Notes on General State Space Markov Chains and MCMC Algorithms

Yu Hang Jiang, Tong Liu, Zhiya Lou,
Shanshan Shangguan, Fei Wang, Zixuan Wu
Under the supervision of Jeffrey S. Rosenthal

May 2020

1 Introduction

This is a supplementary reading for the paper *General State Space Markov Chains and MCMC Algorithms* by Gareth O. Roberts and Jeffrey S. Rosenthal. It is meant to help readers have better understanding of the original paper. The paper provides introduction to MCMC (algorithms, motivations and relevant theory) $\rightarrow$ conditions for geometric and uniform ergodicity with quantitative bounds on the rate of convergence (using coupling constructions based on minorisation and drift conditions).

A Markov chain consists of
(i) a measurable state space $\mathcal{X}$
(ii) an initial distribution μ_0 on $\mathcal{X}$
(iii) transition probabilities $P(x, dy)$
(iv) $\int_A(x) = P(x, A)$
is a measurable function of $x \in \mathcal{X}$ for each fixed set $A \subset \mathcal{X}$
(v) Multi-step transition probability distribution:
Definition of μ_k on $\mathcal{X}$, the probabilities of where the Markov chain will be after k steps $\mu_k(X) = \int_{\mathcal{X}} P(x, A)\mu_{k-1}(dx)$. In discrete case, we can write $\mu_k(y) = \sum_x P(x, y)\mu_{k-1}(x)$. If we write μ_k as a row-vector, and P as a matrix with $[P]_{xy} = P(x, y)$, then $\mu_k = \mu_{k-1}P = \cdots = \mu_0 P^k$.

Markov chain Monte Carlo algorithms - such as the Metropolis-Hastings algorithms and the Gibbs sampler are ways of sampling from complicated probability distributions. It allows sampling from posterior distributions in Bayesian inference. These algorithms raise two important questions: what is the stationary distribution of the Markov chain involved in the algorithm and what is the rate of convergence.

Plan. Section 2: Introduction to MCMC
Section 3: Convergence Rate Results Commonly used in MCMC
Section 4: The eigenvalue connection
Section 5: Asymptotic Convergence
Section 6: Uniform Ergodicity
Section 7: Geometric Ergoidicity
Section 8: Quantitative Convergence Rate
Section 9: Convergence Proof using Coupling Constructions
Section 10: Appendix

1.1 The problem.

Suppose we have a density function[1] π_u on the state space $\mathcal{X}$. Then we can define a probability measure on $\mathcal{X}$, by

$$\pi(A) = \frac{\int_A \pi_u(x)dx}{\int_{\mathcal{X}} \pi_u(x)dx}$$

We want to estimate expectation of functions $f : \mathcal{X} \to \mathbb{R}$ with respect to π:

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

The problem is if $\mathcal{X}$ is high-dimensional and π_u is complicated, then it is infeasible to directly compute the integral.
The classical Monte Carlo solution to this problem is simulate random variables $Z_1, \ldots, Z_n \sim \pi(\cdot)$, and then estimate $\pi(f)$ by

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

Note this is a random variable with expected value $\pi(f)$ (i.e. the estimator is unbiased) and standard deviation of order $O(1/\sqrt{N})$. Then by CLT, the error $\hat{\pi}(f) - \pi(f)$ will have a limiting normal distribution as $N \to \infty$. Therefore we can compute $\pi(f)$ by computing samples (plus some regression techniques?). But the problem is if π_u is complicated, then it is very difficult to simulate i.i.d. random variables from $\pi(\cdot)$.
The MCMC solution is to construct a Markov chain on $\mathcal{X}$ which has $\pi(\cdot)$ as a stationary distribution, i.e.

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

Then for large n the distribution of X_n will be approximately stationary. We can set $Z_1 = X_n$ and get $Z_2, Z_3, \ldots, Z_n$ repeatedly.

Remark. In practice instead of starting a fresh Markov chain every time we take the successive X_n's, for example, $(N-B)^{-1}\sum_{i=B+1}^{N} f(X_i)$. We tend to ignore the dependence problem as many of the mathematical issues are similar in either implementation.

Remark. We have other ways of estimation, such as "rejection sampling" and "importance sampling". But MCMC algorithms is applied most widely.

[1] A function $f : R^n \to R$ is a (unnormalized) density function if it is non-negative and $\int_{\mathcal{X}} f < \infty$. The integral is usually taken with respect to Lebesgue measure, although we could have other settings. If the integral is 1, then it is called a normalized density function.

Summary Estimate $\pi(f)$ needs to take samples; taking sampling needs MCMC.

1.2 Motivation: Bayesian Statistics Computations

Let $L(y|\theta)$ be the likelihood function.(density of y given θ). Let the "prior" density of θ be p. The the "posterior" distribution of θ given y is then density which is proportional to

$$\pi_u(\theta) \equiv L(y|\theta)p(\theta)$$

(as $p(y)$ is a fixed constant, which we do not care about because it does not affect the probability measure π induced by π_u). The "posterior mean" of any function f is given by

$$\pi(f) = \frac{\int_{\mathcal{X}} f(x)\pi_u(x)dx}{\int_X \pi_u(x)dx}$$

For this reason, Bayesians are anxious to estimate such $\pi(f)$. MCMC is extremely helpful for such estimates.

Example of Bayesian inference. Assume the average number of years women spend in education in country μ_i follows normal distribution with mean 13 and standard deviation 0.5. μ is the French average. You're told that the number of years in France follows a normal distribution with $\sigma = 1$ year; you want to estimate μ. Then we know $p(\mu)$. Let x be the samples. Once we know μ, we get $L(x|\mu)$. Then we can get π_u

1.3 Bounds on Markov Chain Convergence Times

Definition The n-step transition law of the Markov chain is defined as:

$$P^n(x, A) = P[X_n \in A|X_0 = x]$$

Fundamental questions
(A) Does there exist a probability distribution π on $\mathcal{X}$ such that $||\mu_k - \pi|| \to 0$ as $k \to \infty$
(B) If so, then given $\epsilon > 0$, how large should k be to ensure $||\mu_k - \pi|| < \epsilon$

We will introduce two approaches:
Section 4: The eigenvalue connection, this is a simple case for finite state space Markov chain.
Section 5 and onwards: we'll focus more on the second way: using Coupling technique. This can be applied to general state space Markov chains.

2 MCMC and its construction

This section will explain how MCMC algorithm is constructed. Now we introduce *reversibility*.
Definition. A Markov Chain on state space $\mathcal{X}$ is reversible with respect to a probability distribution $\pi(\cdot)$ on $\mathcal{X}$, if

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

Proposition. A Markov Chain is reversible with respect to $\pi(\cdot)$, then $\pi(\cdot)$ is the stationary distribution for the chain.

Proof. By reversibility, we have

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

□

Now the simplest way to construct a MCMC algorithm which satisfies reversibility is using Metropolis-Hastings algorithm.

2.1 The Metropolis-Hastings Algorithm.

Suppose that $\pi(\cdot)$ has a (possibly unnormalized) density π_u. Let $Q(x,\cdot)$ be essentially any other Markov Chain, whose transitions also have a (possibly unnormalized) density, i.e. $Q(x,dy) \propto q(x,y)dy$.

First choose some X_0. Then given X_n, generate a *proposal* Y_{n+1} from $Q(X_n,\cdot)$. In the meantime we flip a independent bias coin with probability of heads equals to $\alpha(X_n, Y_{n+1})$, where

$$\alpha(x,y) = \min\left[1, \frac{\pi_u(y)q(y,x)}{\pi_u(x)q(x,y)}\right], \quad \pi(x)q(x,y) \neq 0$$

And $\alpha(x,y) = 1$ when $\pi(x)q(x,y) = 0$. Then if the coin is heads, we accept the *proposal* and set $X_{n+1} = Y_{n+1}$. If the coin is tails, then we reject the proposal and set $X_{n+1} = X_n$. Then we replace n by $n+1$ and repeat. The reason we take $\alpha(x,y)$ as above is explain as follow.

Proposition. The Metropolis-Hastings Algorithm produces a Markov Chain $\{X_n\}$ which is reversible with respect to $\pi(\cdot)$.

Proof. We want to show for any $x,y \in \mathcal{X}$,

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$$

Since if $x = y$ the solution is trivial. Then we can automatically assume that $x \neq y$. Let c be the normalizing constant. i.e. $c = \int_{\mathcal{X}} \pi_u(x)dx$. Then

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

By symmetry (i.e. simply switching x and y), we can also obtain

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

□

Here are some examples under specific conditions.

• **Symmetric Metropolis Algorithm.** Here $q(x,y) = q(y,x)$. Then the acceptance probability can be simplified as

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

• **Random Walk Metropolis-Hastings.** Here $q(x,y) = q(y-x)$ And $Q(x,\cdot) \sim \mu$ which can be any distribution.

• **Independence Sampler.** Here $q(x,y) = q(y)$ which means that $Q(x,\cdot)$ is independent of x.

• **Langevin algorithm.** Here the proposal is generated by

$$Y_{n+1} \sim \mathcal{N}(X_n + (\delta/2)\nabla \log \pi(X_n), \delta)$$

for some small $\delta > 0$

2.2 MCMC in practice

• **Running algorithm on a computer**. It is not complicated to run Metropolis-Hastings algorithm on computer. We need to run the proposal chain $Q(x,\cdot)$, and then do the accept/reject step. For example we can set $u \sim \text{Uniform}(u;0,1)$. If $u < \alpha$, then we accept it. Otherwise we reject it. Here we only need to compute the ratio of density such as $\pi_u(y)/\pi_u(x)$, so we do not need normalizing constant.

Meanwhile, it is also tricky to choose an appropriate *proposal distribution* $Q(x,\cdot)$ In the future section, we will discuss how the choice of proposal distribution affect its convergence.

• **Combining Chains.** If P_2 and P_2 are two different chains, each has a stationary distribution $\pi(\cdot)$, then we can combining them by considering the new chain P_1P_2 which also has stationary distribution $\pi(\cdot)$. Thus, it is perfectly acceptable, and quite common to make new MCMC algorithms out of old ones, by specifying that the new algorithm applies first the chain P_1, then the chain P_2, then the chain P_1 again, etc. And, more generally, it is possible to combine many different chains in this manner.

Note that, even if each of P_1 and P_2 are reversible, the combined chain P_1P_2 will in general not be reversible. It is for this reason that it is important, when studying MCMC, to allow for non-reversible chains as well.

2.3 The Gibbs Simpler

Suppose $\pi_u(\cdot)$ is d-dimensional density, with $\mathcal{X}$ an open subset of $\mathbf{R}^d$, and write $\mathbf{x} = (x_1, ..., x_d)$.

And we define the i^{th} **component Gibbs Sampler** as P_i such that we replace i^{th} component by a draw from the full conditional distribution of $\pi_u(\cdot)$ conditional on all the other components, which leaves all components besides i unchanged.

More formally, let

$$S_{x,i,a,b} = \{y \in \mathcal{X}; y_j = x_j \text{ for } j \neq i, \text{and } a \leq y_j \leq b\}.$$

Then

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

It follows immediately that P_i is reversible with respect to $\pi_u(\cdot)$. Since we can treat P_i as a special case of Metropolis-Hastings algorithm with $\alpha(x, y) \equiv 1$. Therefore P_i has a stationary distribution $\pi_u(\cdot)$.

Then we have two ways to construct a full Gibbs sampler.

• **The deterministic-scan Gibbs samplers**

$$P = P_1P_2..P_d.$$

It performs the d different Gibbs Sampler components in sequential order.

• **The random-scan Gibbs samplers**

$$P = \frac{1}{d}\sum_{i=1}^{d} P_i.$$

It does one of the d different Gibbs sampler components, chosen uniformly at random.

Either version produces an MCMC algorithm having $\pi_u(\cdot)$ as its stationary distribution. And the components get updated once at a time. Also the random-scan Gibbs sampler is reversible, while the deterministic-scan Gibbs sampler usually is not.

2.4 Variance Components Model (A Detailed Bayesian Example)

Variance Components Model is a typical example of a target density $\pi_u(\cdot)$ that arises in Bayesian statistics, in an effort to illustrate the problems and issues which arise.

Now we introduce a fixed constants μ_0 and positive constant a_1, b_1, a_2, b_2 and σ_0^2. Also we must have three hyper-parameters, $\sigma_\theta^2, \sigma_e^2$, and μ, each having prior distributions as follow

$$\sigma_\theta^2 \sim \mathbf{IG}(a_1, b_1); \quad \sigma_e^2 \sim \mathbf{IG}(a_2, b_2); \quad \mu \sim \mathcal{N}(\mu_0, \sigma_0^2).$$

where **IG** is the *Inverse Gaussian distribution*. Furthermore, we must have K conditionally independent parameters $\theta_1, \theta_2 ..., \theta_K$, with

$$\theta_i \sim \mathcal{N}(\mu, \sigma_\theta^2).$$

In terms of these parameters, we have data $\{Y_{ij}\}$, with the assumption that

$$Y_{ij} \sim \mathcal{N}(\theta_i, \sigma_e^2) \text{ conditionally independently, and } 1 \le i \le K; \quad 1 \le j \le J.$$

Then we can clearly see three hierarchies with all parameter conditioning on $\{Y_{ij}\}$.
Now consider the joint $K+3$ parameters given this data, then we have this distribution

$$\pi(\cdot) = \mathcal{L}(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, .., \theta_K | \{Y_{ij}\}),$$

defined on $\mathcal{X} = (0, \infty)^2 \times \mathbf{R}^{K+1}$. We would like to sample from this distribution. Now we compute the unnormalized density

$$\begin{aligned}\pi_u(\sigma_\theta^2, \sigma_e^2, \mu, \theta_1, .., \theta_K) \propto\ & e^{-b_1/\sigma_\theta^2} {\sigma_\theta^2}^{-a_1-1} e^{-b_2/\sigma_e^2} {\sigma_e^2}^{-a_2-1} e^{-(\mu-\mu_0)^2/2\sigma_0^2} \\ & \times \Pi_{i=1}^K [e^{-(\theta_i-\mu)^2/2\sigma_\theta} / \sigma_\theta^2] \times \Pi_{i=1}^K \Pi_{j=1}^J [e^{-(Y_{ij}-\theta_i)^2/2\sigma_e^2} / \sigma_e]\end{aligned}$$

We now begin with the Gibbs Samplers. Then we need additional distribution as follows:

$$\mathcal{L}(\sigma_\theta^2 | \sigma_e^2, \mu, \theta_1, .., \theta_K, Y_{ij}) = \mathbf{IG}\left(a_1 + \frac{1}{2}K, b_1 + \frac{1}{2}\sum_i (\theta_i - \mu)^2\right);$$

$$\mathcal{L}(\sigma_e^2 | \sigma_\theta^2, \mu, \theta_1, .., \theta_K, Y_{ij}) = \mathbf{IG}\left(a_2 + \frac{1}{2}KJ, b_2 + \frac{1}{2}\sum_{i,j} (Y_{ij} - \theta_i)^2\right);$$

$$\mathcal{L}(\mu | \sigma_e^2, \sigma_\theta^2, \theta_1, .., \theta_K, Y_{ij}) = \mathcal{N}\left(\frac{\sigma_\theta^2 \mu_0 + \sigma_0^2 \sum_i \theta_i}{\sigma_\theta^2 + K\sigma_0^2}, \frac{\sigma_\theta^2 \sigma_0^2}{\sigma_\theta^2 + K\sigma_0^2}\right);$$

$$\mathcal{L}(\theta_i | \mu, \sigma_e^2, \sigma_\theta^2, \theta_1, .., \theta_{i-1}, \theta_{i+1}, .., \theta_K, Y_{ij}) = \mathcal{N}\left(\frac{J\sigma_\theta^2 \bar{Y_i} + \sigma_e^2 \mu}{J\sigma_\theta^2 + \sigma_e^2}, \frac{\sigma_\theta^2 \sigma_e^2}{J\sigma_\theta^2 + \sigma_e^2}\right), \quad 1 \le i \le K.$$

where $\bar{Y_i} = \frac{1}{J}\sum_j Y_{ij}$. The Gibbs sampler then proceeds by updating the $K+3$ variables according to the above conditional distributions. This is feasible since the conditional distributions are all easily simulated (**IG** and $\mathcal{N}$).

We can also use **Metropolis-Hastings Algorithm** here. For example, we can choose symmetric random walk algorithm with the proposal distribution being $\mathcal{N}(X_n, \sigma^2 I_{K+3})$ for some appropriate $\sigma^2 > 0$. Then given X_n, we will proceed the algorithm as follows:

1 Choose $Y_{n+1} \sim \mathcal{N}(X_n, \sigma^2 I_{K+3})$;

2 Choose $U_{n+1} \sim \text{Uniform}(0, 1)$;

3 If $U_{n+1} < \frac{\pi_u(Y_{n+1})}{\pi_u(X_n)}$, we accept the proposal and set $X_{n+1} = Y_{n+1}$. Otherwise we reject the proposal and set $X_{n+1} = X_n$.

We now conclude that, for such typical target distributions $\pi(\cdot)$, both the Gibbs sampler and appropriate Metropolis-Hastings algorithms perform well in practice, and allow us to sample from $\pi(\cdot)$.

3 Total variance distance

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

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

This leads to two questions: is $\lim_{n\to\infty} ||P^n(x, \cdot) - \pi(\cdot)|| = 0$ and how large must n be so that $||P^n(x, \cdot) - \pi(\cdot)|| < \epsilon$?. In order to answer these questions, we need a digression on the properties of total variation distance.

Proposition 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) $||\nu_1(\cdot) - \nu_2(\cdot)|| = \frac{1}{b-a} \sup_{f:\mathcal{X}\to[a,b]} |\int f d\nu_1 - \int f d\nu_2|$ 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 $n \in \mathbb{N}$
(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 $m, n \in \mathbb{N}$
(f) If $\mu(\cdot)$ and $\nu(\cdot)$ have densities[2] g and h, respectively, with respect to some $\sigma-$finite measure $\rho(\cdot)$, and $M = \max\{g, h\}$ and $m = \min\{g, h\}$, then

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

[2]see appendix for definition

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

Proof. (a): Let $p(.)$ be any δ-finite measure such that $\nu_1 \ll \rho$, $\nu_2 \ll \rho$. Set $g = d\nu_1/d\rho$ *and* $h = d\nu_2/d\rho$. Then

$$\left| \int f d\nu_1 - \int f d\nu_2 \right| = \left| \int f(g-h) d\rho \right|$$

This is maximized (over all $0 \leq f \leq 1$) when $f = 1$ on $\{g > h\}$ and $f = 0$ on $\{h > g\}$ (We could also let $f = 1$ on $\{g < h\}$. It is easy to verify in both cases they are equal to $|\nu_1(A) - \nu_2(A)|$.) Then note

$$\sup_A |\nu_1(A) - \nu_2(A)| = \sup_A |\int_A g d\rho - \int_A h d\rho| = \sup_A |\int_A (g-h) d\rho|$$

is also maximized when $A = \{g > h\}$ and it gives the same maximum value. This shows the equivalence.

(b): The proof of (b) is very similar to (a). Let $\rho(.)$ be any δ-finite measure such that $\nu_1 \ll \rho$, $\nu_2 \ll \rho$. Let $g = d\nu_1/d\rho$ and $h = d\nu_2/d\rho$. Consider again:

$$|\int f d\nu_1 - \int f d\nu_2| = |\int f(g-h) d\rho|$$

This integral is maximized when $f = b$ on $\{g > h\}$ and $f = a$ on $\{h > g\}$, because then f will be the greatest when it's multiplied to a positive value and smallest when it's multiplied with a negative value (Compare the choice of f with (a), where we used the indicator function). Let $A = \{g > h\}$, then

$$\begin{aligned}
|\int f(g-h) d\rho| &= |b \int_A (g-h) d\rho + a \int_{\mathcal{X} \setminus A} (g-h) d\rho| \\
&= |b[v_1(A) - v_2(A)] + a[(1 - v_1(A)) - (1 - v_2(A))]| \\
&= |b(v_1(A) - v_2(A) - a(v_1(A) - v_2(A))| \\
&= (b-a)|v_1(A) - v_2(A)| \\
\implies \frac{1}{b-a} |\int f(g-h) d\rho| &= |\nu_1(A) - \nu_2(A)|
\end{aligned}$$

We showed in (a) that the set A also maximized $|v_1(A) - v_2(A)|$, thus proving the equivalence.

(c): For any measurable set A,

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

(the last inequality follows from (a)). Then clearly

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

(d): For any measurable set A,

$$|v_1P(A) - v_2P(A)| = |\int_{y\in\mathcal{X}} v_1(dx)P(y,A) - \int_{y\in\mathcal{X}} v_2(dx)P(y,A)| \leq ||v_1(\cdot) - v_2(\cdot)||$$

Then d follows clearly. (Note (c) is a actually special case of (d), with $v_1 = P^n(x,\cdot)$ and $v_2 = \pi(\cdot)$.)

(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

$$\begin{aligned}
\hat{P}\hat{Q}f(x) &\equiv \int_{y\in\mathcal{X}} f(y) \int_{z\in\mathcal{X}} [P^n(x,dz) - \pi(dz)][P^m(z,dy) - \pi(dy)] \\
&= \int_{y\in\mathcal{X}} f(y) \int_{z\in\mathcal{X}} (P^n(x,dz)P^m(z,dy) - \pi(dz)P^m(z,dy)) \\
&- \int_{y\in\mathcal{X}} f(y)\pi(dy) \int_{z\in\mathcal{X}} P^n(x,dz) - \pi(dz) \\
&= \int_{y\in\mathcal{X}} f(y)[P^{n+m}(x,y) - \pi(dy) - \pi(dy)(1-1)] \\
&= \int_{y\in\mathcal{X}} [P^{n+m}(x,dy) - \pi(dy)]
\end{aligned}$$

In other words, if we define $\hat{H}(x,\cdot) = P^{m+n}(x,\cdot) - \pi(\cdot)$, then

$$\hat{P}\hat{Q}f(x) = \hat{H}f(x) \equiv \int_{y\in\mathcal{X}} \hat{H}(x,dy)f(y)$$

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

$$|g(x)| = \left| \int_y P^m(x,dy)f(y) - \int_y \pi(dy)f(y) \right| \leq ||P^m(x,\cdot) - \pi(\cdot)|| \text{ (by part(a))}$$

And

$$g^* \leq \sup_{x\in\mathcal{X}} ||P^m(x,\cdot) - \pi(\cdot)|| = \frac{1}{2}t(m)$$

It is easy to verify $\hat{P}\hat{Q}f = \hat{P}(\hat{Q}f)$ (We left it as an exercise. See 1.1.4.). Then clearly if $g^* = 0$, then $\hat{P}\hat{Q}f = \hat{P}(\hat{Q}f) = 0$. Otherwise

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

Since $-1 \le g/g^* \le 1$(because $g < g*$), we have

$$\sup_{x\in\mathcal{X}}|(\hat{P}[g/g^*])(x)| \le 2||\hat{P}|| = 2||P(x,\cdot) - \pi(\cdot)||$$

by part (b). Hence

$$\sup_{x\in\mathcal{X}}|(\hat{P}[g/g^*])(x)| \le 2\sup_{x\in\mathcal{X}}||P(x,\cdot) - \pi(\cdot)|| = t(n)$$

$$\Longrightarrow 2\sup_{x\in\mathcal{X}}|(\hat{P}\hat{Q}f)(x)| \le t(m)t(n)$$

Recall for every $x \in \mathcal{X}$ there exists some function f_x such that $\hat{P}\hat{Q}f_x(x) = ||\hat{H}(x,\cdot)|| = ||P^{m+n}(x,\cdot) - \pi(\cdot)||$(by part(a)). Then

$$2||P^{m+n}(x,\cdot) - \pi(\cdot)|| \le t(m)t(n)$$

for every x in $\mathcal{X}$. Therefore

$$2\sup_{x\in\mathcal{X}}||P^{m+n}(x,\cdot) - \pi(\cdot)|| \le t(m)t(n)$$

In other words

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

(f): According to (b),

$$||\mu(\cdot) - \nu(\cdot)|| = \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}\left|\int f d\mu - \int f d\nu\right| = \frac{1}{2}\sup_{f:\mathcal{X}\to[-1,1]}\left|\int f(g-h)d\rho\right|$$

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

Since $M = \max(g,h)$ and $m = \min(g,h)$, $M + m = g + h$. Hence $\int_{\mathcal{X}}(M+m)dp = \int_{\mathcal{X}}(g+h)dp = \int_{\mathcal{X}} dv_1 + \int_{\mathcal{X}} dv_2 = 1 + 1 = 2$.
Thus

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

(g): Let $a = \int_{\mathcal{X}} m dp,\ b = \int_{\mathcal{X}} (g-m) dp$ *and* $c = \int_{\mathcal{X}} (h-m) dp$. If any of a, b, c equals zero, the statement is trivial. To be more specific, if $a = 0$, then X and Y have disjoint supports. If $b = 0$ or $c = 0$, then set $X = Y = Z$, where $Z \sim \mu(\cdot)$. So assume they are all positive.

Construct Z, U, V, I such that Z has density m/a, U has density $(g-m)/b$, V has density $(h-m)/c$ and I is independent of Z, U, V.
Case 1: $I = 1,\ P(I = 1) = a,\ X = Y = Z$
Case 2: $I = 0,\ P(I = 0) = 1 - a,\ X = U,\ Y = V$
U, V have disjoint support so $P(U = V) = 0$. Thus

$$P(X = Y) = P(X = Y, I = 1) + P(X = Y, I = 0) = P(I = 1) + 0 = a = \int_X m dp = 1 - ||\mu(.) - \nu(.)||$$

It remains to show $X \sim \mu(\cdot),\ Y \sim \nu(\cdot)$. First note that $b = 1 - a$. It follows that

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

So $X \sim \mu(\cdot)$. Similarly $Y \sim \nu(\cdot)$ □

4 The eigenvalue connection

4.1 Introduction

Consider finite space $\mathcal{X}$. Since $\mu_k = \mu_0 P^k$, we do not want it to blow up. Naturally we should consider eigenvalue. Since left eigenvalues and right eigenvalues are the same (because eigenvalues of its transpose are the same)

4.2 The simplest non-trivial example

Settings

$$\mathcal{X} = \{0, 1\},\ \mu_0 = (1, 0)$$

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

Then

$$\mu_k(0) = \frac{q}{p+q} + (1 - \frac{q}{p+q})(1-p-q)^k$$

$$\mu_k(1) = \frac{p}{p+q} - (1 - \frac{q}{p+q})(1-p-q)^k$$

Observations

(1) Let $\pi = (\frac{q}{p+q}, \frac{p}{p+q})$. Assume $|1-p-q| < 1$, then

$$||\mu_k - \pi|| = |(\frac{q}{p+q})(1-p-q)^k| \to 0$$

(decrease exponentially quickly to 0, with rate governed by (1 - p - q).

(2) The limiting distribution π is a stationary distribution: $\pi P = P$, and thus corresponds to a left-eigenvector of the matrix P with eigenvalue 1. It is easily seen that any limiting distribution π for any Markov chain must be a stationary distribution(since $\mu_k = \mu_{k-1}P$)

(3) The convergence fails when $p = q \in \{0, 1\}$. If $p = q = 0$ the Markov chain is decomposible, meaning that the space $\mathcal{X}$ contains two-empty disjoint closed subsets. If $p = q = 1$ then this Markov chain is periodic. (different from class, we say it is periodic if the space contains disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_d$ such that for any $x \in \mathcal{X}_j$, $P(x, \mathcal{X}_{j+1} = 1)$). If our Markov chain is indecomposible and aperiodic, then it converges exponentially quickly.

(4) The eigenvalues of the matrix P are 1 and $1 - p - q$. We have a connection between trivial eigenvalues and non-trivial eigenvalues

(5) Define

$$\beta = \sum_y \min_x P(x, y)$$

Then $\beta = \min\{p+q, 2-p-q\}$. Then $1 - \beta = |1-p-q|$ is the absolute value of the non-trivial eigenvalue as above. The relationship can be explored via the method of "coupling"

(6) This Markov chain is reversible. It guarantees all eigenvalues will be real (so diagonalizable). But not all Markov chain has such property.

(7) When $p = q$, this corresponds to a simple random walk on group $Z/2Z$ with step distribution $Q(1) = p$ and $Q(0) = 1 - p$. Then $E_Q((-1)^x) = -p + (1-p) = 1 - p - q$. So for simple random walk on groups, the eigenvalues can be computed by taking expected values with respect to Q. This is a connection to Random walks in groups.

4.3 Fact 1

Any stochastic matrix P has an eigenvalue 1.

Proof. the vector u with $u_1 = \cdots = u_n = 1$ is a right-eigenvector corresponding to eigenvalue 1 of P. □

4.4 Fact 2

Suppose we have eigenvalues $\lambda_0, \ldots, \lambda_{n-1}$ such that $\lambda_0 = 1$. Consider $\lambda_* := \max\{\lambda_1, \ldots, \lambda_{n-1}\}$. Then $\lambda_* \leq 1$. Furthermore, if $P(x, y) > 0$ for all x, y, we have $\lambda_* < 1$

Proof. Let $v(x)$ be the biggest entry. Then we have

$$|\lambda v(x)| = (Pv)_x = |\sum_{y=1}^{n} P(x, y)v(y)| \leq |v(x)|$$

So $\lambda \leq 1$.
Suppose $P(x, y) > 0$, the equality holds only if $v(x)$ is a constant. But in this case it is v_0. Therefore $\lambda_* < 1$ □

4.5 Fact 3

Suppose P satisfies $\lambda_* < 1$. Then there is a unique stationary distribution π on $\mathcal{X}$ and, given an initial distribution μ_0 and any point $x \in \mathcal{X}$, there is a constant $C_x > 0$ such that

$$|\mu_k(x) - \pi(x)| \leq C_x k^{J-1} (\lambda_*)^{k-J+1}$$

If P is diagonalizable, we have

$$|\mu_k(x) - \pi(x)| \leq \sum_{m=1}^{n-1} |a_m v_m(x)||\lambda_m|^k \leq (\sum_{m=1}^{n-1} |a_m v_m(x)|)(\lambda_*)^k$$

where initial distribution can be expressed as linear combination of corresponding eigenvectors

$$\mu_0 = a_0 v_0 + \cdots + a_{n-1} v_{n-1}$$

If the eigenvectors v_i are orthogonal in $L^2(\pi)$, i.e. if $\sum_x v_i(x)\overline{v_j(x)}\pi(x) = \sigma_{ij}$, we get the further bound,

$$\sum_x |\mu_k(x) - \pi(x)|^2 \pi(x) = \sum_{m=1}^{n-1} |a_m|^2 |\lambda_m|^{2k} \leq (\sum_{m=1}^{n-1} |a_m|^2)(\lambda_*)^k$$

Interpretation: Now we easily get a new bound using eigenvalues. When $\lambda_* < 1$, $|\mu_k(x) - \pi(x)| \to 0$ as $k \to \infty$, the Markov Chain converges exponentially quickly. $\mu_k(x) \to a_0 v_0 = \pi(x)$. Since $\sum_x \pi(x) = 1$, $a_0 = (\sum_y v_0(y))^{-1}$. Indeed, the stationary distribution does

not depend on the initial distribution μ_0. Meanwhile, $\lambda_* < 1$ means the eigenvalue 1 has only one multiplicity, so the Markov chain has a unique stationary distribution π as the corresponding eigenvector. Conversely, if $\lambda_* = 1$, the eigenvalue 1 has at least 2 multiplicity, and also assume P is diagonalizable, then each multiplicity has a corresponding eigenvector, contradicts the property of unique stationary distribution.

4.6 Fact 4

A finite Markov chain satisfies $\lambda_* < 1$ if and only if it is both indecomposible and aperiodic

Proof. Necessity: assume it is decomposible, with disjoint subspaces $\mathcal{X}_1$ and $\mathcal{X}_2$. Then P is like

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Then there are two vectors with eigenvalue 1
Assume it is periodic. Then the matrix is like

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Let

$$v = (e^{2\pi i/3}, e^{4\pi i/3}, 1)$$

Then $e^{2\pi i/d}$ is an eigenvalue.(Intuitively, the matrix is a permutation, which moves x_i to x_{j+1}, so we can just make $\frac{v(x_{j+1})}{v(x_j)}$ to be a fixed number)(can not pick a random ratio because we want $v(x_1)/v(x_n) = r$) In this case $\lambda_* = 1$
Sufficiency: 2.4.4. If a chain is indecomposible and aperiodic, then for any state $i, j \in S$, there is a $n_0(i, j)$ such that $p_{ij}^n > 0$ for any $n > n_0$. Find the biggest n_0. Then P^n has all positive entries. □

Note: Irreducible is a stronger property than indecomposible. Here is an example of indecomposible Markov chain but not irreducible:

$$\begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Next, we move to General State Space Markov Chain.

5 Asymptotic Convergence

For a Markov chain to converge to stationarity, several conditions are required.

Definition. 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 $P^n(x, A) > 0$.

Intuitively, a chain is said to be *ϕ-irreducible*, if there exists a probability measure ϕ and with this measure we can get from x to A in finite steps for all x. The reason why *ϕ-irreducible* is used instead of *irreducible* as in discrete case is that the probability of getting from one state to another state is zero in continuous state space.

Running Example. Here we give an example of a ϕ-irreducible Markov chain defined by the Metropolis-Hastings Algorithm.

Suppose that $\pi(\cdot)$ is a probability measure having unnormalised density function π_u with proposal density $q(\mathbf{x}, \cdot)$ with respect to d-dimensional Lebesgue measure. Then if $q(\cdot, \cdot)$ is positive and continuous on $\mathbf{R}^d \times \mathbf{R}^d$, and π_u is positive everywhere, then the algorithm is π-irreducible (simply follows from the definition of π-irreducible). Then we have $\pi(A) > 0$, and specifically there exists a $R > 0$ such that $\pi(A_R) > 0$, where $A_R = A \cap B_R(0)$, and $B_R(0)$ represents the ball of radius R centered at $\mathbf{0}$. Then by continuity, for any $\mathbf{x} \in \mathbf{R}^d, \inf_{\mathbf{y} \in A_R} \min\{q(\mathbf{x}, \mathbf{y}), q(\mathbf{y}, \mathbf{x})\} \geq \epsilon$ for some $\epsilon > 0$, this implies that $q(\mathbf{x}, \mathbf{y}) \geq \epsilon$ and $q(\mathbf{y}, \mathbf{x}) \geq \epsilon$. Therefore, we have that

$$
\begin{aligned}
P(\mathbf{x}, A) \geq P(\mathbf{x}, A_R) &= P(\mathbf{x}, A_R | \, \textbf{accept}) + P(\mathbf{x}, A_R | \, \textbf{reject}) \\
&\geq \int_{A_R} q(\mathbf{x}, \mathbf{y}) \min\left[1, \frac{\pi_u(\mathbf{y}) q(\mathbf{y}, \mathbf{x})}{\pi_u(\mathbf{x}) q(\mathbf{x}, \mathbf{y})}\right] d\mathbf{y} \\
&= \int_{A_R} \min\left[q(\mathbf{x}, \mathbf{y}), \frac{\pi_u(\mathbf{y})}{\pi_u(\mathbf{x})} q(\mathbf{y}, \mathbf{x})\right] d\mathbf{y} \\
&\geq \epsilon \int_{A_R} \min\left[1, \frac{\pi_u(\mathbf{y})}{\pi_u(\mathbf{x})}\right] d\mathbf{y} \\
&= \epsilon \int_{A_1} 1 \, d\mathbf{y} + \epsilon \int_{A_2} \frac{\pi_u(\mathbf{y})}{\pi_u(\mathbf{x})} d\mathbf{y} \\
&\geq \epsilon \text{Leb}(A_1) + \frac{\epsilon}{\pi_u(\mathbf{x})} \int_{A_2} \pi_u(\mathbf{y}) d\mathbf{y} \\
&= \epsilon \text{Leb}(A_1) + \frac{\epsilon K}{\pi_u(\mathbf{x})} \pi(A_2)
\end{aligned}
$$

where $K = \int_{\mathcal{X}} \pi_u(y) dy > 0$, $A_1 = \{\mathbf{y} \in A_R : \pi_u(\mathbf{y}) \geq \pi_u(\mathbf{x})\}$, $A_2 = \{\mathbf{y} \in A_R : \pi_u(\mathbf{y}) < \pi_u(\mathbf{x})\}$ (The last equality follows from the definition of π_u : $\pi(A) = \frac{\int_A \pi(y) dy}{K}$ for any

measurable set $A \subset \mathcal{X}$.)

To see $P(x, A) > 0$, it suffices to show $\text{Leb}(A_1)$ and $\pi(A_2)$ cannot both be 0. Suppose by contrary, that $\text{Leb}(A_1) = \pi(A_2) = 0$. Then since $\pi(\cdot)$ is absolutely continuous with respect to Lebesgue measure, $\pi(A_1) = 0$. This contradicts the fact that $\pi(A_R) = \pi(A_1) + \pi(A_2) > 0$. Therefore $P(x, A) > 0$, and the chain is π-irreducible

Example 2. Periodic example. Suppose $\mathcal{X} = \{1, 2, 3\}$ with $\pi\{1\} = \pi\{2\} = \pi\{3\} = 1/3$. Let $P(1, \{2\}) = P(2, \{3\}) = P(3, \{1\}) = 1$. Then $\pi(.)$ is stationary, and the chain is ϕ-irreducible (any state is reachable from all states). However, if $X_0 = 1$, then $X_n = 1$ whenever n is a multiple of 3, so $P(X_n) = 1$ oscillates between 0 and 1, it doesn't converge to 1/3.

To avoid this problem, we require *aperiodicity*.

Definition. A Markov chain with stationary distribution $\pi(\cdot)$ is *aperiodic* if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d \subseteq \mathcal{X}$ with $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i (1 \leq i \leq d-1)$, 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 equal to the largest such value of d, and corresponding periodic decomposition $\mathcal{X}_1, \ldots, \mathcal{X}_d$.)

It is reasonable to ask if we can define this aperiodicity without assuming the existence of stationary distribution. The answer is actually yes. There is another definition of aperiodicity using maximal irreducible measure (usually denoted by $\psi(\cdot)$), where the measure $\psi(\cdot)$ is "maximal" in the sense that $\phi \ll \psi$ for any probability measure $\phi(\cdot)$ on $\mathcal{X}$ which makes P $\phi-$irreducible. It might be surprising that such ψ exists for any $\phi-$irreducible chain. Then this allows us to define aperiodicity without using $\pi(\cdot)$, by requiring $\psi((\bigcup \mathcal{X}_i)^c) = 0$ instead of $\pi(\mathcal{X}_1) = 0$. To see the equivalence of these two definitions, see [1].

Running Example Continued. There is no additional assumptions necessary to ensure aperiodicity. To see this, suppose that $\mathcal{X}_1$ and $\mathcal{X}_2$ are disjoint subsets of $\mathcal{X}$ both of positive π measure, with $P(x, \mathcal{X}_2) = 1$ for all $x \in \mathcal{X}_1$. But just take any $x \in \mathcal{X}_1$, then since $\mathcal{X}_1$ must have positive Lebesgue measure

$$P(x, \mathcal{X}_1) \geq \int_{y \in X_1} q(x, y)\alpha(x, y)dy > 0$$

This contradicts $P(x, \mathcal{X}_1) \leq 1 - P(x, \mathcal{X}_2) = 0$. Therefore aperiodicity must hold.(It is rather rare for MCMC algorithms to be periodic, as aperiodicity)

Now we can state the main asymptotic convergence theorem, which assumes that the state space's σ-algebra is *countably generated*, but this is a very weak assumption which is true for e.g. any countable state space, or any subset of R^d with the usual Borel σ-algebra.

Theorem 4. If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(.)$, then for $\pi - a.e.\ x \in \mathcal{X}$.

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

In particular, $\lim_{n\to\infty} P^n(x,A) = \pi(A)$ *for all measurable* $A \subseteq \mathcal{X}$
Remark: Sometimes π-a.e. might not be as strong as it looks like. We will discuss this problem in Example 3.

Fact: If Markov Chain is ϕ irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, if $h : \mathcal{X} \to R$ with $\pi(|h|) < \infty$, then a "strong law of large numbers" also holds:

$$P[\lim_{n\to\infty} (1/n) \sum_{i}^{n} h(X_i) = \pi(h)] = 1$$

Recall that the statement of the classical strong law of large numbers theorem is: For infinite i.i.d. sequence X_i of copies of X, with $|E(x)| < \infty$. Then $P[\lim_{n\to\infty} \bar{X}_n = E(X)] = 1$. Here, we consider positive recurrent chain possessing invariant probability π, the "Strong LLN" states that $(1/n)\sum_i^n h(X_i)$ converges to $\pi(h) = E_\pi(h(X))$, the steady state expectation of $h(X)$.

Example 3. This example illustrates the importance of π-a.e for Theorem 4. Let $\mathcal{X} = 1, 2, ...$ Let $P(1,1) = 1$, and for $x \geq 2. P(x,1) = 1/x^2$ and $P(x, x+1) = 1 - (1/x^2)$. Then chain has stationary distribution $\pi(\cdot) = \delta_1(\cdot)$, that is for state 1, $\pi_1 = 1$ and other state $\pi_i = 0$ for $i \geq 2$, and this chain is π-irreducible and aperiodic. On the other hand, if we do not start from 1, then $P(X_n = x + n \text{ for all n}) = \Pi_{j=x}^{\infty}(1 - (1/j^2)) > 0$, so that $\lim_{n\to\infty} ||P^n(x,.) - \pi(.)|| \neq 0$. To check this *infinite product* $\Pi_{j=x}^{\infty} 1 - (1/j^2) < 1$ does not converge to 0, suppose j is large, then $1 - (1/j^2) \approx e^{-1/j^2}$ (since for small x, $e^x \approx 1 + x$ by Taylor's expansion). The trick is that j can increase, the whole product is not an infinite product of a constant. $\Pi(1 - 1/j^2) = e^{-\sum 1/j^2} > 0$ since exponential functions are always positive, and $\sum 1/j^2$ is convergent. This is also why we construct this chain by $P(x,1) = 1/x^2$, instead of $1/x$. (See Exercise 1.2.1.) In conclusion, Theorem 4 do holds for π-a.e. $x \in \mathcal{X}$, but does not hold for $x \geq 2$. This Markov chain is not Harris Recurrent because it doesn't go to another state with probability 1, it may have chance to reach far (big number state), but with only positive probability.

Harris Recurrent Markov Chain: Note that MCMC algorithms are created precisely so that $\pi(\cdot)$ is stationary. And obviously ϕ irreducible and aperiodic, so it applies to theorem well (for π - $a.e$). Furthermore, Metropolis algorithm holds the conclusion of the theorem for all $x \in \mathcal{X}$, not just π- $a.e.$ Since the transition kernels $P(x,\cdot)$ with proposal distribution

$Q(x,\cdot)$ are all absolutely continuous with respect to $\pi(\cdot)$. It is also *Harris recurrent*: for all $B \subseteq \mathcal{X}$ with $\pi(B) > 0$, $P[\exists n : X_n \in B | X_0 = x] = 1$. Equivalently, for discrete Markov Chain, Harris recurrent holds *iff* there exists $y \in \mathcal{X}$ such that $P_x[\tau(y) < \infty] = 1$ for some y, $\tau(y) = \inf(n \geqslant 1 : X_n = y)$. This implies that the chain for sure will eventually reach B/y from x. Harris Recurrence is a stronger condition than π irreduciblity, and it can be proved that Harris recurrent Markov chain satisfies minorization condition as we'll discuss later.

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

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

and also the "strong law of large number" still holds without change.

Proof. Let this Markov Chain P have periodic decomposition $\mathcal{X}_1, ..., \mathcal{X}_d \subseteq \mathcal{X}$. And we define a new Chain P' as a d-step chain P^d restrict to $\mathcal{X}_1$. Then it is obvious that P' is ϕ-irreducible and aperiodic on $\mathcal{X}_1$. And now we define $\pi'(A) = d\pi(A \cap \mathcal{X}_1)$ for all measurable $A \subseteq \mathcal{X}$. Because with disjoint $\mathcal{X}_1, ..., \mathcal{X}_d \subseteq \mathcal{X}$, for all $i = 1, .., d-1, P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$. And we observe that

$$\begin{aligned}
\pi(\mathcal{X}_2) &= \int_{\mathcal{X}} \pi(dx) P(x, \mathcal{X}_2) \\
&\leq \int_{\mathcal{X}_1} \pi(dx) P(x, \mathcal{X}_2) \\
&= \pi(\mathcal{X}_1)
\end{aligned}$$

Similarly, we have $\pi(\mathcal{X}_1) \leq \pi(\mathcal{X}_2) \leq \cdots \leq \pi(\mathcal{X}_d) \leq \pi(\mathcal{X}_1)$. Therefore we can conclude that $\pi(\mathcal{X}_1) = \cdots = \pi(\mathcal{X}_d) = 1/d$.
Since $\pi'(\mathcal{X}_1) = d\pi(\mathcal{X}_1) = d * (1/d) = 1$, π' is a probability measure. Now let $A \subseteq \mathcal{X}_1$, we have

$$\begin{aligned}
\int_{\mathcal{X}_1} \pi'(dx) P^d(x, A) &= \int_{\mathcal{X}_1} d\pi(dx) P^d(x, A) \\
&= d \int_{\mathcal{X}_1} \pi(dx) P^d(x, A) = d\pi(A) = \pi'(A)
\end{aligned}$$

Therefore π' is stationary for P'. Now we observe that for any measurable $A \subseteq \mathcal{X}$,

$$\frac{1}{d} \sum_{j=0}^{d-1} (\pi' P^j * P)(A) = \frac{1}{d} \sum_{j=1}^{d} (\pi' P^j)(A) = \frac{1}{d} \sum_{j=0}^{d-1} (\pi' P^j)(A) + \pi'(A) = \frac{1}{d} \sum_{j=0}^{d-1} (\pi' P^j)(A)$$

And since

$$\frac{1}{d}\sum_{j=0}^{d-1}(\pi' P^j)(\mathcal{X}) = \frac{1}{d} d \sum_{j=0}^{d-1}(\pi P^j)(\mathcal{X}_1) = d\pi(\mathcal{X}_1) = d * \frac{1}{d} = 1$$

Therefore by uniqueness, it is true that π' satisfy that $\pi(\cdot) = (1/d)\sum_{j=0}^{d-1}(\pi' P^j)(\cdot)$. Now without loss of generality, we assume $x \in \mathcal{X}_1$. And let $n = md$ with $m \to \infty$, then for $j \in \mathbb{N}$, by **Proposition 3(d)**, we have

$$\left\| P^{md+j}(x,\cdot) - (\pi' P^j)(\cdot) \right\| \leq \left\| P^{md}(x,\cdot) - \pi'(\cdot) \right\|$$

Then we have

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

Now we apply **Theorem 4** to P', we then obtain that $\lim_{m\to\infty} \left\| P^{md}(x,\cdot) - \pi'(\cdot) \right\| = 0$. for π'-a.e. $x \in \mathcal{X}_1$. By replacing md with n, we have

$$\lim_{n\to\infty} \left\| (1/d) \sum_{i=n}^{n+d-1} P^i(x,\cdot) - \pi(\cdot) \right\| \leq (1/d) \sum \lim_{m\to\infty} \left\| P^{md}(x,\cdot) - \pi'(\cdot) \right\| = 0$$

Furthermore, let $h : \mathcal{X} \to \mathbb{R}$, with $\pi(|h|) < \infty$, and let $\bar{P}$ be the transit kernel of $\mathcal{X}_1 \times ... \times \mathcal{X}_d$ which corresponding to the sequence $\{(X_{md}, X_{md+1}..., X_{md+d-1})\}_{m=0}^{\infty}$. Now let $\bar{h}(x_0, ..., x_{d-1}) = (1/d)\sum_{j=0}^{d-1} h(x_j)$. Similar to P', we can see that $\bar{P}$ is ϕ-irreducible and aperiodic since each components in the product are so. And we have the stationary distribution

$$\bar{\pi} = \pi' \times (\pi' P) \times ... \times (\pi' P^{d-1})$$

Therefore with the assumption of h, by applying the above **Fact** to $\bar{P}$ and $\bar{h}$, we can establish that

$$\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} \bar{h}(X_{id}, X_{id+1}..., X_{id+d-1}) = \bar{\pi}(\bar{h}) \quad w.p.\ 1.$$

□

Remark. If we have a irreducible Markov Chain with *finite* state space, we do not need the assumption of periodicity to establish above corollary.

To consider the quantitative rate of convergence on $\left\|(1/n)\sum_{i=1}^{n} P^1(x,\cdot) - \pi(\cdot)\right\|$, we need to use *shift-coupling* technique.

6 Uniform Ergodicity

Definition. A Markov chain with stationary $\pi(.)$ is *uniformly ergodic* if

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

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

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

Proof. If the Markov chain is uniformly ergodic, it is very easy to see that since $\rho < 1, M < \infty$, there will be a n large enough such that:

$$||P^n(x,.) - \pi(.)|| \leq M\rho^n < \frac{1}{2}$$

for all $x \in \mathcal{X}$.
Now, assume $\sup_{x\in\mathcal{X}} ||P^n(x,.) - \pi(.)|| < \frac{1}{2}$ for some $n \in \mathbb{N}$.
Define $T(n) = 2\sup_{x\in\mathcal{X}} ||P^n(x,.) - \pi(.)|| = \beta < 1$. Then by Proposition 3(e), $T(n)$ satisfies the sub-multiplicity property such that for all $j \in \mathbb{N}$, $T(jn) \leq (T(n))^j = \beta^j$. Furthermore, setting $j = \lfloor \frac{m}{n} \rfloor$ and using Proposition 3(c):

$$||P^m(x,.) - \pi(.)|| \leq ||P^{(jn)}(x,.) - \pi(.)|| \leq \frac{1}{2}T(jn) \leq (T(n))^j \leq \beta^j \leq \beta^{-1}(\beta^{\frac{1}{n}})^m$$

with $T(jn) = 2\sup_{x\in\mathcal{X}} ||P^n(x,.) - \pi(.)||$. Then the chain is uniformly ergodic with $M = \beta^{-1}$ and $\rho = \beta^{\frac{1}{n}}$. □

Remark. The same statement is true when $\frac{1}{2}$ is substituted with any $0 < \delta < \frac{1}{2}$ as we can apply the same proof above.

In order to develop further conditions on uniform ergodicity, we need some definitions.

Definition. A subset $C \subseteq \mathcal{X}$ is *small* (or (n_0, ϵ_{n_0}, v)-small) if there exists a $n_0 \in \mathbb{Z}, \epsilon > 0$, and a probability measure $v(.)$ on $\mathcal{X}$ such that the following *minorization condition* holds:

$$P^{n_0}(x,.) \geq \epsilon v(.) \quad x \in C$$

Remark. Some other versions of this definition also require that C have positive stationary measure. However, we do not require this since $\pi(C) > 0$ will be satisfied under the assumption of the drift condition considered in the next section.

Intuitively, we can think of this n_0-step transition will have a ϵ-overlap in probability. Now, realize that if $\mathcal{X}$ is countable, and if

$$\sum_{y \in \mathcal{X}} \inf_{x \in C} P^{n_0}(x, \{y\}) > 0,$$

then we can set $\epsilon_{n_0} = \sum_{y \in \mathcal{X}} \inf_{x \in C} P^{n_0}(x, \{y\})$, and for $y \in \mathcal{X}$, let $v(\{y\}) = \epsilon_{n_0}^{-1} \inf_{x \in C} P^{n_0}(x, \{y\})$, it's easy to see that C is (n_0, ϵ_{n_0}, v)-small. This fact is useful in that for an irreducible (or simply indecomposible) and aperiodic chain on a finite state space, we will have $\epsilon_{n_0} > 0$ for sufficiently large n_0. Generalizing to a continuous space, if the transition probabilities have densities with respect to some measure $\eta(.)$, i.e. if $P^{n_0}(x, dy)$, then we can take $\epsilon_{n_0} = \int_{y \in \mathcal{X}} (\inf_{x \in \mathcal{X}} p_{n_0}(x, y)) \eta(dy)$. This leads us to the concluding theorem of this section, whose proof uses a technique called coupling, which will be introduced in the next section.

Theorem 8. Consider a Markov chain with invariant probability distribution $\pi(.)$. Suppose $\mathcal{X}$ is (n_0, ϵ_{n_0}, v)-small, then the chain is uniformly ergodic. Furthermore,

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

Therefore, Theorem 8 allows us to bound the distance to stationarity quantitatively!

Definition. A subset $C \subseteq \mathcal{X}$ is *pseudo-small* if there exists a $n_0 \in \mathbb{Z}$, $\epsilon > 0$, and a probability measure $v_{xy}(.)$ on $\mathcal{X}$ such that for all $x, y \in C$, $P^{n_0}(x, .) \geq \epsilon v_{xy}(.)$ and $P^{n_0}(y, .) \geq \epsilon v_{xy}(.)$.

Remark. Theorem 8 will remain true if the assumption of the state space being small is replaced by pseudo-small because we will use pairwise coupling in our proofs.

Running example, continued. In the previous section, we applied some of our results to a MCMC running example. Now, we look at whether we can apply the definition of small to it. It turns out that in the general set up of our running example, all compact sets on which π_u is bounded are small. To see this, suppose C is a compact set on which π is bounded by $k < \infty$. Take any point $\mathbf{x}$ in C, and let D be any compact set with positive π and Lebesgue measure. Let $\epsilon = \inf_{\mathbf{a}, \mathbf{b} \in C \cup D} q(\mathbf{a}, \mathbf{b})$, $M = \sup_{\mathbf{x} \in C, \mathbf{y} \in D} q(\mathbf{x}, \mathbf{y})$. Then, we

can conclude that for any $\mathbf{x} \in C$, $\mathbf{y} \in \mathcal{X}$, $M \geq q(\mathbf{x}, \mathbf{y}) \geq \epsilon$, $q(\mathbf{y}, \mathbf{x}) \geq \epsilon$, and $\pi(\mathbf{x}) \leq k$. Then

$$P(\mathbf{x}, d\mathbf{y}) \geq q(\mathbf{x}, \mathbf{y}) \min\{1, \frac{\pi_u(\mathbf{y}) q(\mathbf{y}, \mathbf{x})}{\pi_u(\mathbf{x}) q(\mathbf{x}, \mathbf{y})}\} \geq \epsilon \min\{1, \frac{\pi_u(\mathbf{y})\epsilon}{Mk}\}$$

Thus for any $y \in \mathcal{X}$,

$$P(\mathbf{x}, d\mathbf{y}) \geq \epsilon \mathbf{1}_D \min\{1, \frac{\pi_u(\mathbf{y})\epsilon}{Mk}\}$$

We see that the right hand side is a positive measure that does not depend on $\mathbf{x}$, so the compact set C is small, and we can easily bound the convergence rate.

However, most MCMC algorithms are not uniformly ergodic, so we will have to use a more general theorem to bound their rates of convergence.

7 Geometric Ergodicity

A weaker condition than uniform ergodicity.
Definition. A Markov chain with stationary distribution $\pi(.)$ is geometrically ergodic if

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

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

If $\mathcal{X}$ is finite, then all irreducible and aperiodic Markov chains are geometrically (in fact, uniformly) ergodic. If $\mathcal{X}$ is infinite, however, this is not always the case. So it is reasonable to question that in which case the geometric ergodicity is guaranteed. (Spoiler alert: we will assume two more conditions, minorization condition and drift condition)

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

Definition. 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:

$$PV \leq \lambda V + b\mathbf{1}_C,$$

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

The main result guaranteeing geometric ergodicity is:

Theorem 9. Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(.)$. Suppose the minorisation condition is satisfied for some $C \subset \mathcal{X}$ and $\epsilon > 0$ and probability measure $v(.)$. Suppose further that the drift condition is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, 0]$ for at least one (and hence for π-a.e.) $x \in \mathcal{X}$. Then the chain is geometrically ergodic.
(why these conditions can guarantee geometric ergodicity?)

Fact 10. In fact, it follows from Theorems 15.0.1, 16.0.1, and 14.3.7 of Meyn and Tweedie[1], and Proposition 1 of [2], that the minorisation condition and drift condition of Theorem 9 are equivalent (assuming ϕ-irriducibility and aperiodicity) to the apparently stronger property of "V-uniform ergodicity" i.e. that there is $C < \infty$ and $p < 1$ such that

$$\sup_{|f| \leq V} |P^n f(x) - \pi(f)| \leq CV(x)p^n, \ x \in \mathcal{X}$$

Example 4. This is a example that demonstrates the geometric ergodicity of Metropolis algorithms on R (section 2.1 of the paper). Let $\mathcal{X} = \mathbb{R}^+$ and $\pi_u(x) = e^{-x}$. Let the proposal distribution $q(x, y) = q(|y - x|)$ with support contained in $[x - a, x + a]$. Let $V(x) = e^{cx}$, for some $c > 0$, be the drift function, and the proposal distribution $q(x, y) = q(|y - x|)$ be symmetric. For $x \geq a$, Note

$$\alpha(x, y) = \min\{1, \frac{\pi_u(y)q(y, x)}{\pi_u(x)q(x, y)}\} = \min\{1, \frac{\pi_u(y)}{\pi_u(x)}\} \text{ (since } q \text{ is symmetric)}$$

Then since $\pi_u(x, y) = e^{-x}$ is decreasing,

$$\alpha(x, y) = \begin{cases} 1, \ y < x \\ \frac{\pi_u(y)}{\pi_u(x)}, y \geq x \end{cases}$$

Then we can try to piece together a drift condition by:

$$\begin{aligned}
PV(x) &= \int_{x-a}^{x+a} V(y)P(x,dy) + V(x)\int_{x-a}^{x+a}(1-P(x,dy)) \\
&= \int_{x-a}^{x+a} V(y)q(x,y)\alpha(x,y)dy + \int_{x-a}^{x+a} V(x)q(x,y)(1-\alpha(x,y))dy \\
&= (\int_{x-a}^{x} V(y)q(x,y)dy + \int_{x}^{x+a} V(y)q(x,y)\frac{\pi_u(y)}{\pi_u(x)}dy) \\
&+ (0 + \int_{x}^{x+a} V(x)q(x,y)(1-\frac{\pi_u(y)}{\pi_u(x)})dy) \\
&= \underbrace{\int_{x}^{x+a} V(2x-y)q(x,y)dy}_{\text{change of variable}} + \int_{x}^{x+a} V(y)q(x,y)\frac{\pi_u(y)}{\pi_u(x)}dy \\
&+ \int_{x}^{x+a} V(x)q(x,y)(1-\frac{\pi_u(y)}{\pi_u(x)})dy \\
&= \int_{x}^{x+a} q(x,y)[\frac{V(y)\pi_u(y)}{\pi_u(x)} + V(2x-y) + V(x)(1-\frac{\pi_u(y)}{\pi_u(x)})]dy
\end{aligned}$$

Now we try to bound the equation inside the integral, let $u = y - x$:

$$\begin{aligned}
\frac{V(y)\pi_u(y)}{\pi_u(x)} + V(2x-y) + V(x)(1-\frac{\pi_u(y)}{\pi_u(x)}) &= e^{cx}[e^{(c-1)u} + e^{-cu} + 1 - e^{-u}] \\
&= 2(1 - \frac{(1+e^{(c-1)u})(1-e^{-cu})}{2})e^{cx} \\
&= 2(1-\epsilon)V(x)
\end{aligned}$$

where $\epsilon = \frac{(1+e^{(c-1)u})(1-e^{-cu})}{2}$. Notice if $c < 1$, then $0 < \epsilon < 1$. We then take a particular $\epsilon \in (0,1)$ such that

$$\frac{V(y)\pi_u(y)}{\pi_u(x)} + V(2x-y) + V(x)(1-\frac{\pi_u(y)}{\pi_u(x)}) \leq 2(1-\epsilon)V(x)$$

for any u (this is possible as u has a compact support), so that

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

Furthermore, it is easy to show that $PV(x)$ is bounded on $[0,a]$, and since $[0,a]$ is compact, it is a small set. Thus we've shown that both the minorization condition and drift condition holds. The algorithm is geometrically ergodic by Theorem 9.

Remark. Geometric ergodicity is often an useful property; however, it doesn't prove to be too helpful in certain situations.

Example 5. Consider an independent sampler, with $\pi(.)$ an Exponential(1) distribution, and $Q(x,.)$ an Exponential(λ). If we run the MCMC with $0 < \lambda \leq 1$, then the sampler is geometrically ergodic (as shown above with the case $\lambda = 1$), has central limit theorems, and have nice properties. However, if $\lambda > 1$, the sample is no longer geometrically ergodic. In fact, simulations indicate that when $\lambda = 5$, when started at stationarity and averaged over the first million iterations, the sampler will miss return a value that typically misses the target and occasionally becomes very large, leading to very unstable behaviour. This is a scenario where the property of geometric ergodicity is useful in telling us the convergence behaviours of MCMC algorithms.

Example 6. Let $\mathcal{X} = [0,1]$, $\delta = 10^{-100}$, $0 < a < 1 - \delta$, $\pi_u(\mathbf{x}) = \delta + \mathbf{1}_{[a,a+\delta]}(x)$. Then:

$$\begin{aligned}\pi([a, a+\delta]) &= \frac{\int_{[a,a+\delta]} \pi_u(x)dx}{\int_{[0,1]} \pi_u(x)dx} \\ &\approx \frac{\delta}{2\delta} = \frac{1}{2}\end{aligned}$$

Now consider an algorithm with an uniformly distributed proposal function. Since the interval $[a, a+\delta]$ is really small, unless the sampler gets really lucky, the result will appear as converging to Uniform($\mathcal{X}$) distribution. Nevertheless, this algorithm is still geometrically ergodic (even uniformly ergodic since it's irreducible, aperiodic, and the state space is finite).

Example 7. Let $\mathcal{X} = \mathbb{R}, \pi_u(x) = \frac{1}{1+x^2}$. Then a random walk Metropolis algorithm with proposal function distributed Uniform$[x-1, x+1]$ is ergoidc but is not geometrically ergodic. However, is we truncate the tail of π_u to $\pi'_u(x) = \pi_u(x)\mathbf{1}_{|x|\leq 10^{100}}$, then the algorithm is geometrically and uniformly ergodic. However, the two algorithms are essentially the same, and in this case geometric ergodicity don't tell us much at all.

These examples demonstrates the limitation of qualitative convergence properties and forces us to come up with more quantitative bounds on Markov chain convergence.

8 Quantitative Convergence Rates

Seeing the limitations of uniform ergodicity and geometric ergodicity, we begin to search for a *quantitative* bound on convergence rates, i.e. $||P^n(x,.) - \pi(.)|| \leq g(x,n)$ for some *explicit* function $g(x,n)$, which is small for large n (hopefully), so we can bound convergence rate efficiently.

Definition. Bivariant Drift condition:

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

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

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

($\bar{P}$ represents running two independent copies of the chain) This equation of $\bar{P}$ represents the expectation of $h(\cdot,\cdot)$ after one move, starting from the point (x,y). The bivariate drift condition can be understood as: if the pair of Markov chains is both outside of the pair of small sets $C \times C$ (the region where they can couple), then both chains will drift back towards it. The expression seems very difficult to compute, but there is actually a connection between univariate drift condition introduced above, which is easier to find, and the bivariant drift condition:

Proposition 11. Suppose the univariant drift condition is satisified for some $V : \mathcal{X} \to [1,\infty]$, $C \in \mathcal{X}$, $0 < \lambda < 1$ and $b < \infty$. Let $d = \inf_{x \in C^c} V(x)$. If $c = d > [b/(1-\lambda)] - 1$, then the bivariate drift condition is satisified for the same C, with $h(x,y) = \frac{1}{2}[V(x) + V(y)]$ and $\alpha^{-1} = \lambda + b/(d+1) < 1$. That is $\bar{P}h(x,y) \leq \frac{1}{2}[V(x) + V(y)][\lambda + b/(d+1)]$.

Proof. For drift condition, consider if $(x,y) \notin C \times C$, then either $x \notin C$, or $y \notin C$, or both. In all cases, since $d = \inf_{x \in C^c} V(x)$, and at most one chain out of the pair (x,y) is in C, we can bound $h(x,y) \geq (1+d)/2$. Since the univariant drift condition holds, $PV \leq \lambda V + b1_C$, then $PV(x) + PV(y) \leq \lambda V(x) + \lambda V(y) + b$:

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

□

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

$$\bar{R}h(x,y) \equiv \int_{\mathcal{X}} \int_{\mathcal{X}} (1-\epsilon)^{-2} h(z,w)(P(x,dz) - \epsilon v(dz))(P(y,dw) - \epsilon v(dw))$$

The expression of $\bar{R}h$ could be interpreted as the residual expectation, or the expectation of h in the event that two chains fail to couple with probability ϵ. In terms of these assumptions made above, we can state our main result for a qualitative bound on the convergence rate:

Theorem 12. Consider a Markov chain on a state space $\mathcal{X}$, having transition kernel P. Suppose there is $C \subset \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty))$, a probability distribution $v(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, and $\epsilon > 0$, B as defined above. If these conditions in the section all hold, Then for any joint initial distribution $\mathcal{L}(X_0, X_0')$, and any integer $1 \leq j \leq k$, if $\{X_n\}$ and $\{X_n'\}$ are two copies of the Markov chain started in the joint distribution $\mathcal{L}(X_0, X_0')$, then

$$||\mathcal{L}(X_k) - \mathcal{L}(X_k')||_{TV} \leq (1-\epsilon)^j + \alpha^{-k} B_{n0}^{j-1} E[h(X_0, X_0')].$$

The formal proof will be given at next section using coupling constructions.

Convergence Diagnostics In many complicated Markov Chain, it may be difficult to satisfy the condition of theorem 12 and apply successfully. In such cases, MCMC practitioners will instead use "convergence diagnostics". In statistic, diagnostics involve doing statistical analysis of the realised output $X_1, X_2, ...$, to check if the distribution of X_n appear to be stable for large enough n. This is like "Law of Large numbers", and in idea of Monte Carlo Simulation. For MCMC, we Run the Markov chain repeatedly from different initial states, and check if the chain all converge to approximately the same distribution. In practice, this works well, but also it can introduce bias into the resulting estimates.

9 Convergence Proofs using Coupling Constructions.

9.1 The Coupling Inequality.

Suppose we have two *random variables* X and Y which are defined jointly on a common state space $\mathcal{X}$, their probability distribution $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ have property that

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

Therefore we have

$$\|\mathcal{L}(X) - \mathcal{L}(Y)\| \leq P(X \neq Y) \tag{1}$$

9.2 The Coupling Loop on Small Set

Suppose now that C is a small set. We shall use the following coupling construction by using *splitting technique*. We will run two copies $\{X_n\}$ and $\{X_n'\}$ of the Markov Chain. And we update them as follow.

To start with splitting technique, we let $X_0 = x$ and $X_0' \sim \pi(\cdot)$. And $n = 0$, then we repeat the following loop.

Beginning of the loop. Given X_n and X_n',

1. If $X_n = X_n'$, we 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 w.p. ϵ, we choose X_{n+n_0} and X_{n+n_0}' and w.p. $1-\epsilon$, we conditionally independently choose (since we have minorisation condition)

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

If $n_0 > 1$, for completeness, we construct $X_{n+1}, \ldots, X_{n+n_0-1}$ from their correct conditional distributions given X_0 and X_{n+n_0}. And similarly for $X_{n+1}', \ldots, X_{n+n_0-1}'$. And we 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 $n+1$.

Then we return to the beginning of the loop.

The above construction let two chains $\{X_n\}$ and $\{X_n'\}$ update over transit kernel P. So we have,

$$P[X_n \in A] = P^n(x, \cdot); \quad P[X_n'] = \pi(A) \text{ for all } n$$

We can also see that the two chains are run independently until they both enter C. And we will provide way later to obtain a better bound on $\|P^n(X_n, \cdot) - \pi(\cdot)\|$ other than the result from coupling inequality.

9.2.1 Proof of Theorem 8.

Recall Theorem 8: Consider a Markov chain with invariant probability distribution $\pi(.)$. Suppose $\mathcal{X}$ is (n_0, ϵ_{n_0}, v)-small, then the chain is uniformly ergodic. Furthermore,

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

Proof. In this case, we have $C = \mathcal{X}$. Therefore we will only consider step 2 of the above loop. Since for every n_0, we have at least the probability of ϵ to make $X_n = X_n'$. Consider a n_0-step chain, by coupling inequality, let $n = n_0 m$ (have m chances to couple), so

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

Now for any n, we have $\lfloor n/n_0 \rfloor \leq m$, let $z = n_0 \lfloor n/n_0 \rfloor$, so $z \leq n$. And by Proposition 3(c), we have that

$$\|P^n(X_n, \cdot) - \pi(\cdot)\| \leq \|P^z(X_n, \cdot) - \pi(\cdot)\| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$$

□

9.3 Proof of Theorem 12.

Recall Theorem 12: Consider a Markov chain on a state space $\mathcal{X}$, having transition kernel P. Suppose there is $C \subset \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty))$, a probability distribution $v(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, and $\epsilon > 0$, B as defined above. If these conditions in the section all hold, Then for any joint initial distribution $\mathcal{L}(X_0, X_0')$, and any integer $1 \leq j \leq k$, if $\{X_n\}$ and $\{X_n'\}$ are two copies of the Markov chain started in the joint distribution $\mathcal{L}(X_0, X_0')$, then

$$||\mathcal{L}(X_k) - \mathcal{L}(X_k')||_{TV} \leq (1-\epsilon)^j + \alpha^{-k} B_{n0}^{j-1} E[h(X_0, X_0')].$$

Proof. First we choose X_0 and X_0' from the initial distribution and doing the coupling loop as above. Therefore we have the following inequality

$$\left\|\mathcal{L}(X_k) - \mathcal{L}(X_k')\right\| \leq P[X_k \neq X_k'] \tag{2}$$

Now let

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

Therefore N_k is the cardinality the set of times where $\{(X_n, X_n')\}$ successfully visit to $C \times C$ and we names this set's elements $\tau_1, \tau_2, ...$ Then for any integer j with $1 \leq j \leq k$,

$$P[X_k \neq X_k'] = P[X_k \neq X_k', N_{k-1} \geq j] + P[X_k \neq X_k', N_{k-1} < j] \tag{3}$$

Since $\{X_k \neq X_k', N_{k-1} > j\}$ including the event that first $1-\epsilon$ event happens. Therefore we have

$$P[X_k \neq X_k', N_{k-1} \geq j] \leq (1-\epsilon)^j. \tag{4}$$

Now we only need to bound $P[X_k \neq X_k', N_{k-1} < j]$.
Let

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

where $N_{-1} = 0$. Now we claim following lemma.

9.3.1 Lemma 13. $\{M_k\}$ is a *supermartingale*

We want to show that

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

Now we consider two scenarios. First if $(X_k, X_k') \notin C \times C$, we know that $X_{k+1} \neq X_{k+1}'$ and $N_k = N_{k-1}$, so

$$\begin{aligned} E[M_{k+1} | X_k, X_k'] &= \alpha^{k+1} B^{-N_{k-1}} E[h(X_{k+1}, X_{k+1}') | X_k, X_k'] \\ &= M_k \frac{E[h(X_{k+1}, X_{k+1}') | X_k, X_k']}{h(X_k, X_k')/\alpha} \end{aligned} \tag{5}$$

So the *drift condition* implies that the fraction part of (5) is less then 1, so

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

Second scenario is when $(X_k, X'_k) \in C \times C$, then $N_k = N_{k-1} + 1$. Since if $X_k = X'_k$, we have $X_{k+1} = X'_{k+1}$ and $E[M_{k+1}] = E[M_k] = 0$,so here we assume $X_k \neq X'_k$. Now we have

$$\begin{aligned} E[M_{k+1}|X_k, X'_k] &= \alpha^{k+1} B^{-N_{k-1}-1} E[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_k \neq X'_k)|X_k, X'_k] \\ &= M_k \frac{\alpha E[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_k \neq X'_k)|X_k, X'_k]}{Bh(X_k, X'_k)} \end{aligned} \tag{6}$$

Since $X_k \neq X'_k$ and we have a ϵ-coin to decide whether $X_{k+1} \neq X'_{k+1}$ or $X_{k+1} = X'_{k+1}$, so

$$E[h(X_{k+1}, X'_{k+1})|X_k, X'_k] = (1-\epsilon)\bar{R}h(X_k, X'_k) + \epsilon * 0$$

Therefore the fraction part of (6) becomes

$$\begin{aligned} \frac{\alpha(1-\epsilon)\bar{R}h(X_k, X'_k)}{Bh(X_k, X'_k)} &= \frac{\alpha(1-\epsilon)\bar{R}h(X_k, X'_k)}{\max[1, \alpha(1-\epsilon)\sup_{C\times C}\bar{R}h]h(X_k, X'_k)} \\ &\leq \frac{\alpha(1-\epsilon)\sup_{C\times C}\bar{R}h(X_k, X'_k)}{\max[1, \alpha(1-\epsilon)\sup_{C\times C}\bar{R}h]h(X_k, X'_k)} \\ &= \min[1, \alpha(1-\epsilon)\sup_{C\times C}\bar{R}h]\frac{1}{h(X_k, X'_k)} \\ &\leq \frac{1}{h(X_k, X'_k)} \leq 1 \end{aligned} \tag{7}$$

Therefore by (7) we have that under the second scenario,

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

Now we combine two scenarios and conclude that $\{M_k\}$ is a supermartingale. And we come back to the second term of (3). Since $B > 1$, we have

$$\begin{aligned} P[X_k \neq X'_k, N_{k-1} < j] &= P[X_k \neq X'_k, N_{k-1} \leq j-1] \\ &\leq P[X_k \neq X'_k, B^{-N_{k-1}} \geq B^{-(j-1)}] \\ &= P[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}} \geq B^{-(j-1)}] \end{aligned} \tag{8}$$

By applying *Markov's Inequality* (P(X≥a) $\leq \frac{E(X)}{a}$) to (8), we get

$$P[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}} \geq B^{-(j-1)}] \leq E[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}}]B^{j-1} \tag{9}$$

Since $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, and the property of supermartingale, we have

$$E[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}}] \leq E[\mathbf{1}(X_k \neq X'_k)B^{-N_{k-1}}h(X_k, X'_k)] = E[M_k]\alpha^{-k} \leq E[M_0]\alpha^{-k} \tag{10}$$

By combining (8), (9) and (10), we have

$$P[X_k \neq X'_k, N_{k-1} < j] \leq B^{j-1}\alpha^{-k}E[M_0] = B^{j-1}\alpha^{-k}E[h(X_0, X'_0)] \tag{11}$$

By combining (2), (3), (4) and (11), we conclude for $n_0 = 1$,

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

Now if $n_0 > 1$, we will adjust a few details of the above process. Here we will not count the visit time which corresponding to filling in times during the construction of $X_{n+1}, ..., X_{n+n_0}$ (and similarly for X'), becasue in these filling times we cannot couple the joint chain. And we let N_k be the time of visit to $C \times C$ and with its $\{\tau_i\}$, avoiding filling in times. Now we replace N_{k-1} in (3) with N_{k-n_0}. And we claim $M_{t(k)}$ is a supermartingale where $t(k)$ is the latest time $\leq k$ such that it does not correspond to filling in times. And we can complete proof by continue the previous process with these changes. □

9.4 Proof of Theorem 9.

Recall Theorem 9: Consider a ϕ-irreducible, aperiodic Markov chain with stationary distribution $\pi(.)$. Suppose the minorisation condition is satisfied for some $C \subset \mathcal{X}$ and $\epsilon > 0$ and probability measure $v(.)$. Suppose further that the drift condition is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \mathcal{X} \to [1, 0]$ for at least one (and hence for π-a.e.) $x \in \mathcal{X}$. Then the chain is geometrically ergodic.

We will prove by using Theorem 12, shows that it satisfy bivariate drift condition. Now set $h(x, y) = \frac{1}{2}[V(x) + V(y)]$. Now we will introduce a few statements which will be useful for the proof.

9.4.1 Lemma 14.

We may assume with out the loss of generality that

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

Specifically given a small set C and drift function V satisfying both minorisation and drift condition, we can find a smaller set $C_0 \subseteq C$ such that it can still satisfy those conditions. Then for C_0 we have λ_0 instead of λ. And the above inequality still hold.

Proof. The drift condition gives that for $1 < \lambda < 1$, and $b < \infty$ and $V : \mathcal{X} \to [1, \infty)$, we have

$$PV \leq \lambda V + b\mathbf{1}_c$$

Now we choose δ with $0 < \delta < 1 - \lambda$, and let $\lambda_0 = 1 - \delta$; $K = b/(1 - \lambda - \delta)$, and set

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

It is clear that the minorisation condition still holds for C_0 since $C_0 \subseteq C$. And drift condition still holds for $x \in C_0$ and $x \in \mathcal{X} \setminus C$ because $\lambda \leq \lambda_0$. Now for $x \in C \setminus C_0$, we have

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

Therefore we have

$$(PV)(x) \leq \lambda_0 V(x) + b\mathbf{1}_{C_0}(x)$$

for all $x \in \mathcal{X}$. □

9.4.2 Proposition 15.

For a geometrically ergodic Markov Chain, with small set C, and drift function V satisfying both minorisation and drift conditions, there does not exist a drift function $V_0 : \mathcal{X} \to [0, \infty)$, with the property upon replacing V with V_0 while both conditions and $\sup_{x \in C} V(x) < \infty$ still holds.

Proof. Consider the Markov Chain on $\mathcal{X} = (0, \infty)$, define as follow. For $0 < x \leq 1$, $P(x, \cdot) = \frac{1}{2}\lambda(\cdot) + \frac{1}{2}\delta_{h(x)}(\cdot)$, where λ is the Lebesgue measure on $(0, 1)$ and $h(x) = 1 + \sqrt{\log(1/x)}$. For $1 < x \leq 2$, $P(x, \cdot)$ is uniform on $[1/2, 1]$. For $x > 2$, $P(x, \cdot) = \delta_{x-1}(\cdot)$. For this chain at interval $C = (0, 1)$, we have $\int_0^1 h(x)dx = 1 + \sqrt{\pi}/2 < \infty$, since C have a finite mean of returning times, so we conclude that C has a stationary distribution (Renewal theory). Let $V(x) = \max(e^x, x^{-1/2})$ be drift function, we can compute $(PV)(x)$.

$$\begin{aligned}(PV)(x) &= \frac{1}{2}V(h(x)) + \frac{1}{2}\int_0^1 V(y)dy \\ &= \frac{1}{2}\max(\exp(1 + \sqrt{\log(1/x)}), (1 + \sqrt{\log(1/x)})^{-1/2}) + \frac{1}{2}\int_0^{0.462} y^{-1/2}dy + \frac{1}{2}\int_{0.462}^1 e^y dy \\ &\approx 1.298\mathbf{1}_c(x) + 1.36\exp(\sqrt{\log(1/x)})\end{aligned}$$

Since we know $\operatorname{argmin}_c V(x) = 0.462$ and we calculate $1.36\exp(\sqrt{\log(1/0.462)}) \approx 2.461$, and
$0.8V(0.462) \approx 1.27$. Therefore we have

$$\begin{aligned}(PV)(x) &\approx 1.298\mathbf{1}_c(x) + 0.8V(x) + 1.191\mathbf{1}_c(x) \\ &\approx 2.49\mathbf{1}_c(x) + 0.8V(x)\end{aligned}$$

So we can also verify that $(PV)(x) \leq 0.8V(x) + 4\mathbf{1}_c(x)$, therefore by theorem 9, we have that this chain is geometrically ergodic.
Now suppose we have some V_0 satisfying drift condition with $\sup_{x\in C} V_0(x) < \infty$. And since $(PV_0)(x) = \frac{1}{2}V_0(h(x)) + \frac{1}{2}\int_0^1 V_0(y)dy$, we have $\sup_{x\in C} V_0(h(x)) < \infty$,i.e. $V_0(h)$ is bounded on $(0, 1]$ which implies V_0 is bounded on entire $\mathcal{X}$. By Fact 10, we conclude that this Chain is uniform ergodic. By contradiction, we have completed this proof. □

With lemma 14 and proposition 15, we know that we can choose a smaller subset of small set and still keep the drift condition true, but we cannot replace new V_0 when keeping the same small set.
Now we can assume $\sup_{x\in C} V(x) < \infty$. And with drift condition, we have

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

which ensure that B_{n_0} is finite.
Now let $d = \inf_{C^c} V(x)$, then by Proposition 11, the bivariate drift condition still hold and gives $d > b/(1 - \lambda) - 1$, which will simply gives theorem 9.
Note. $d \leq b/(1 - \lambda) - 1$ will not let this argument go through. In fact, $d > b/(1 - \lambda) - 1$ ensure that the chain is aperiodic. Without this we must assume the periodicity of the Chain.
Next, our plan is to enlarge C so that the new d will satisfy $d > b/(1 - \lambda) - 1$, and show C is still small. Then, we can show it satisfies bivariate drift condition, by theorem 9 it is geometrically ergodic.
Note. We will have no direct control over the new values of n_0 and ϵ, which is why this approach does not provide a quantitative convergence rate bound. Let us choose any $d' > b/(1 - \lambda) - 1$ and let $S = \{x \in \mathcal{X}, V(x) \leq d'\}$. Now set $C' = C \cup S$, this ensure that $\inf_{C'^c} V(x) \geq d' > b/(1 - \lambda) - 1$. Since V is bounded on S, we have $\sup_{x\in C'} V(x) \leq \infty$. Therefore the bound on bivariate drift condition implies $B_{n_0} < \infty$ with C' instead of C.
Now to complete the proof of Theorem , we still need following lemmas.

9.4.3 Lemma 16.

C' is a small set.

Definition. *A subset $C \subseteq \mathcal{X}$, is a **petite** set 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,*

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

Intuitively, petite set allow the different state in C to cover the minorsation measure $\epsilon\nu(\cdot)$ at different time i. Clearly, any petite set is small, and the converse is generally false.

Note that petite set does not rule out the possibilities of the periodic behavior of a chain. For example some of the state $x \in C$, only cover the minorsation measure at odd times, and others only at even times. For an aperiodic and ϕ-irreduciable chain, we have another result as follow.

9.4.4 Lemma 17.

For an aperiodic and ϕ-irreduciable Markov Chain, all petite sets are small.
The proof of this lemma is in appendix.

9.4.5 Lemma 18.

Let $C' = C \cup S$ where $S = \{x \in \mathcal{X}, V(x) \leq d'\}$ for some $d' < \infty$, then C' is petite.

Proof. Let N large enough that $r \equiv 1 - \lambda^N d' > 0$. Let $\tau_C = \inf n \geq 0; X_n \in C$ be the first return to C. Let $Z_n = \lambda^{-n} V(X_n)$ and $W_n = Z_{\min(n,\tau_C)}$ By drift condition, we know that W_n is a supermartingale. If $n \geq \tau_C$, we have

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

And if $n \leq \tau_C$, then we know that $X_n \notin C$. Then we have

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

Moreover, for $x \in S$, since $V > 1$, by Markov's inequality, we have

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

Therefore $\mathbf{P}[\tau_C < N|X_0 = x] \geq r$.
Since C is (n_0, ϵ, ν)-small, we have $P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot)$ for $x \in C$. Note

$$\sum_{i=1+n_0}^{N+n_0} P^i(x, \cdot) = \sum_{i=1}^{N} P^{n_0}P^i(x, \cdot) = P^{n_0}(\sum_{i=1}^{N} P^i(x, \cdot))$$

Since $P[\tau_C < N|x_0 = x] \geq r$ for any $x \in S$,

$$P^{n_0}(\sum_{i=1}^{N} P^i(x, \cdot)) \geq rP^{n_0}(x, \cdot) \geq r\epsilon v(\cdot)$$

Hence for $x \in C \cup S$, we have the previous inequality which shows that C' is petite. □

Since by previous lemma, we have proven C' is small, therefore we have proven theorem 9.

9.5 Proof of Theorem 4.

Recall Theorem 4: If a Markov chain on a state space with countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution $\pi(.)$, then for $\pi - a.e.\ x \in \mathcal{X}$.

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

Since Theorem 4 does not assume the existence of small set, therefore we need the following Theorems 19. The key idea is to extract the part of $P^{n_0}(x,\cdot)$ which is absolutely continuous with respect to the measure ϕ. And then find C with $\phi(C) > 0$ such that the density of this is at least $\delta > 0$ throughout C.

9.5.1 Theorem 19.

Every ϕ-irreducible Markov Chain on a state space with countably generated σ-algebra, contain a small set $C \subseteq \mathcal{X}$ such that $\phi(C) > 0$. Moreover for each $B \subseteq \mathcal{X}$ with $\phi(B) > 0$, we have a $C \subseteq B$ with $\phi(C) > 0$. And we take minorsation measure $\nu(\cdot)$ such that $\nu(C) > 0$.

Now if we can show (X_n, X'_n) will hit $C \times C$ infinitely often, then we will have infinite opportunities of coupling with probability $\geq \epsilon > 0$. Then it will eventuality coupling w.p. 1 due to construction. Hence it will prove Theorem 4.

Now following above idea, we have following lemma.

9.5.2 Lemma 20.

Consider a Markov Chain on state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$. Suppose for some $A \subseteq X$, we have $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$, then for π-a.e. $x \in \mathcal{X}$, $\mathbf{P}_x(\tau_A < \infty) = 1$.

Proof. Recall that $\tau_A = \inf\{n \geq 1, X_n \in A\}$. i.e. the first time when X_n enters A. Now we suppose the contrary which is

$$\pi\{x \in \mathcal{X} : \mathbf{P}_x(\tau_A = \infty) > 0\} > 0 \tag{12}$$

i.e. there exists $x \in \mathcal{X}$ such that it will never return to A.

Step 1 First we want to find δ_1 and $B_1 \subseteq \mathcal{X}$ with $\pi(B_1) > 0$, such that $\mathbf{P}_x(\tau_A < \infty) \leq 1 - \delta_1$ for all $x \in B_1$. Let $A_n \equiv \{x \in \mathcal{X} : P_x(\tau_A < \infty) \leq 1 - \frac{1}{n}\}$ and $A \equiv \{x \in \mathcal{X} : P_x(\tau_A < \infty) < 1\}$. Clearly A_n converges A as n goes to infinity. Then since $\pi(A) = \pi\{x \in \mathcal{X} : \mathbf{P}_x(\tau_A = \infty) > 0\} > 0$, there exists $n \in \mathbb{Z}$ such that $\pi(A_n) > 0$. We then set $B_1 = A_n$ and $\delta_1 = 1/n$.

Since $\mathbf{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathcal{X}$, then we can find $\ell_0 \in \mathbb{N}$ and $\delta_2 > 0$ and $B_2 \subseteq B_1$ with $\phi(B_2) > 0$ such that $P^{\ell_0}(x, A) \geq \delta_2$ for all $x \in B_2$. To be more specific, let $U_i = \{x \in B_2, P^i(x, A) > 0\}$. Since $P_x(\tau_A < \infty) > 0$ for all $x \in B_2$, we have

$$\bigcup_{i=1}^{\infty} U_i = B_2$$

It follows that there must exist one such U_i with $\pi(U_i) > 0$ (otherwise $\pi(B_2) = 0$). So we can choose $U_{l_0} \subset B_2$ such that $\pi(U_{l_0}) > 0$ and $P^{l_0}(x, A) > 0$ for all $x \in U_{l_0}$. Thus we can find our desired B_2 and δ_2 (in the similar way we define B_1).

Now let us set $\eta = \#\{k \geq 1 : X_{k\ell_0} \in B_2\}$. Since for $x \in B_2$, $P^{\ell_0}(x, A) \geq \delta_2$, we have $\mathbf{P}_x(\tau_A = \infty) \leq (1 - \delta_2)$. Then for any $r \in \mathbb{N}$ and $x \in \mathcal{X}$, we have $\mathbf{P}_x(\tau_A = \infty, \eta = r) \leq (1 - \delta_2)^r$. In particular $\mathbf{P}_x(\tau_A = \infty, \eta = r) = 0$. Therefore, for $x \in B_2$

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

Note $\eta < \infty$ if and only if $\sup\{k \geq 1, X_{k\ell_0} \in B_2\} < \infty$. So

$$\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1, X_{k\ell_0} \in B_2\} < \infty) = \mathbf{P}_x(\tau_A = \infty, \eta < \infty) \geq \delta_1$$

Let $D_n \equiv (\sup\{k \geq 1, X_{k\ell_0} \in B_2\} < n)$ be an event for any positive integer n. It follows that $\mathbf{P}_x(\tau_A = \infty, D_n)$ converges to $\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1, X_{k\ell_0} \in B_2\} < \infty)$ as n goes to infinity. So there exists $n \in \mathbb{Z}$ and $\delta > 0$ such that $\mathbf{P}_x(\tau_A = \infty, D_n) > \delta$. Let $\ell = n$, then

$$\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1, X_{k\ell_0} \in B_2\} \leq \ell) \geq \delta \quad x \in B$$

Since $B \in B_2$, we have $\sup\{k \geq 1, X_{k\ell_0} \in B\} \leq \sup\{k \geq 1, X_{k\ell_0} \in B_2\}$, then for some $\ell, \ell_0 \in \mathbb{N}$, $\delta > 0$, we have

$$\mathbf{P}_x(\tau_A = \infty, \sup\{k \geq 1, X_{k\ell_0} \in B\} \leq \ell) \geq \delta \quad x \in B \subseteq \mathcal{X}$$

Step 2 Following step 1, let $L = \ell\ell_0$, and $S = \sup\{k \geq 1; X_{kL} \in B\}$. Note that we let $S = -\infty$ if $\{k \geq 1; X_{kL} \in B\} = \varnothing$
If $S = r$, then after r steps, X_{kL} will never return to A. Also since for any $j \in \mathbb{N}$,by stationarity we have

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

Now for all integers $1 \leq r \leq j$, Then we have

$$\begin{aligned}
\int_{x\in\mathcal{X}} \pi(dx)\mathbf{P}_x[S = r, X_{jL} \notin A] &= \int_{x\in\mathcal{X}} \pi(dx) \int_{y\in B} P^{rL}(x, dy)\mathbf{P}_x[S = -\infty, X_{(j-r)L} \notin A] \\
&= \int_{y\in B}\int_{x\in\mathcal{X}} \pi(dx)P^{rL}(x, dy)\mathbf{P}_y[S = -\infty, X_{(j-r)L} \notin A] \\
&= \int_{y\in B} \pi(dy)\mathbf{P}_y[S = -\infty, X_{(j-r)L} \notin A] \\
&\geq \int_{y\in B} \pi(dy)\mathbf{P}_x[\sup\{k \geq 1; X_{kL} \in B\} \leq l, \tau_A = \infty] \\
&\text{(as } [\sup\{k \geq 1; X_{kL} \in B\} < l] \subset [S = -\infty] \text{ and } [\tau_A = \infty] \subset [X_{(j-r)L} \notin A]) \\
&\geq \int_{y\in B} \pi(dy)\delta = \pi(B)\delta
\end{aligned}$$

Now we have

$$\pi(A^C) \geq j\pi(B)\delta$$

Then for $j > \pi(B)\delta$, we have $\pi(A^c) > 1$ which is impossible. Therefore by contradiction, we have proven Lemma 20. □

Now continue on the proof of Theorem 4, we let C be a small set from Theorem 19. Consider the coupling construction $\{(X_n, Y_n)\}$). Let $G \subseteq \mathcal{X} \times \mathcal{X}$ be the set of (x, y) such that $\mathbf{P}_{(x,y)}[\exists n \geq 1 : X_n = Y_n] = 1$. So if $(X_0, Y_0) \equiv (x, Y_0) \in G$, then we have

$$\lim_{n\to\infty} \|P^{n_0}(x, \cdot) - \pi(\cdot)\| = 0$$

proving Theorem 4. And it is sufficient to show that for π-a.e. $x \in \mathcal{X}$, we have $\mathbf{P}[(x, Y_0) \in G] = 1$.
Now let $G_x = \{y \in \mathcal{X}; (x, y) \in G\}$ for $x \in \mathcal{X}$. And let $\bar{G} = \{x \in \mathcal{X}; \pi(G_x) = 1\}$. We have following lemma

9.5.3 Lemma 21. $\pi(\bar{G}) = 1$.

Proof. Since $\nu(C) > 0$ by the assumption in Theorem 19, along with Lemma 34 in appendix, we are able to show that for any $(x, y) \in \mathcal{X} \times \mathcal{X}$ have positive probability of eventually hitting $C \times C$. First, there is some n and a with $P^n(x, C) = a > 0$ by irreducibility. Then, since C is a (n_0, ϵ, ν)-small set,

$$P^n n + n_0(x, \cdot) \geq a\epsilon\nu(\cdot).$$

Then, by Lemma 34, for any $r \geq n^*$,

$$\begin{aligned}
P^{n+n_0+r}(x, \cdot) &\geq a\epsilon \int_x \nu(dz)P^r(z, \cdot) \\
&\geq a\epsilon\delta_r\nu(\cdot)
\end{aligned}$$

Then, choose $r, s \geq n^*$ to make $n + n_0 + r = m + n_0 + s = N$. Then, since X_n and Y_n are independent until they reach $C \times C$, we must have

$$\begin{aligned} P[X_N \in C, Y_N \in C] &\geq P[X_N \in C]P[Y_N \in C] \\ &\geq [a\epsilon\delta_r\nu(C)][b\epsilon\delta_s\nu(C)] \\ &> 0 \end{aligned}$$

This last inequality follows from the fact that $\nu(C) > 0$.
Now by Lemma 20, we know that the joint chain will return to $C \times C$ w.p. 1 from $(\pi \times \pi)$-a.e. $(x, y) \notin C \times C$. Once the chain reaches $C \times C$, by construction since $\bar{R}$ is absolute continuous with respect to $\pi \times \pi$. Therefore by Lemma 20, this chain will keep return repeatedly to $C \times C$ w.p. 1 until $X_n = Y_n$. Since we know from the construction that each time in $C \times C$, the chain have probability $\geq \epsilon$ of $X_n = Y_n$. Therefore, we will eventually have $X_n = Y_n$. So we have $(\pi \times \pi)(G) = 1$
Now we assume $\pi(\bar{G}) < 1$, the by stationarity, we have

$$\begin{aligned} \pi(G^C) &= \int_{\mathcal{X}} \pi(dx)\pi(G_x^C) \\ &= \int_{\bar{G}} \pi(dx)(1 - \pi(G_x)) + \int_{\bar{G}^C} \pi(dx)(1 - \pi(G_x)) \\ &= \int_{\bar{G}^C} \pi(dx)(1 - \pi(G_x)) > 0 \end{aligned}$$

While $(\pi \times \pi)(G^C) = 0$. Therefore by contradiction, we have proven Lemma 21. □

Therefore we have proven Theorem 4.

10 Appendix: background

10.1 Absolute continuity of measures

Definition If μ and v are two measures on the same measurable space $(X, \mathcal{A})$, μ is said to be absolutely continuous with respect to v if $\mu(A) = 0$ for every measurable set for which $v(A) = 0$. This is written as $u \ll \nu$

$$u \ll v \iff \forall A \in \mathcal{A}\ (v(A) = 0 \implies \mu(A) = 0)$$

10.2 Radon–Nikodym theorem

If $v \ll \mu$, then there is a measurable function $f : X \to [0, \infty)$, such that for any measurable set $A \subset X$,

$$v(A) = \int_A f d\mu$$

The function f is called the **Radon–Nikodym derivative** and is denoted by $\frac{dv}{d\mu}$.

10.3 Formal definition of the density

A random variable x with values in a measurable space $\mathcal{X}, \mathcal{A}$(usually $\mathbb{R}^n$ with the Borel sets as measurable subsets) has a probability distribution X_*P on $(\mathcal{X}, \mathcal{A})$: the density of X with respect to a reference measure μ on $(\mathcal{X}, \mathcal{A})$ is the **Radon–Nikodym derivative**

$$f = \frac{dX_*P}{d\mu}$$

10.4 Lebesgue Measure

Definition. For each subset E of R we define its Lebesgue outer measure $\mu^*(E)$ by

$$\mu^*(E) = \inf\{\sum_{k=1}^{\infty} l(I_k) : \{I_k \text{ is a sequence of open intervals with } E \subset \cup_{k=1}^{\infty} I_k\}$$

Note that $0 \leq \mu^*(E) \leq \infty$ for any set $E \subset \mathbb{R}$.
Definition. A set $E \subset \mathbb{R}$ is called Lebesgue measurable if for every subset A of $\mathbb{R}$,

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^C)$$

Definition. If E is a Lebesgue measurable set, then the Lebesgue measure of E is defined to be its outer measure $\mu^*(A)$ and is written $\mu(A)$.

10.5 Proof of Lemma 17.

To prove Lemma 17, we need to introduce following lemma.

Lemma 22. 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 $\nu(\cdot) \ll \pi(\cdot)$, and that for all $x \in \mathcal{X}$, there is $n = n(x) \in \mathbb{N}$ and $\delta = \delta(x) > 0$ such that $P^n(x, \cdot) > \delta\nu(x)$(for example, this always holds if $\nu(\cdot)$ is a minorisation measure for a small or petite set which is reachable from all states). Let $T = \{n \geq 1; \exists \delta_n > 0 s.t. \int \nu(dx) P^n(x, \cdot) \geq \delta\nu(\cdot)\}$ and assume that T is non-empty. Then there is $n_* \in \mathbb{N}$ with $T \supseteq \{n_*, n_* + 1, n_* + 2, ...\}$.

Proof. Let us first interpret T. We can see that T is the set of steps which some $x \in \mathcal{X}$ which satisfy n-step minorsation condition with positive measure $\nu(\cdot)$. Since we know that for all $x \in \mathcal{X}$.

$$P^{(n(x))}(x, \cdot) \geq \delta(x)\nu(\cdot)$$

And we assume that T is not empty.
Therefore if $n, m \in T$, then we have

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

We can see that T is close under addition.
Now we will show that $gcd(T) = 1$. Suppose the contrary that $gcd(T) = d > 1$. For $1 \leq i \leq d$, let

$$\mathcal{X}_i = \{x \in \mathcal{X}; \exists \ell \in \mathbb{N}, \delta > 0 \text{ s.t. } P^{\ell d-i}(x,\cdot) \geq \delta\nu(\cdot)\}$$

Since for all $x \in \mathcal{X}$, x satisfies minorsation condition. Therefore we have $\bigcup_{i=1}^d \mathcal{X}_i = \mathcal{X}$. Now let

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

And let $\bar{S} = S \cup \{x \in \mathcal{X}; \exists m \in \mathbb{N} \text{ s.t. } P^m(x, S) > 0\}$. So $\bar{S}$ is S union the set of all elements which can reach S. Now let

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

Therefore $\mathcal{X}_1', \ldots, \mathcal{X}_d'$ are disjoint. And we know that for $x \in \mathcal{X}_i'$, we have $P(x, \bar{S}) = 0$. Therefore by construction, we have $P(x, \bigcup_i \mathcal{X}_i') = 1$. In fact we must have $P(x, \mathcal{X}_{i+1}') = 1$ for $i < d$, and $P(x, \mathcal{X}_1') = 1$ for $i = d$. Otherwise we would have x in two $\mathcal{X}_j'$ which would contradict the disjointedness.

Now let us consider for all $m > 0$, $\nu(P^m(\mathcal{X}_i \cap \mathcal{X}_j))$ for $i \neq j$. Now if $\nu(P^m(\mathcal{X}_i \cap \mathcal{X}_j)) > 0$, then we have $S' \subseteq \mathcal{X}$, such that for $x \in S'$, we have $\ell_1, \ell_2 \in \mathbb{N}$, and $\delta > 0$ such that for all $x \in S'$,

$$P^{\ell_1 d+i}(x,\cdot) \geq \delta\nu(\cdot);\ P^{\ell_1 d+j}(x,\cdot) \geq \delta\nu(\cdot).$$

Therefore by additive property of T we have $\ell_1 d + i + m \in T$ and $\ell_2 d + i + m \in T$, which contradict the assumption that $gcd(T) = d$. Therefore we have for all $m > 0$, $\nu(P^m(\mathcal{X}_i \cap \mathcal{X}_j)) = 0$ for $i \neq j$.
Therefore by construction of $\bar{S}$, we have $\nu(\bar{S}) = 0$. Therefore we have $\nu(\bigcup_{i=1}^d \mathcal{X}_i') = 1$, so $\nu(\bigcup_{i=1}^d \mathcal{X}_i) = \nu(\mathcal{X}) = 1$. Since we have $\nu(\cdot) \ll \pi(\cdot)$, therefore $\pi(\bigcup_{i=1}^d \mathcal{X}_i') > 0$.

Since $\mathcal{X}_i'$ are disjoint and are subsets of a positive measure $\pi(\bigcup_{i=1}^d \mathcal{X}_i')$. We conclude that this chain has periodic behavior. Since by assumption we know our chain is aperiodic. Therefore by contradiction, we have that $gcd(T) = 1$. Since T is additive and non-empty, we must have $n_* \in \mathbb{N}$ such that $T \supseteq \{n_*, n_* + 1, n_* + 2, ..\}$. □

10.5.1 Proof of lemma 17.

Proof. Let R be $(n_0, \epsilon, \nu(\cdot))$-petite under the assumption of a chain in lemma 17. Therefore by definition, for all $x \in R$ we have

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

Now Let T be the same as Lemma 34. Then we have

$$\sum_{i=1}^{n_0} \int_{x\in\mathcal{X}} \nu(dx)P^i(x, \cdot) \geq \epsilon\nu(\cdot)$$

. Therefore there must exist some $1 \leq i \leq n_0$ and δ such that $\int \nu(dx)P^i(x, \cdot) \geq \delta\nu(\cdot)$. Therefore $T \neq \varnothing$. Hence by assumption in lemma 17 and lemma 34, we can find n_* and $\delta_n > 0$ such that $\int \nu(dx)P^n(x, \cdot) \geq \delta_n\nu(\cdot)$ for all $n \geq n_*$. Now let $N = n_* + n_0$, and let $r = \min\{\delta : n_* \leq n \leq N - 1\}$. Therefore for $x \in R$ we have we have

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

Therefore we conclude that R is $(N, \frac{r\epsilon}{n_0}, \nu(\cdot))$-small. □