\
&= \sup_{|g| \leq V} \sup_{x \in \mathcal{X}} \frac{|P^n(x, g) - \pi(g)|}{V(x)} \\
&= \sup_{|g| \leq V} |P^n(\cdot, g) - \pi(g)|_V \\
&= \sup_{|g|_V \leq 1} |P^n(\cdot, g) - \pi(g)|_V
\end{aligned}$$

which is, by definition, the operator norm of $P^n - \pi$.

□

Definition. An ergodic chain is *V-uniformly ergodic* if

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

Note: A chain is uniformly ergodic if it is V-uniformly ergodic in the special case where $V \equiv 1$, i.e. $\lim_{n\to\infty} \sup_{x\in\mathcal{X}} ||P^n(x,\cdot) - \pi(\cdot)|| = 0$.

Proposition 6.2. Let $\{X_n\}$ be a Markov chain with a stationary distribution π, and let for some n_0, $|||P - \pi|||_V < \infty$ and $|||P^{n_0} - \pi|||_V < 1$. Then the chain is V-uniformly ergodic.

Proof: Since $|||\cdot|||$ is an operator norm, then $\forall n, m \in N$ we have that

$$\begin{aligned} |||P^{n+m} - \pi|||_V &= |||(P-\pi)^n (P-\pi)^m|||_V (\text{since } \pi \text{ is stationary}) \\ &\leq |||P^n - \pi|||_V |||P^m - \pi|||_V \end{aligned}$$

Now, $\forall n \in N$ we write $n = jn_0 + i$, $1 \leq i \leq n_0$. Since $|||P - \pi|||_V < \infty$ without lost of generality we can find $1 \leq M < \infty$ such that $|||P - \pi|||_V \leq M$. Then since $|||P^{n_0} - \pi|||_V = \gamma < 1$ we have that

$$\begin{aligned} |||P^n - \pi|||_V &= |||P^{kn_0+i} - \pi|||_V = |||(P^{n_0} - \pi)^k (P-\pi)^i|||_V \\ &\leq |||P - \pi|||_V^i |||P^{n_0} - \pi|||_V^k \leq M^i \gamma^k \leq M^{n_0} \gamma^{(\frac{n}{n_0} - \frac{i}{n_0})} \\ &\leq M^{n_0} \gamma^{\frac{n}{n_o} - 1} = M^{n_0} \gamma^{-1} \left(\gamma^{\frac{1}{n_0}}\right)^n \end{aligned}$$

Thus, since $\gamma^{\frac{1}{n_0}} < 1$ and $M < \infty$, we have that $|||P^n - \pi|||_V \to 0$ as $n \to \infty$. By definition, it means that the Markov chain is V-uniformly ergodic.

$\square$

Note that when $V \equiv 1$, Proposition 6.2 is equivalent to Proposition 4.2.1, though the proofs are quite different. So we can say that the proof of Proposition 6.2 is an equivalent proof of Proposition 4.2.1 in the case when $V \equiv 1$.

6.1 Open Problem

Now we are ready to state an interesting fact and an existing problem, that we'll try to solve in the near future.

Fact 1. The minorisation condition (5) and the drift condition (6) are equivalent to the V-uniform ergodicity.

To prove this fact we would need to combine Proposition 1 of [4] and a few theorems from Meyn and Tweedie [2].

Since the Proposition 6.2 is a part of the proof of one of those theorems, let us state the theorem itself.

Theorem 6.1.1(see theorem 16.0.1 in [2]). Let a Markov chain $\{X_n\}$ be ϕ-irreducible and aperiodic. Then the following are equivalent for any $V \geq 1$:

(i) $\{X_n\}$ is V-uniformly ergodic.

(ii) There exists $r > 1$ and $R < \infty$ such that for all $n \in N$

$$|||P^n - \pi|||_V \leq Rr^{-n}.$$

(iii) There exists some $n > 0$ such that $|||P^i - \pi|||_V < \infty$ for $i \leq n$ and

$$|||P^n - \pi|||_V < 1.$$

(iv) The drift condition holds for some small set C and some V_0, where V_0 is equivalent to V in the sense that for some $c \geq 1$

$$c^{-1}V \leq V_0 \leq cV.$$

And the *open problem* consists in proving the above fact 1 using the coupling method described in the previous sections.

To be continued...

References

[1] Gareth O. Roberts and Jeffrey S. Rosenthal, March 2004 *General State Space Markov Chains and MCMC Algorithms.*

[2] S.P. Meyn and R.L. Tweedie, 1994 *Markov Chains and Stochastic Stability.*

[3] W.R. Gilks, S. Richardson and D.J. Spiegelhalter, 1996 *Markov Chain Monte Carlo in practice.*

[4] Gareth O. Roberts and Jeffrey S. Rosenthal, August 1,1996; revised, April 11,1997 *Geometric Ergodicity and Hybrid Markov Chains.*