Markov chain Monte Carlo Methods and Applications

Swetha U. Lal
MS14162

A dissertation submitted for the partial fulfillment of BS-MS dual degree in Science

Indian Institute of Science Education and Research, Mohali
April 2019

Certificate of Examination

This is to certify that the dissertation titled "Markov chain Monte Carlo Algorithms and Applications" submitted by Ms. Swetha U. Lal (Reg. No. MS14162) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Amit Kulshrestha Dr. Lingaraj Sahu Dr. Neeraja Sahasrabudhe

(Supervisor)

Dated: April 25, 2019

Declaration

The work in this dissertation has been carried out by me under the guidance of Dr. Neeraja Sahasrabudhe at the Indian Institute of Science Education and Research, Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Swetha U. Lal

(Candidate)

Dated: April 25, 2019

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Neeraja Sahasrabudhe
(Supervisor)

Acknowledgement

I thank both of my committee members Dr. Amit Kulshrestha and Dr. Lingaraj Sahu, for coming and sitting through the presentations and their valuable inputs and suggestions. I would like to thank my guide Dr. Neeraja Saharabudhe for her endless support and encouragement; for patiently going through the one year journey with me, constantly helping and pushing us to work. I am ever so grateful to Amma, Achan, Chechi, Chettan for their love, support and for believing in me. Honorary mentions to Vishal and Pinku without whom my presentations would not have happened. I thank Vaitheesh, Priyasha, Adarsh, Varun, Awani for all those tea breaks and laughter. I thank Balu, Adheena, Abin, Aswathy, Jain, Arya for all the love and food. I express my love to Darsana, Greeshma, Shiny for being the best friends there could be, for their love, for hearing me out and all of our never-ending gabfests. I send my thanks and love to Somak for all the help, conversations, and all those jokes shared when we were almost at a breaking point. Finally, my love and gratitude to Sachin, for being the best; for being the much needed help and inspiration.

Swetha U. Lal

Contents

Abstract

Markov Chain Monte Carlo (MCMC) Methods are used extensively in various problems across physics, engineering and applied mathematics. In this thesis, we study the convergence results as well as the two standard but very important Markov Chain Monte Carlo algorithms, namely, the Gibbs Sampler and the Metropolis algorithm. The theory of Markov chain convergence is vast and a lot of work has been done recently on mixing times of Markov chains. A large part of thesis focuses on the conditions required for uniform as well as geometric ergodicity of Markov chains and thus providing quantitative bounds to the convergence of the Markov chain to stationarity. A brief idea of how MCMC algorithms work is also presented. Finally, we consider an application of MCMC to covariance realization problem for a discrete random process.

Introduction

Markov Chain Monte Carlo methods find applications across applied sciences. It finds use in optimization, when there are difficult numerical integrals with complicated boundary conditions, and mostly in places where we need to sample from a probability distribution. Markov Chain Monte Carlo methods are used to address problems of the following kind: Consider a density function π_u such that $0 < \int_\Omega \pi_u < \infty$ where $\Omega \subset \mathbb{R}$. We get a probability measure on Ω from this density:

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

Suppose we have a function $f : \Omega \longrightarrow \mathbb{R}$ and we want to compute its expectation with respect to $\pi(.)$. That is, we want to estimate :

$$\pi(f) = \mathbb{E}_\pi[f(X)] = \frac{\int_\Omega f(x)\pi_u(x)dx}{\int_\Omega \pi_u(x)dx}$$

There might be cases where Ω is of very high dimensional and $\pi(.)$ fairly complicated. These complications make the estimation of the integral computationally difficult and inefficient. This is where Markov chain Monte Carlo comes into the picture. The Monte Carlo solution to this problem is to simulate independently identically generated random variables $Y_1, Y_2, ...Y_N \sim \pi(.)$ and then, estimate $\pi(f)$ as follows :

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

The idea behind the Markov Chain Monte Carlo algorithm is to generate a Markov chain that can be easily simulated and is such that it converges, reasonably fast, to a stationary distribution that matches the distribution from which we are trying to obtain our samples. That is, we want the Markov chain to have transition probabilities $P(x, dy)$ for states $x, y \in \Omega$

satisfying

$$\int_{x\in\Omega} \pi(dx)P(x,dy) = \pi(dy)$$

We run the Markov chain long enough so that the distribution of X_n gets very close to the stationary distribution and then, take $Y_1 = X_n$. Similarly, to obtain $Y_2, Y_3, \ldots$ and so on, we do multiple runs of the Markov chain. We can then estimate the expectation of the given function as described above.

It usually takes more than a few runs before the Markov chain gets reasonably close to the stationary distribution. Hence, the usual practice is to ignore the first few values obtained via the Markov chain. This is commonly known as the "burning period" of a Markov chain. Note that for an irreducible, aperiodic Markov chain on a finite state space, there is a unique stationary distribution and the convergence does not depend on the state we are starting in. Often, instead of creating or starting a Markov chain afresh, the tail of an already existing Markov chain is used for estimation purposes. That is, we now have the estimates as $\frac{1}{N-B}\sum_{i=B+1}^{N} f(X_i)$, where B, large enough, is the burning period.

In applying the Markov chain Monte Carlo algorithms to a covariance realization problem, we also briefly illustrate the effectiveness (in terms of speed of convergence) of the adaptive Markov chain Monte Carlo methods. While we do not provide any explicit theoretical bounds on the convergence rates, it is observed that when the state space is very large, using adaptive methods are very useful.

The thesis is organized as follows : The first chapter concentrates on the preliminaries of Markov chain theory that are required in the subsequent chapters. Basic definitions and concepts are discussed in this chapter. A few examples are also discussed for better understanding of the concepts. The next chapter gives a brief overview of two most popular Markov chain Monte Carlo algorithms : the Metropolis-Hastings algorithm and the Gibbs sampler algorithm. It provides the basic idea and an example to understand the Metropolis algorithm, which is used in the application part of the thesis, better. Chapter 3 is where we discuss the convergence results for Markov chains. It states the Asymptotic convergence theorem, an elaborate proof for it and also reproduces the main result of the paper [GORJSR01]. The main results and proofs are from [GORJSR01] and have been included for the sake of completeness. Some proofs of results in *Chapter 3* are omitted in the interest of space. These proofs can be found in [GORJSR01]. Finally, in the last chapter, we talk about an application of the MCMC algorithm. We illustrate how MCMC algorithm can be used to efficiently obtain a probability distribution of a finite discrete random vector given its variance-covariance matrix. We also demonstrate the effectiveness of the adaptive Markov chain Monte Carlo algorithms.

Chapter 1

Preliminaries

In this chapter, we discuss some basic definitions associated with Markov chains and properties of transition matrices. We start by defining a Markov chain on a state space Ω.

Definition 1. ***Markov chain***

A discrete time Markov chain $(X_i)_{i\geq 0}$ is a sequence of random variables X_i where each X_i takes its value from the state space Ω associated to the Markov chain, where, given that we are at time n, the next state X_{n+1} depends only on the current state X_n. The memory of how we got to X_n doesn't matter.

That is,

$$P(X_{n+1}|X_0, X_1, ..., X_n) = P(X_{n+1}|X_n) \quad \textit{where } X_0, X_1, ..., X_n, X_{n+1} \in \Omega$$

Figure 1.1 is an example of a Markov chain on a state space of cardinality 5. Note that at a given time, the next move depends only on the present state.

Definition 2. ***State space***

A state space associated to a Markov chain is the set of states that the Markov chain is allowed to take values from.

In the Markov chain shown below, the associated state space is $\Omega = \{1, 2, 3, 4, 5\}$.

Definition 3. ***Transition probability***

Transition probability is the probability that our Markov chain goes from a state i to a state j where $i, j \in \Omega$, is given by

$$p_{ij} = \mathbf{P}(X_{t+1} = j|X_t = i); i, j \in \Omega$$

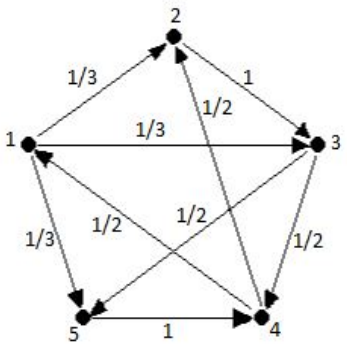

Figure 1.1: A Markov chain.

Consider the example of the Markov chain $(X_i)_{i\geq 0}$ on the state space $\Omega = \{1, 2, 3, 4, 5\}$ discussed above. Here, $X_i \in \Omega$ for each i with the transition probabilities is shown in the figure (*Figure 1.1*).
Each arrow points from one state (say X_i) to the next state (say X_{i+1}) and has an associated probability with it, which is the transition probability of going from X_i to X_{i+1}. that is, $p_{X_i,X_{i+1}} = P(X_{i+1}|X_i)$.

Definition 4. ***k-th step Transition probability***
k-th step transition probability is the probability that the Markov chain will reach state j from state i in exactly k steps. The expression for this is given by

$$p_{ij}^{(k)} = \mathbf{P}(X_{t+k} = j|X_t = i); i, j \in \Omega$$

Definition 5. ***Transition probability matrix***

The transition probability matrix is essentially the collection of all possible transition probabilities of the Markov chain.

$$P = (p_{ij})$$

Similarly, the transition probability matrix for the k-step transition is given by $P^k = (p_{ij}^{(k)})$.

The following is the transition probability matrix associated to the Markov chain mentioned in the example shown in *Figure 1.1* above.

$$P = \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

The 3-step transition probability of this Markov chain will be given by P^3.

Definition 6. ***Stationary distribution***

Stationary distribution of a Markov chain is a probability distribution (π) *that remains constant as the Markov chain runs. Thus, it satisfies the equation*

$$\pi = \pi P$$

Consider the following example.

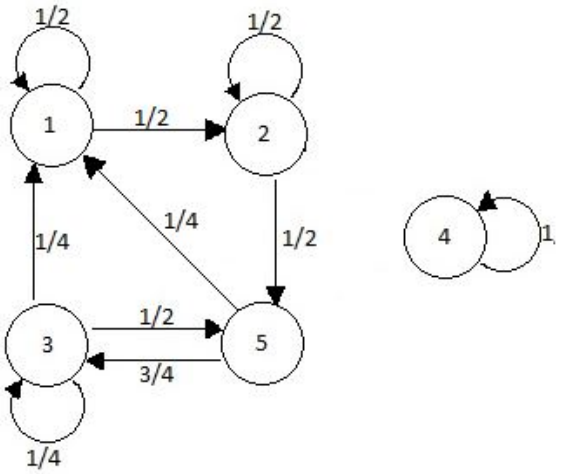

Figure 1.2: A Markov chain and its corresponding transition probability matrix is shown above.

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1/4 & 0 & 1/4 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \end{bmatrix}$$

Note that the probability distribution $\pi = [1/5\ 1/5\ 1/5\ 1/5\ 1/5]$ is stationary with respect to this Markov chain as it satisfies the equation $\pi = \pi P$.
However, we can see that the Markov chain does not converge to π. If the Markov chain has initial distribution $X_0 \in \{1, 2, 3, 5\}$, then, as time progresses, X_n will be restricted to $X_n \in \{1, 2, 3, 5\}$. That is, X_n will never visit the state 4 and hence the probability for visiting that state will be 0.
Hence, we see that the Markov chain will not converge to the given stationary distribution π. This occurs due to reducible property of the Markov chain. Irreduciblity of a Markov chain is defined as follows:

Definition 7. ***Irreducible Markov chain***
A Markov chain where there is a positive probability of reaching any state from any other state is called an irreducible Markov chain. That is, for every $i, j \in \Omega$ there exists a k such that

$$p_{ij}^{(k)} > 0$$

However, irreducibility is not enough to ensure the convergence to a stationary distribution. Given below is an example of an irreducible Markov chain which does not converge to any stationary distribution. This happens due to the periodicity of the chain.

In *Figure1.3*, we see that the Markov chain is irreducible since one can go from any state to any other state in finitely many steps. However, starting from a particular state, we can come back to the exact same state only in steps of multiples of 5. That is, starting from $X_n = i$, where $i \in \Omega$, we have that for $X_m = i$, m will be of the form $n + 5j$; $j \in \mathbb{N}$. In such a case, we can see that the Markov chain would not converge to any stationary distribution.

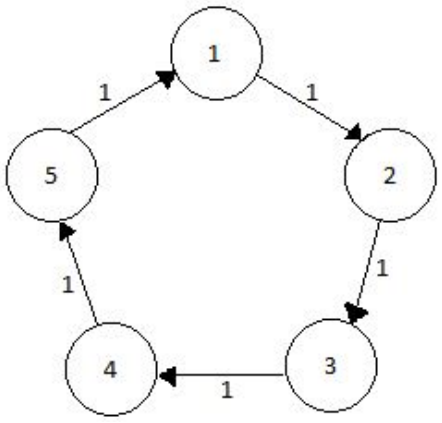

Figure 1.3: Figure depicting a Markov chain with period 5. Its corresponding transition probability matrix is shown below.

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

Definition 8. ***Periodicity of a Markov chain***

An irreducible Markov chain is called periodic if there exists a state in the state space such that the greatest common divisor of all possible number of steps required to start from that state and come back to the same state is greater than 1.

A Markov chain is said to have period d where

$$d = gcd\{k > 0 : p_{ij}^{(k)} > 0\}$$

The Markov chain is said to be aperiodic if $d = 1$.

The above example in *Figure1.3* depicts a Markov chain with period 5 and *Figure1.4* shows an aperiodic Markov chain.

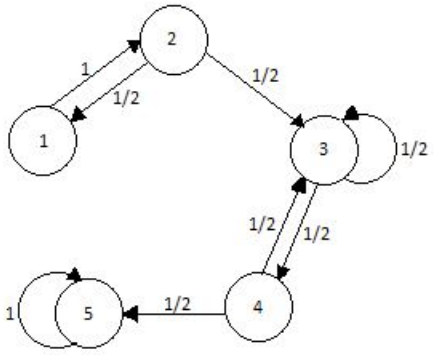

Figure 1.4: An aperiodic Markov chain and its corresponding transition probability matrix is given below.

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

Definition 9. ***Aperiodicity of a Markov chain*** *[GORJSR01]*

A Markov chain with stationary distribution $\pi(.)$ *is aperiodic if there do not exist* $d \geq 2$ *and disjoint subsets* $\Omega_1, \Omega_2, ..., \Omega_d \subseteq \Omega$ *with* $P(x, \Omega_{i+1}) = 1$ *for all* $x \in \Omega_i$ *where* $1 \leq i \leq d - 1$,

and $P(x, \Omega_1) = 1$ *for all* $x \in \Omega_d$ *such that* $\pi(\Omega_1) > 0$ *for all* i.
Otherwise, the chain is periodic with period d*, with periodic decomposition* $\Omega_1, \Omega_2, ..., \Omega_d$.

Definition 10. ***Reversibility of a Markov chain***
A Markov chain is called reversible if the following holds

$$\pi(i)p_{ij} = \pi(j)p_{ji}$$

Consider the following example of a Markov chain.

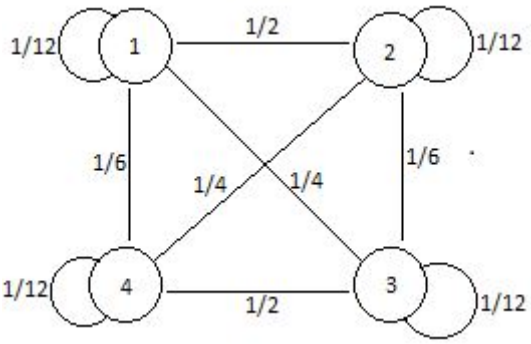

Figure 1.5: A reversible Markov chain

The transition probability matrix of the following Markov chain looks like :

$$P = \begin{bmatrix} 1/12 & 1/2 & 1/4 & 1/6 \\ 1/2 & 1/12 & 1/6 & 1/4 \\ 1/4 & 1/6 & 1/12 & 1/2 \\ 1/6 & 1/4 & 1/2 & 1/12 \end{bmatrix}$$

Now, consider the uniform stationary distribution on this chain. That is, $\pi = [1/4, 1/4, 1/4, 1/4]$. Clearly, we can see that the Markov chain is reversible with respect to this stationary distribution as it satisfies

$$\pi(i)p_{ij} = \pi(j)p_{ji} \qquad \forall\ 1 \leq i, j \leq 4.$$

Remark 1.0.1. *Throughout this thesis,* $\|.\|$ *refers to the standard Euclidean norm.*

Chapter 2

Markov chain Monte Carlo Methods

2.1 Gibbs Sampler

Consider the problem of generating samples from a very large set according to a given probability distribution. Several sampling techniques are known to be useful. We refer the interested reader to [NM02]. In this chapter, we focus on Markov chain Monte Carlo (MCMC) method. The central idea is to construct a reversible Markov chain that converges to the required distribution. Constructing a suitable Markov chain depends on the problem at hand. We discuss two important algorithms to sample using MCMC.

We first discuss the Gibbs Sampler method [IY12], [GORJSR01]. Suppose that we have a d-dimensional density $\pi_u(.)$ on Ω, an open subset of $\mathbb{R}^d$ and we write $\mathbf{x} = (x_1, x_2, ..., x_d)$. The Gibbs Sampler works as follows:

- Start with $\mathbf{x}^{(0)} = \mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^d$.
- For the iterations $i = 1, 2, ...$ from now on; a single iteration is as shown below :
- Begin iteration:
  - $x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, ...X_d = x_d^{(i-1)})$
  - $x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, ...X_d = x_d^{(i-1)})$

    .

    .

    .
  - $x_d^{(i)} \sim p(X_d = x_d | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, ...X_{d-1} = x_{d-1}^{(i)})$
- End iteration

$P_j = p(X_d = x_d | X_1 = x_1^{(i)}, .., X_{j-1} = x_{j-1}^{(i)}, .., X_{j+1} = x_{j+1}^{(i-1)}), .., X_{d-1} = x_{d-1}^{(i-1)})$, then,

The deterministic scan Gibbs sampler is:

$$P = P_1 P_2 ... P_d$$

The random scan Gibbs sampler is:

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

The output of Gibbs sampler will result in a zig-zag pattern because of the construction of its updation.

2.2 The Metropolis-Hastings method of MCMC algorithm

As we saw, the idea behind MCMC algorithm is the construction of a Markov chain with a given stationary distribution π. The key notion for this construction is, *reversibility*. That is, $\pi(i)p_{i,j} = \pi(j)p_{j,i}$.

Proposition 2.2.1. *[GORJSR01] If a Markov chain is reversible with respect to $\pi(.)$, then $\pi(.)$ is stationary for the Markov chain.*

Proof:A stationary distribution π satisfies the equation $\pi = \pi\mathbf{P}$. Thus, we just need to show that the equation of stationary distribution is satisfied.

$$\begin{aligned} \sum_{i \in \Omega} \pi(i) p_{ij} &= \sum_{i \in \Omega} \pi(j) p_{ji} \\ &= \pi(j) \sum_{i \in \Omega} p_{ji} \\ &= \pi(j) \qquad \text{(Since } \sum_{i \in \Omega} p_{ji} = 1\text{).} \end{aligned}$$

Hence, from here, we can see that it suffices to create a Markov chain that can be easily run on a computer and that is reversible.

The Metropolis-Hastings algorithm

Consider a stationary distribution $\pi(.)$ having a density π_u which is possibly unnormalized. Now, consider another Markov chain $Q(.,.)$ called the proposal Markov chain whose

transitions also have some density.

The Metropolis-Hastings algorithm:
First, choose some initial distribution X_0. Then, given X_n, we generate a proposal Y_{n+1} from the proposal chain $Q(Xn, .)$ having a probability distribution $q(., .)$. We define the acceptance probability as : $\alpha(x, y) = min\left[1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right]$ (set $\alpha(x, y) = 1$ when $\pi(x)q(x, y) = 0$). Now, either accept the proposal with probability $\alpha(X_n, Y_{n+1})$ and set $X_{n+1} = Y_{n+1}$ or reject the proposal with probability $1 - \alpha(X_n, Y_{n+1})$ by setting $X_{n+1} = X_n$.
According to the Metropolis-Hastings algorithm, the **The Markov chain thus obtained is reversible**.

Proof. We will show that the Markov chain obtained by the Metropolis-Hastings method is reversible. For showing this, we need to show that $\pi(x)P(x, y) = \pi(y)P(y, x)$.

$$\begin{aligned}
\pi(x)P(x, y) &= \pi(x)q(x, y)\alpha(x, y) \\
&= \pi(x)q(x, y)min\left[1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right] \\
&= min\left[\pi(x)q(x, y), \pi(y)q(y, x)\right] \\
&= \pi(y)q(y, x)min\left[\frac{\pi(x)q(x, y)}{\pi(y)q(y, x)}, 1\right] \\
&= \pi(y)q(y, x)\alpha(y, x) \\
&= \pi(y)P(y, x).
\end{aligned}$$

■

The following example depicts how the Metropolis algorithm works. Consider the following proposal chain on $\Omega = \{1, 2, 3, 4, 5\}$:

$$\begin{aligned}
&q_{i,i+1} = \frac{1}{2} = q_{i.i-1} \qquad \forall\ 2 \leq i \leq 4. \\
&q_{1,2} = 1 \qquad q_{5,4} = 1
\end{aligned}$$

Let $\pi(.)$ be the Poisson distribution with parameter $\lambda = 7.3$. Then, the transition probabilities of the Metropolis chain derived from the proposal chain is given by

$$p_{ij} = \begin{cases} q_{ij}min\left\{1, \frac{\pi_j}{\pi_i}\right\} & \text{if} j \neq i, \\ 1 - \sum_{k:k\neq i} q_{ik}min\left\{1, \frac{\pi_k}{\pi_i}\right\} & \text{if} j = i. \end{cases}$$

An extension of this example is given in [NM02].

Chapter 3

Convergence of Markov chains

Disclaimer : This chapter heavily relies on [GORJSR01]
The convergence of a Markov refers to the probability distribution of the Markov chain converging to the desired stationary distribution. That is, we see after how many runs does the probability distribution converge to the stationary distribution. We want to look at

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

Total variation distance is one such way to measure the convergence.

Definition 11. ***Total variation distance****[GORJSR01]*
Total variation distance between two probability measures μ_a, μ_b is defined as

$$\delta(\mu_a, \mu_b) = \|\mu_a(.) - \mu_b(.)\| = sup_{\mathbf{A}}|\mu_a(A) - \mu_b(A)|.$$

Now, how fast the Markov chain will converge will depend on how large n is. As n increases, we get closer to the stationary distribution. We need to know how large n should be, to get a fair enough convergence.

Proposition 3.0.1. *[GORJSR01] The following holds true*
(a) $\|\mu_1(.) - \mu_2(.)\| = sup_{f:\Omega \longrightarrow [0,1]} \left| \int f d\mu_1 - \int f d\mu_2 \right|$
(b) $\|\mu_1(.) - \mu_2(.)\| = \frac{1}{b-a} sup_{f:\Omega \longrightarrow [a,b]} \left| \int f d\mu_1 - \int f d\mu_2 \right|$ *for any* $a < b$.
In particular, $\|\mu_1(.) - \mu_2(.)\| = \frac{1}{2} sup_{f:\Omega \longrightarrow [-1,1]} \left| \int f d\mu_1 - \int f d\mu_2 \right|$
(c) If $\pi(.)$ *is a stationary for a Markov Chain kernel P, then,* $\|P^n(x,.) - \pi(.)\|$ *is non-decreasing in n, That is,* $\|P^n(x,.) - \pi(.)\| \leq \|P^{n-1}(x,.) - \pi(.)\|$ *for* $n \in \mathbb{N}$.
(d)More generally, letting $(\mu_i P)(A) = \int \mu_i(dx) P(x, A)$*, we always have*
$\|(\mu_1 P)(.) - (\mu_2 P)(.)\| \leq \|\mu_1(.) - \mu_2(.)\|$
(e)Let $t(n) = 2sup_{x \in \Omega} \|P^n(x,.) - \pi(.)\|$ *where* $\pi(.)$ *is stationary. Then t is sub-multiplicative. That is,* $t(m+n) \leq t(m)t(n)$ *for* $m, n \in \mathbb{N}$.
(f)If $\mu(.)$ *and* $\nu(.)$ *have densities g and h respectively with respect to some* σ*-finite measure* $\rho(.)$ *and* $M = max(g,h)$ *and* $m = min(g,h)$*, then,* $\|\mu(.) - \nu(.)\| = \frac{1}{2} \int_\Omega (M - m) d\rho = 1 - \int_\Omega m d\rho$.
(g)Given probability measures $\mu(.)$ *and* $\nu(.)$*, there are jointly defined random variables X and Y such that* $X \sim \mu(.)$ *and* $Y \sim \nu(.)$*, and* $P[X = Y] = 1 - \|\mu(.) - \nu(.)\|$.

Proof. The proof follows majorly from [GORJSR01] with explanations.
(a)Let $A = \{x \in \Omega : \mu_1(x) \geq \mu_2(x)$ and $\overline{A} = \Omega \backslash A$.
Now, note that, over $0 \leq f \leq 1$, the expression $\left| \int f d\mu_1 - \int f d\mu_2 \right|$ is maximized when
$(i) f = 1$ on A and $f = 0$ on $\overline{A}$ or it could be maximized when
$(ii) f = 1$ on $\overline{A}$ and $f = 0$ on A
In case (i) we have

$$\begin{aligned}
&sup_{f:\Omega \longrightarrow [0,1]} \left| \int_\Omega f d\mu_1 - \int_\Omega f d\mu_2 \right| \\
&= sup_{f:\Omega \longrightarrow [0,1]} \left| \int_A f d\mu_1 - \int_A f d\mu_2 + \int_{\overline{A}} f d\mu_1 - \int_{\overline{A}} f d\mu_2 \right| \\
&= \left| \int_A d\mu_1 - \int_A d\mu_2 + \int_{\overline{A}} 0 * d\mu_1 - \int_{\overline{A}} 0 * d\mu_2 \right| \\
&= |\mu_1(A) - \mu_2(A)|
\end{aligned}$$

Note that

$$\begin{aligned}
\left| \int_A d\mu_1 - \int_A d\mu_2 \right| &= \left| \{1 - \int_{\overline{A}} d\mu_1\} - \{1 - \int_{\overline{A}} d\mu_2\} \right| \\
&= \left| - \int_{\overline{A}} d\mu_1 + \int_{\overline{A}} d\mu_2 \} \right| \\
&= \left| \mu_1(\overline{A}) \mu_2(\overline{A}) \right|
\end{aligned}$$

Similarly, for case (ii) we have

$$\begin{aligned}
&sup_{f:\Omega\longrightarrow[0,1]}\left|\int_\Omega f d\mu_1 - \int_\Omega f d\mu_2\right| \\
&= sup_{f:\Omega\longrightarrow[0,1]}\left|\int_A f d\mu_1 - \int_A f d\mu_2 + \int_{\overline{A}} f d\mu_1 - \int_{\overline{A}} f d\mu_2\right| \\
&= \left|\int_A 0 * d\mu_1 - \int_A 0 * d\mu_2 + \int_{\overline{A}} d\mu_1 - \int_{\overline{A}} d\mu_2\right| \\
&= \left|\mu_1(\overline{A}) - \mu_2(\overline{A})\right| \\
&= |\mu_1(A) - \mu_2(A)|
\end{aligned}$$

Thus, we arrive at the following result.

$$sup_{f:\Omega\longrightarrow[0,1]}\left|\int_\Omega f d\mu_1 - \int_\Omega f d\mu_2\right| = |\mu_1(A) - \mu_2(A)| = |\mu_1(\overline{A}) - \mu_2(\overline{A})|.$$

Hence, we get that

$$\begin{aligned}
\|\mu_1(.) - \mu_2(.)\| &= sup_{x\in\Omega}|\mu_1(.) - \mu_2(.)| \\
&= sup_{f:\Omega\longrightarrow[0,1]}\left|\int f d\mu_1 - \int f d\mu_2\right|.
\end{aligned}$$

(b)This has a similar proof as that of part (a)
We again have two cases :
$(i) f = b$ on A and $f = a$ on $\overline{A}$ or it could be maximized when
$(ii) f = b$ on $\overline{A}$ and $f = a$ on A
Case (i) :

$$\begin{aligned}
&sup_{f:\Omega\longrightarrow[a,b]}\left|\int_\Omega f d\mu_1 - \int_\Omega f d\mu_2\right| \\
&= sup_{f:\Omega\longrightarrow[0,1]}\left|\int_A f d\mu_1 - \int_A f d\mu_2 + \int_{\overline{A}} f d\mu_1 - \int_{\overline{A}} f d\mu_2\right| \\
&= \left|\int_A b * d\mu_1 - \int_A b * d\mu_2 + \int_{\overline{A}} a * d\mu_1 - \int_{\overline{A}} a * d\mu_2\right| \\
&= \left|b(\mu_1(A) - \mu_2(A)) + a(\mu_1(\overline{A}) - \mu_2(\overline{A}))\right| \\
&= |b(\mu_1(A) - \mu_2(A)) + a(1 - \mu_1(A)) - a(1 - \mu_2(A))| \\
&= |(b-a)||(\mu_1(A) - \mu_2(A))| \\
&= |(b-a)||(\mu_1(\overline{A}) - \mu_2(\overline{A}))|
\end{aligned}$$

The last equality comes from a similar calculation as to that done in part (a).

Case (ii) :

$$
\begin{aligned}
&sup_{f:\Omega\longrightarrow[a,b]}\left|\int_\Omega f d\mu_1 - \int_\Omega f d\mu_2\right| \\
&= sup_{f:\Omega\longrightarrow[0,1]}\left|\int_A f d\mu_1 - \int_A f d\mu_2 + \int_{\overline{A}} f d\mu_1 - \int_{\overline{A}} f d\mu_2\right| \\
&= \left|\int_A a * d\mu_1 - \int_A a * d\mu_2 + \int_{\overline{A}} b * d\mu_1 - \int_{\overline{A}} b * d\mu_2\right| \\
&= \left|a(\mu_1(A) - \mu_2(A)) + b(\mu_1(\overline{A}) - \mu_2(\overline{A}))\right| \\
&= \left|a(1 - \mu_1(\overline{A})) - (1 - \mu_2(\overline{A})) + b(\mu_1(\overline{A}) - \mu_2(\overline{A}))\right| \\
&= |(a-b)||(\mu_1(\overline{A}) - \mu_2(\overline{A}))| \\
&= |(b-a)||(\mu_1(A) - \mu_2(A))|
\end{aligned}
$$

Hence we have $\|\mu_1(.) - \mu_2(.)\| = \frac{1}{(b-a)} sup_{f:\Omega\longrightarrow[a,b]}\left|\int f d\mu_1 - \int f d\mu_2\right|$.

(c)We need to prove : $\|P^n(x,.) - \pi(.)\| \leq \|P^{n-1}(x,.) - \pi(.)\|$ for $n \in \mathbb{N}$.

$$
\begin{aligned}
|P^n(x,A) - \pi(A)| &= \left|\int_{z\in\Omega} P^{n-1}(x,dy)P(y,A) - \int_{z\in\Omega} \pi(dy)P(y,A)\right| \\
&= \left|\int_{z\in\Omega} P^{n-1}(x,dy)f(y) - \int_{z\in\Omega} \pi(dy)f(y)\right| \\
&\leq \|P^{n-1}(x,.) - \pi(.)\|
\end{aligned}
$$

$f(y) = P(y,A)$ and hence $f : \Omega \longrightarrow [0,1]$ and the last inequality follows from (a) as :

$$
\begin{aligned}
&\left|\int_{z\in\Omega} P^{n-1}(x,dy)f(y) - \int_{z\in\Omega} \pi(dy)f(y)\right| \\
&\leq sup_{f:\Omega\longrightarrow[0,1]}\left|\int_{z\in\Omega} P^{n-1}(x,dy)f(y) - \int_{z\in\Omega} \pi(dy)f(y)\right| \\
&= \|P^{n-1}(x,.) - \pi(.)\| \qquad\qquad (from(a)).
\end{aligned}
$$

Hence we have that $|P^n(x,A) - \pi(A)| \leq \|P^{n-1}(x,.) - \pi(.)\|$ for all A, which implies,

$$
\begin{aligned}
\|P^n(x,.) - \pi(.)\| &= sup_A|P^n(x,A) - \pi(A)| \\
&\leq \|P^{n-1}(x,.) - \pi(.)\|.
\end{aligned}
$$

(d)Given that $(\mu_i P)(A) = \int \mu_i(dx)P(x, A)$. Then, by definition,

$$\begin{aligned}
\|(\mu_1 P)(A) - (\mu_2 P)(A)\| &= \left| \int_{x\in\Omega} \mu_1(dx)P(x, A) - \int_{x\in\Omega} \mu_2(dx)P(x, A) \right| \\
&= \left| \int_{x\in\Omega} \mu_1(dx)f(x) - \int_{x\in\Omega} \mu_2(dx)f(x) \right| \\
&\leq sup_{f:\Omega\longrightarrow[0,1]} \left| \int_{x\in\Omega} \mu_1(dx)f(x) - \int_{x\in\Omega} \mu_2(dx)f(x) \right| \\
&= \|\mu_1(A) - \mu_2(A)\|
\end{aligned}$$

The last equality comes from part (a). We can also get the result by directly applying part (c) to the LHS.

(e)Given $t(n) = 2sup_{x\in\Omega}\|P^n(x, .) - \pi(.)\|$. We need to show that
$t(m + n) \leq t(m)t(n)$ where $t(m + n) = 2sup_{x\in\Omega}\|P^{m+n}(x, .) - \pi(.)\|$.
Let $\hat{P}(x, .) = P^n(x, .) - \pi(.)$ and $\hat{Q}(x, .) = P^m(x, .) - \pi(.)$. Also, let f be a function such that $f : \Omega \longrightarrow [0, 1]$. Then,

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

Since $\int_{y\in\Omega} \pi(dy)P^m(y, dz) = \pi(dz)$ by the property of stationary distribution. Also, $\int_{y\in\Omega} \pi(dy) = 1$ and $\int_{y\in\Omega} P^n(x, dy) = 1$.
Now, let $g(x) = (\hat{Q}f)(x) = \int_{y\in\Omega} \hat{Q}(x, dy)f(y)$. Then,

$$\begin{aligned}
g^* = sup_{x\in\Omega}|g(x)| &= sup_{x\in\Omega} \left| \int_{y\in\Omega} \hat{Q}(x, dy)f(y) \right| \\
&\leq sup_{x\in\Omega} sup_{f:\Omega\longrightarrow[0,1]} \left| \int_{y\in\Omega} (P^m(x, dy) - \pi(dy)f(y) \right| \\
&= sup_{x\in\Omega}\|P^m(x, .) - \pi(.)\| = \frac{1}{2}t(m)
\end{aligned}$$

The last equality comes from the definition of t(m).
Now, if $g^* = 0$, then by definition, $sup_{x\in\Omega}|g(x)| = 0$ which in turn implies $sup_{x\in\Omega}(\hat{Q}f)(x) =$

0. Thus we have, $(\hat{P}\hat{Q}f)(x) = 0$. So, taking $g^* \neq 0$, we have that

$$\begin{aligned} 2sup_{x\in\Omega}|(\hat{P}\hat{Q}f)(x)| &= 2sup_{x\in\Omega}|\hat{P}(g)(x)| \\ &= 2g^* sup_{x\in\Omega}|\hat{P}(g/g*)(x)| \\ &\leq t(m)sup_{x\in\Omega}|\hat{P}(g/g*)(x)|. \end{aligned}$$

Since $g^* \leq \frac{1}{2}t(m)$.
Now, note that since $-1 \leq \hat{Q}(x, .) \leq 1$ and $0 \leq f \leq 1$
we have $-1 \leq g(x) \leq 1$ as $g(x) = (\hat{Q}f)(x)$. Also, since $g^* = sup_{x\in\Omega}|g(x)|$,

$$-1 \leq \frac{g(x)}{g^*(x)} \leq 1.$$

It follows from part (b) that,

$$\begin{aligned} \hat{P}(g/g*)(x) &\leq sup_{\frac{g}{g^*}:\Omega\longrightarrow[-1,1]}|(P^n(x, .) - \pi(.))(\frac{g}{g^*})| \\ &= 2\|P^n(x, .) - \pi(.)\| \end{aligned}$$

Thus, the above equation becomes,

$$sup_{x\in\Omega}|\hat{P}(g/g*)(x)| \leq 2sup_{x\in\Omega}\|P^n(x, .) - \pi(.)\| = t(n).$$

Recall that we got

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

Taking the supremum of this over x will give us the same result as taking the supremum over f. Hence, taking supremum of the above equation and applying (b) to this, we get that

$$sup_{x\in\Omega}|(\hat{P}\hat{Q}f)(x)| = sup_{f:\Omega\longrightarrow[0,1]}|(\hat{P}\hat{Q}f)(x)| = \|P^{m+n}(x, .) - \pi(.)\|$$

Similarly,

$$\begin{aligned} sup_{x\in\Omega}|\hat{P}(g/g*)(x)| &= sup_{\frac{g}{g^*}:\Omega\longrightarrow[-1,1]}|\hat{P}(g/g*)(x)| \\ &= 2\|P^n(x, .) - \pi(.)\| \end{aligned}$$

Hence,

$$\begin{aligned}
2sup_{x\in\Omega}\|P^{m+n}(x,.)-\pi(.)\| &= 2sup_{x\in\Omega}sup_{x\in\Omega}|(\hat{P}\hat{Q}f)(x)| \\
&= 2sup_{x\in\Omega}g^{*}sup_{x\in\Omega}|\hat{P}(g/g*)(x)| \\
&\leq 2sup_{x\in\Omega}\frac{1}{2}t(m)\{sup_{x\in\Omega}|\hat{P}(g/g*)(x)|\} \\
&= t(m)sup_{x\in\Omega}2\|P^{n}(x,.)-\pi(.)\| \\
&= t(m)t(n).
\end{aligned}$$

(f)Given that $\mu(.)$ has density g and $\nu(.)$ has density h. That is, we have $g = \frac{d\mu}{d\rho}$ $h = \frac{d\nu}{d\rho}$ and also, $M = max(g,h)$ and $m = min(g,h)$. Consider a function $f : \Omega \longrightarrow [-1,1]$. Let A be the subset of Ω in which $\mu(.) \geq \nu(.)$ Now, applying (b), we get

$$\begin{aligned}
\|\mu(.)-\nu(.)\| &= \frac{1}{2}sup_{f:\Omega\longrightarrow[-1,1]}\left|\int_{\Omega}fd\mu-\int_{\Omega}fd\nu\right| \\
&= \frac{1}{2}sup_{f:\Omega\longrightarrow[-1,1]}\left|\int_{\Omega}fgd\rho-\int_{\Omega}fhd\rho\right| \\
&= \frac{1}{2}sup_{f:\Omega\longrightarrow[-1,1]}\left|\int_{A}fgd\rho-\int_{A}fhd\rho+\int_{\overline{A}}fgd\rho-\int_{\overline{A}}fhd\rho\right| \\
&= \frac{1}{2}\left|\int_{A}(g-h)d\rho+\int_{\overline{A}}(h-g)d\rho\right| \\
&= \frac{1}{2}\left|\int_{A}(M-m)d\rho+\int_{\overline{A}}(M-m)d\rho\right| \\
&= \frac{1}{2}\int_{\Omega}(M-m)d\rho
\end{aligned}$$

For the next inequality, note that $M + m = g + h$. Hence,

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

Therefore, we can write

$$\begin{aligned}
\frac{1}{2}\int_{\Omega}(M-m)d\rho &= 1-\frac{1}{2}\left(2-\int_{\Omega}(M-m)d\rho\right) \\
&= 1-\frac{1}{2}\left(\int_{\Omega}(M+m)d\rho-\int_{\Omega}(M-m)d\rho\right) \\
&= 1-\frac{1}{2}\int_{\Omega}2md\rho = 1-\int_{\Omega}md\rho
\end{aligned}$$

(g)For proving this part, let us first look at the following lemma.(Lemma 2 from [CD11]).

Lemma 3.0.2. *(**Coupling Lemma**) Let μ and ν be two probability measures over a finite set Ω. Then, for any coupling ω of (μ, ν), if the random variable (X, Y) is distributed according to ω, then $P(X \neq Y) \geq \|\mu - \nu\|$.*

Proof. Fix any coupling ω of μ and ν and let X and Y be distributed with respect to this coupling. Then, for any $z \in \Omega$,

$$\begin{aligned}\mu(z) = P(X = z) &= P(X = z, Y = X) + P(X = z, Y \neq X) \\ &\leq P(Y = z) + P(X = z, Y \neq X) \\ &= \nu(z) + P(X = z, Y \neq X).\end{aligned}$$

Hence, $\mu(z) - \nu(z) \leq P(X = z, Y \neq X)$. Similarly, we have that $\nu(z) - \mu(z) \leq P(Y = z, X \neq Y)$

$$\begin{aligned}\nu(z) = P(Y = z) &= P(Y = z, X = Y) + P(Y = z, X \neq Y) \\ &\leq P(X = z) + P(Y = z, X \neq Y) \\ &= \mu(z) + P(Y = z, X \neq Y).\end{aligned}$$

Therefore, we can write

$$\begin{aligned}2\|\mu - \nu\| &= \sum_{z \in \Omega} |\mu(z) - \nu(z)| \\ &= \sum_{\substack{z \in \Omega; \\ \mu(z) \geq \nu(z)}} \mu(z) - \nu(z) + \sum_{\substack{z \in \Omega; \\ \mu(z) < \nu(z)}} \nu(z) - \mu(z) \\ &\leq \sum_{\substack{z \in \Omega; \\ \mu(z) \geq \nu(z)}} P(X = z, Y \neq X) + \sum_{\substack{z \in \Omega; \\ \mu(z) < \nu(z)}} P(Y = z, X \neq Y) \\ &\leq P(Y \neq X) + P(Y \neq X)\end{aligned}$$

Thus, we get $P(X \neq Y) \geq \|\mu - \nu\|$. ∎

Now, note that part (g) of our proposition now follows directly from the above lemma.

$$\begin{aligned}P(X = Y) &= 1 - P(X \neq Y) \\ &\leq 1 - \|\mu - \nu\|\end{aligned}$$

∎

Lemma 3.0.3. *For two probability measures μ and ν,when μ and ν are two distribution functions corresponding to two probability mass functions $p = \{p_x\}_{x \in \Omega}$ and $q = \{q_x\}_{x \in \Omega}$,*

so that for every measurable A such that $A \subseteq \Omega$, the total variation distance is given by $d_{TV}(\mu, \nu) = max_A|\mu(A) - \nu(A)|$ and we have

$$\mu(A) = \sum_{x \in A} p_x \qquad\qquad \nu(A) = \sum_{x \in A} q_x,$$

Then,

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |p_x - q_x|.$$

Proof. We know

$$d_{TV}(\mu, \nu) = \|\mu(.) - \nu(.)\| = max_A|\mu(A) - \nu(A)|.$$

Let $A = \{x \in \Omega : p(x) \geq q(x)\}$. Then,

$$\begin{aligned} d_{TV}(\mu, \nu) = max_A|\mu(A) - \nu(A)| &= max_A \left| \sum_{x \in A} p_x - \sum_{x \in A} q_x \right| \\ &= max_A \left| \sum_{x \in A} p_x - q_x \right| \\ &= max_A \sum_{x \in A} |p_x - q_x| \qquad \text{(Since } p(x) \geq q(x) \; \forall \; x \in A) \quad (\blacklozenge) \end{aligned}$$

Similarly, note that

$$\begin{aligned} d_{TV}(\mu, \nu) &= max_A|\mu(A) - \nu(A)| \\ &= max_A|(1 - \mu(A^C)) - (1 - \nu(A^C))| = max_A|\mu(A^C) - \nu(A^C)| \\ &= max_A \left| \sum_{x \in A^C} p_x - \sum_{x \in A^C} q_x \right| \\ &= max_A \left| \sum_{x \in A^C} p_x - q_x \right| \\ &= max_A \sum_{x \in A^C} |p_x - q_x| \qquad \text{(Since } p(x) \leq q(x) \; \forall \; x \in A^C) \quad (\lozenge) \end{aligned}$$

Summing ($\blacklozenge$) and ($\lozenge$), we get,

$$\begin{aligned} 2d_{TV}(\mu, \nu) &= max_A \sum_{x \in A} |p_x - q_x| + max_A \sum_{x \in A^C} |p_x - q_x| \\ &= \sum_{x \in \Omega} |p_x - q_x| \end{aligned}$$

■

A similar Lemma to this is :

Lemma 3.0.4. *When F and G are the distribution functions corresponding to two continuous densities $f = \{f(x)\}_{x\in\mathbb{R}}$ and $g = \{g(x)\}_{x\in\mathbb{R}}$, so that for every measurable $A \subseteq \mathbb{R}$,*

$$\mu(A) = \int_{x\in A} f(x)dx \qquad\qquad \nu(A) = \int_{x\in A} g(x)dx,$$

Then,

$$d_{TV}(\mu, \nu) = \frac{1}{2}\int_{-\infty}^{\infty} |f(x) - g(x)|dx.$$

Proof. The proof follows along the lines of *Lemma 3.0.3.* stated above. ∎

Theorem 3.0.5. ***Asymptotic Convergence Theorem*** *[GORJSR01]*
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 π-a.e. $x \in \Omega$,

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

In particular, $lim_{n\longrightarrow\infty}P^n(x, A) = \pi(A)$ for all measurable $A \subseteq \Omega$

Corollary 3.0.5.1. *[GORJSR01] If a Markov chain is ϕ-irreducible with period $d \geq 2$, and has a stationary distribution $\pi(.)$, then for π-a.e. $x \in \Omega$*

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

Proof. Note that the chain is ϕ- irreducible. Let the chain have the periodic decomposition $\Omega_1, \Omega_2, ..., \Omega_d \subseteq \Omega$ and let P' be the d-step chain P^d with its state space restricted to the set Ω_1.
Then, P' is ϕ- irreducible and aperiodic on Ω_1, with stationary distribution π' which satisfies

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

From *Proposition 3.0.1(c)*, it is enough to prove the corollary taking $n = md$ with $m \longrightarrow \infty$ and, without loss of generality, we assume that $x \in \Omega_1$.
Note that using *Proposition 3.0.1(d)*, we can write

$$\|P^{md+j}(x, .) - \pi' P^j(.)\| \leq \|P^{md}(x, .) - \pi'(.)\| \qquad \text{for } j \in \mathbb{N}$$

Then,

$$
\begin{aligned}
&\left\| \frac{1}{d} \sum_{i=md}^{md+d-1} P^i(x,.) - \pi(.) \right\| \\
&= \left\| \frac{1}{d} \sum_{j=0}^{d-1} P^{md+j}(x,.) - \pi(.) \right\| \qquad \text{(take } i = md + j\text{)} \\
&\leq \left\| \frac{1}{d} \sum_{j=0}^{d-1} P^{md+j}(x,.) - \frac{1}{d} \sum_{j=0}^{d-1} (\pi' P^j)(.) \right\| \\
&\leq \frac{1}{d} \left\| \sum_{j=0}^{d-1} \left[P^{md+j}(x,.) - (\pi' P^j)(.) \right] \right\| \quad \text{(Triangle Inequality)} \\
&\leq \frac{1}{d} \left\| \sum_{j=0}^{d-1} \left[P^{md}(x,.) - \pi'(.) \right] \right\| \qquad \text{(using } \textit{Proposition 3.0.1(d)}\text{)}
\end{aligned}
$$

Now, applying *Theorem 3.0.5* to P' will give us

$$\lim_{m \longrightarrow \infty} \| P^{md}(x,.) - \pi'(.) \| = 0 \qquad \text{for } \pi - \text{a.e. } x \in \Omega_1$$

Thus, giving us the desired result :

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

■

3.1 Rate of Convergence

We have so far looked at the convergence of the Markov chain to stationary. Uniform ergodicity is a way to qualitatively measure the rate of convergence.

3.1.1 Uniform Ergodicity

Definition 12. *[GORJSR01] A Markov chain with stationary distrbution $\pi(.)$ is* ***uniformly ergodic*** *if*

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

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

An equivalent form of uniform ergodicity is discussed below in the proposition.

Proposition 3.1.1. *[GORJSR01] A Markov chain with stationary distribution $\pi(.)$ is uniformly ergodic if and only if*

$$sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| < \frac{1}{2} \quad \textit{for some } n \in \mathbb{N}$$

Proof. According to the definition of uniform ergodicity, if the chain is uniformly ergodic, then,

$$\begin{aligned} lim_{n\longrightarrow\infty} sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| &\leq lim_{n\longrightarrow\infty} M\rho^n \\ &= 0 \qquad\qquad (\text{since } \rho < 1) \end{aligned}$$

Thus, we will directly get that

$$sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| < \frac{1}{2}$$

Now, to prove the converse, consider

$$sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| < \frac{1}{2} \qquad \text{for some } n \in \mathbb{N}$$

Let $d(n) = 2sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\|$. Then,

$$\frac{1}{2}d(n) = sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| < \frac{1}{2} \qquad (\text{Initial assumption})$$

$$\Rightarrow d(n) < \frac{1}{4}$$

Thus, $d(n) \equiv \beta < 1$.
Then, using *Proposition 3.0.1 (e)*, we have that $d(n)$ is submultiplicative such that for all $j \in \mathbb{N}$,

$$d(jn) \leq (d(n))^j = \beta^j$$

Hence, using *Proposition 3.0.1(c)*,

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

Hence, the chain is uniformly ergodic with $M = \frac{1}{\beta}$ and $\rho = \beta^{\frac{1}{n}}$. ∎

Remark 3.1.2. *[GORJSR01] The above proposition will still continue to hold even if we replace the $\frac{1}{2}$ in the equation with any $0 < \delta < \frac{1}{2}$. However, it will cease to hold once $\delta > \frac{1}{2}$. To understand this, consider the following example*

Example: [GORJSR01] Note that if $\delta > \frac{1}{2}$, say $\delta = \frac{2}{3}$ then we will have $sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| < \frac{2}{3}$ then, we have

$$\begin{aligned} d(n) &= 2sup_{x\in\Omega}\|P^n(x,.) - \pi(.)\| \\ &< 2 * \frac{2}{3} = \frac{4}{3} \not< 1 \end{aligned}$$

Thus, we cannot use *Proposition 3.0.1 (e)* to prove that the minorisation condition holds. For further conditions that ensures uniform ergodicity, we shall first look at a new definition

Definition 13. *[GORJSR01] A subset $C \subseteq \Omega$ is small (or (n_0, ϵ, ν)- small) if $\exists$ a positive integer n_0, $\epsilon > 0$, and a probability measure $\nu(.)$ on Ω such that the following minoristation condition holds.*

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

That is; $P^{n_o}(x,A) \geq \epsilon\nu(A)\ \forall\ x \in C$ and all measurable $A \subseteq \Omega$. The minorisation condition essentially points out that there is always an ϵ-sized component in common between all of the n_o-step transitions.
Consider Ω to be countable and take

$$\epsilon_{n_o} \equiv \sum_{y\in\Omega} inf_{x\in C} P^{n_o}(x,\{y\}) > 0$$

then, C is $(n_o, \epsilon_{n_o}, \nu)$- small where

$$\begin{aligned} \nu(\{y\}) &= \frac{1}{(\epsilon_{n_o})} inf_{x\in C} P^{n_o}(x,\{y\}) \\ &= \frac{inf_{x\in C} P^{n_o}(x,\{y\})}{\sum_{y\in\Omega} inf_{x\in C} P^{n_o}(x,\{y\}) > 0} \end{aligned}$$

Now, suppose the transition probabilities have probability densities with respect to some measure, say $\eta(.)$, that is,

$$P^{n_o}(x,dy) = p_{n_o}(x,y)\eta(dy)$$

then, we can write our ϵ_{n_o} as

$$\epsilon_{n_o} = \int_{y\in\Omega} \eta(dy) inf_{x\in\Omega} p_{n_o}(x,y)$$

Theorem 3.1.3. *[GORJSR01] Consider a Markov chain with invariant probability distri-*

bution $\pi(.)$. Suppose the minorisation condition is satisfied for some $n_0 \in \boldsymbol{N}$ and $\epsilon > 0$ and the probability measure $\nu(.)$, in the special case $C = \Omega$ (i.e., the entire state space is small). Then the chain is uniform ergodic, and in fact $||P^n(x,.) - \pi(.)|| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \Omega$, where $\lfloor r \rfloor$ is the greatest integer not exceeding r.

In case of a discrete subspace, we will have

$$\|P^n(x,.) - \pi(.)\| \leq (1 - \epsilon_{n_o})^{\lfloor n/n_o \rfloor}$$

where

$$\epsilon_{n_o} \equiv \sum_{y \in \Omega} inf_{x \in C} P^{n_o}(x, \{y\})$$

Observe that we get a qualitative bound for the convergence to stationary from *Theorem 3.1.3*. That is;

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

Thus, once we are able to lay our hands on an n_o such that the above holds, then we can further, find an n_* such that

$$\|P^{n_*}(x,.) - \pi(.)\| \leq 0.01$$

Then, we can say that n_* iterations of the Markov chain is sufficient for convergence.

3.1.2 Geometric Ergodicity

Definition 14. *[GORJSR01] A Markov chain with stationary distribution $\pi(.)$ is **geometrically ergodic** if*

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

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

Remark 3.1.4. *[GORJSR01] Note that if Ω is finite, then all irreducible and aperiodic Markov chains are geometrically as well as uniformly ergodic.*

Definition 15. *[GORJSR01] Given Markov chain transition probabilities P on a state space Ω, and a measurable function $f :\longrightarrow \mathbb{R}$, define the function $Pf :\longrightarrow \mathbb{R}$ such that $(Pf)(x)$ is the* ***conditional expected value*** *of $f(\Omega_{n+1})$, given that $\Omega_n = x$. That is,*

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

Definition 16. *[GORJSR01] A Markov chain satisfies a* ***drift condition*** *if $\exists$ a small set C of invariant measure, constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \Omega \longrightarrow [1, \infty]$ such that*

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

That is, such that

$$\int_\Omega P(x, dy)V(y) \leq \lambda V(x) + b\mathbf{1}_C(x) \;\; \forall x \in \Omega$$

Theorem 3.1.5. *[GORJSR01] Consider a $\phi-$ irreducible, aperiodic Markov chain with stationary distribution $\pi(.)$. Suppose the minorisation condition is satisfied for some $C \subset \Omega$ and $\epsilon > 0$ and probability measure $\nu(.)$. Suppose further that the drift condition is satisfied for some constants $0 < \lambda < 1$ and $b < \infty$, and a function $V : \Omega \to [0, \infty]$ with $V(x) < \infty$ for at least one (and hence for $\pi - a.e.$) $x \in \Omega$. Then the chain is geometrically ergodic.*

Example [GORJSR01] Is Metropolis algorithm on $\mathbb{R}$ geometrically ergodic?.
Consider $\Omega = \mathbb{R}^+$ and $\pi_u(x) = e^{-x}$. Take the proposal distribution to be the symmetric distribution about x. That is, $q(x, y) = q(|y - x|)$ with $y \in [x - a, x + a]$ as the support.
Now, taking the drift function as $V(x) = e^{rx}$ for some $r > 0$. For $x \geq a$, we get $PV(x)$ as

$$\begin{aligned} PV(x) = &\int_{x-a}^{x} V(y)q(x, y)dy + \int_{x}^{x+a} V(y)q(x, y)dy\frac{\pi_u(y)}{\pi_u(x)} \\ &+ V(x)\int_{x}^{x+a} q(x, y)dy\left\{1 - \frac{\pi_u(y)}{\pi_u(x)}\right\} \qquad (*) \end{aligned}$$

Proof: Let $\alpha = min\left\{1, \frac{\pi_u(y)}{\pi_u(x)}\right\}$. We can then write

$$PV(x) = \int_{x-a}^{x+a} V(y)q(x, y)\alpha dy + V(x)\int_{x-a}^{x+a} q(x, y)(1 - \alpha)dy \qquad (**)$$

Note that:

(a) for $y \in [x - a, x]$, $\frac{\pi_u(y)}{\pi_u(x)} = \frac{e^{-y}}{e^{-x}} = e^{x-y}$ where $x \geq y$. This implies that $e^{x-y} \geq 1$.
Hence, $\alpha = min\left\{1, \frac{\pi_u(y)}{\pi_u(x)}\right\} = 1$ and $1 - \alpha = 0$

(b) for $y \in [x, x+a]$, $\frac{\pi_u(y)}{\pi_u(x)} = \frac{e^{-y}}{e^{-x}} = e^{x-y}$ where $x \leq y$. This implies that $e^{x-y} \leq 1$.
Hence, $\alpha = min\left\{1, \frac{\pi_u(y)}{\pi_u(x)}\right\} = \frac{\pi_u(y)}{\pi_u(x)}$ and $1 - \alpha = 1 - \frac{\pi_u(y)}{\pi_u(x)}$

Putting in the respective values of α and $(1-\alpha)$ for the intervals $y \in [x-a, x]$ and $y \in [x, x+a]$ in (**), we get the desired result.
Note that q is symmetric. We have $q(x, y) = q(|y - x|)$. Then, observe that $q(x, x+a) = q(|x + a - x|) = q(x - a - x) = q(x, x - a)$. Also, consider $\epsilon < a$. Then, $x - \epsilon > x - a$ and $x + \epsilon < x + a$ and we get $q(x, x+\epsilon) = q(|x + \epsilon - x|) = q(x - \epsilon - x) = q(x, x - \epsilon)$ hence verifying that q is symmetric.
Now, consider $\int_{x-a}^{x} V(t)q(x,t)dt$ and here, do variable substitution by putting $t = 2x - y$. Then, we get that $dt = -dy$, the lower limit as $x + a$, the upper limit as x, and now, looking at $q(x, t)$, we see that

$$\begin{aligned} q(x,t) = q(x, 2x-y) &= q(|x - (2x-y)|) \\ &= q(|y - x|) = q(x, y) \end{aligned}$$

Thus, we can write equation (*) as follows

$$PV(x) = \int_{x}^{x+a} \left[V(2x-z) + V(z)\frac{\pi_u(z)}{\pi_u(x)} + V(x)\left(1 - \frac{\pi_u(z)}{\pi_u(x)}\right)\right] q(x,z)dz$$

Let $I(x, z) = V(2x - z) + V(z)\frac{\pi_u(z)}{\pi_u(x)} + V(x)\left(1 - \frac{\pi_u(z)}{\pi_u(x)}\right)$. Putting in values $V(x) = e^{rx}$ and $\pi_u(x) = e^{-x}$ in $I(x, z)$ and simplifying, we get

$$\begin{aligned} I(x,z) &= e^{r(2x-z)} + \frac{e^{-z}e^{rz}}{e^{-x}} + e^{rx}\left(1 - \frac{e^{-z}}{e^{-x}}\right) \\ &= e^{x}e^{(r-1)z} + e^{2rx}e^{-rz} + e^{rx}\left(1 - e^{x-z}\right) \\ &= \frac{e^{rx}}{e^{rx}}e^{ry}e^{x-z} + e^{cx}e^{r(x-z)} + e^{rx}(1 - e^{x-z}) \\ &= e^{rx}\left(e^{ru}e^{-u} + e^{-ru} + 1 - e^{-u}\right) \qquad \text{(take } u = z - x\text{)} \\ &= e^{rx}\left(e^{(r-1)u} + e^{-ru} + 1 - e^{-u}\right) \\ &= e^{rx}\left(2 - \left(1 + e^{(r-1)u}\right)\left(1 - e^{-ru}\right)\right) \end{aligned}$$

For $r < 1$, note that $0 < 1 - r < 1$. Hence, for any $k > 0$, we will have $e^{-(1-r)k} < 1$, $e^{-ru} < 1$

and hence $1 - e^{-ru} > 0$. Thus, we can write $I(x, y)$ as

$$\begin{aligned} I(x,y) &= e^{rx}\left(2 - \left(1 + e^{(r-1)u}\right)(1 - e^{-ru})\right) \\ &= 2V(x)\left(1 - \frac{1}{2}\left(1 + e^{-(1-r)u}\right)(1 - e^{-ru})\right) \\ &\leq 2V(x)(1-\epsilon) \qquad \text{(for some positive constant } \epsilon) \end{aligned}$$

Thus, $\forall x > a$,

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

since $q(x, .)$ is a probability measure and also, q is symmetric about x.
Thus, we have $PV(x) \leq (1-\epsilon)V(x)$. Then, for $x \in [0, a]$; $V(0) = 1$ and $V(a) = e^{ca}$. Clearly, $V(a) > V(0)$ and so, $PV(x) \leq (1-\epsilon)V(a)$. Hence, $PV(x)$ is bounded on $[0, a]$.
Now, to show that $[0, a]$ is small.

$$\begin{aligned} P(x, dy) &\geq q(x,y)\, dy\, min\left\{1, \frac{\pi_u(y)}{\pi_u(x)}\right\} \\ &\geq \epsilon\, dy\, min\left\{1, \frac{\pi_u(y)}{K}\right\} \end{aligned}$$

where $\epsilon = inf_{x\in C, y\in\Omega} q(x, y) > 0\ \forall\ y \in \Omega$ and $K = \int_{y\in\Omega} \pi_u(y)dy$
Hence we get that $C = [0, a]$ is small.
This shows that the drift condition holds and thus, the algorithm is geometrically ergodic by *Theorem 3.1.5*.

3.2 Quantitative Convergence Rates

We are looking for a quantitative ways to bound the convergence rates of the Markov chains to the stationary distribution. That is we need something like

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

We will prove a result for the bound of convergence rates and we require the ***bivariate drift condition*** in order to prove it

Definition 17. *[GORJSR01] The **Bivariate drift condition** that we require is of the form*

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

for some funtion $h : \Omega \times \Omega \longrightarrow [1, \infty)$ *and some* $\alpha > 1$*, where*

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

$\bar{P}$ *is essentially running two independent copies of the chain.*

Proposition 3.2.1. *[GORJSR01] Suppose the univariate drift condition is satisfied for some* $V : \Omega \to [1, \infty]$*,* $C \subseteq \Omega$*,* $\lambda < 1$*, and* $b < \infty$*. Let* $d = \inf_{x \in C^c} V(x)$*. Then if* $d > [b/(1-\lambda)]-1$*, then the bivariate drift condition is satisfied for the same C, with* $h(x, y) = \frac{1}{2}[V(x) + V(y)]$ *and* $\alpha^{-1} = \lambda + b/(d + 1) < 1$*.*

Proof. Refer to [GORJSR01]. ∎

Now, we state a few assumptions to state the main result. Let

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

where, for $(x, y) \notin C \times C$

$$\bar{R}h(x, y) = \int_{\Omega} \int_{\Omega} (1 - \epsilon)^{-2} \left(P^{n_0}(x, dz) - \epsilon\nu(dz)\right) \left(P^{n_0}(y, dw) - \epsilon\nu(dw)\right)$$

We now state the main result:

Theorem 3.2.2. *[GORJSR01] Consider a Markov chain on a state space* Ω*, having transition kernel P. Suppose ther is* $C \subseteq \Omega$*,* $h : \Omega \times \Omega \to [1, \infty)$*, a probability distribution* $\nu(.)$ *on* Ω*,* $\alpha > 1$*,* $n_0 \in \mathbb{N}$*, and* $\epsilon > 0$*, such that the minorisation condition*

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

and the bivariate drift condition

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

hold. Define B_{n_0} *as*

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

where, for $(x, y) \notin C \times C$ *. 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 initial distribution* $\mathcal{L}(X_0, X_0')$*, then*

$$\|\mathcal{L}(X_k) - \mathcal{L}(X_k')\|_{TV} \leq (1 - \epsilon)^j + \alpha^{-k}(B_{n_0})^{j-1}\mathbb{E}[h(X_0, X_0')]$$

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

3.3 Coupling Construction for proving Quantitative Convergence Rates and Proofs

The Coupling Inequality[GORJSR01] The main idea of coupling is: consider two random variables, say Y and Z, who has joint distribution on some space Ω. Then, we can write the laws for these variables $\mathcal{L}(Y)$ and $\mathcal{L}(Z)$ for their respective probability distributions and we can calculate the total variation distance between them as :

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

Coupling construction: [GORJSR01] Assume that C is a small set. The following coupling construction is called the "splitting technique" by [EN78] and [KBAPN78]; see also [EN84] and [SPMRLT93]. The basic idea is to generate two copies of the Markov chain, sampling both from the same probability distribution such that the construction of their joint distribution gives a high probability for them to be close to each other.

Start with the initial distribution $X_0 = x$ and $X_0' \sim \pi(.)$, and $n = 0$. Now, the following loop is repeated for long in order to generate the Markov chains.

- If $X_n = X_n'$, choose $X_n = X_n' \sim P(X_n, .)$ and update n to $n + 1$.
- Else, if $(X_n, X_n') \in C \times C$, then:
  - with probability ϵ choose $X_{n+n_0} = X_{n+n_0}' \sim \nu(.)$;
  - or, with probability $1 - \epsilon$, conditionally, independently choose

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

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

For the case when $n_0 > 1$, in order to complete the sequence, we can return to construct

$X_{n+1}X_{n+2}, ..., X_{n+n_0-1}$ from their corresponding conditional probability distributions given X_n and X_{n+n_0}. Similarly, we can also construct $X'_{n+1}X'_{n+2}, ..., X'_{n+n_0-1}$ from their corresponding conditional probability distributions given X'_n and X'_{n+n_0}.

PROOFS OF QUANTITATIVE CONVERGENCE RATES

Here, we give the proofs of a few theorems that were stated above.

Proof of **Theorem 3.1.3** [GORJSR01]

Proof. We have $C = \Omega$. So, according to the coupling construction, in every n_0 iterations we have a probability of ϵ for making the Markov chains equal. Therefore, if $n = n_0 m$, then we have

$$P(X_n \neq X'_n) \leq (1-\epsilon)^m = (1-\epsilon)^{n/n_0}$$

By *Proposition 3.0.1 (c)*, since the probability distribution $\pi(.)$ is stationary for the kernel P, then, $\|P^n(x,.) - \pi(.)\|$ is non increasing. That is,

$$\|P^n(x,.) - \pi(.)\| \leq \|P^{n-1}(x,.) - \pi(.)\| \text{for } n \in \mathbb{N}.$$

■

Proof of **Theorem 3.2.2** [GORJSR01]

Proof. In the minorisation condition, assume $n_0 = 1$ for the small set C. Then, let $B_{n_0} = B$. Consider

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

Also, let the time corresponding to the consecutive visits of (X_k, X'_k) to $C \times C$ be $\tau_1, \tau_2, ...$ Then, for $j \in \mathbb{N}$ such that $1 \leq j \leq k$,

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

Observe that the first event on the right side of equation (1) : $\{X_k \neq X'_k, N_{k-1} \geq j\}$ is contained in : first j consecutive coin flips came up tails. That is, the event of getting more that j tails is contained in the event of getting j tails. Hence,

$$P(X_k \neq X'_k, N_{k-1} \geq j) \leq (1-\epsilon)^j \qquad (2)$$

Now, to find a bound for the second term in equation (1), let

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

Let us look at a lemma now.

Lemma 3.3.1. *[GORJSR01] We have*

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

That is, M_k is a supermartingale.

Proof: Let us proceed by looking at the two cases : $(X_k, X'_k) \notin C \times C$ and $(X_k, X'_k) \in C \times C$.
Case 1 : $(X_k, X'_k) \notin C \times C$: Refer to [GORJSR01]
Case 2 : $(X_k, X'_k) \in C \times C$: There are two sub cases here. That is : i. $X_k = X'_k$ and ii. $X_k \neq X'_k$
Case i: Trivial.
Now, suppose $X_k \neq X'_k$. Then, we get

$$\begin{aligned}
&\mathbb{E}[M_{k+1}|X_0, ...X_k, X'_0, ...X'_k] \\
&= \alpha^{k+1} B^{-N_{k-1}-1} \mathbb{E}[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_{k+1} \neq X'_{k+1})|X_k, X'_k] \\
&= \alpha^{k+1} B^{-N_{k-1}-1}(1-\epsilon)\bar{R}h(X_k, X'_k) \qquad (\star) \\
&= \alpha^{k+1} B^{-N_{k-1}-1}(1-\epsilon)\bar{R}h(X_k, X'_k)\frac{M_k}{\alpha^k B^{-N_{k-1}} h(X_k, X'_k)\mathbf{1}(X_k \neq X'_k)} \\
&= M_k \left[\frac{\alpha(1-\epsilon)\bar{R}h(X_k, X'_k)}{B}\right]\frac{1}{h(X_k, X'_k)} \qquad (\text{Since } \mathbf{1}(X_k \neq X'_k) = 1) \\
&\leq \frac{M_k}{h(X_k, X'_k)} \qquad (\star\star) \\
&\leq M_k \qquad \left(\text{Since } h \text{ is defined as } h : \Omega \times \Omega \to [1, \infty). \text{ Hence } \frac{1}{h} \leq 1\right)
\end{aligned}$$

$(\star\star)$: To prove the second inequality
We know that $B = max\left\{1, \alpha(1-\epsilon)sup_{C\times C}\bar{R}h\right\}$. If $B = 1 \implies \frac{1}{B} = 1$ and $\alpha(1-\epsilon)sup_{C\times C}\bar{R}h < 1$. Thus,

$$\left[\frac{\alpha(1-\epsilon)\bar{R}h(X_k, X'_k)}{B}\right] \leq 1$$

and, if $B = \alpha(1-\epsilon)sup_{C\times C}\bar{R}h$; then again, we have that the above inequality is satisfied.
$(\star)$: To prove : $\mathbb{E}[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_{k+1} \neq X'_{k+1})|X_k, X'_k] = (1-\epsilon)\bar{R}h(X_k, X'_k)$ We have $X_k \neq X'_k$ and, $(X_k, X'_k) \in C \times C$. Thus, we have two possibilities :

(i)We can either choose both X_k and X'_k to be equal

$$X_{k+1} = X'_{k+1} \sim P(X_k, .) \Longrightarrow \mathbf{1}(X_{k+1} \neq X'_{k+1}) = 0$$

(ii)Or we can choose X_k and X'_k from their respective distributions (note that $n_0 = 1$).

$$\begin{aligned} X_{k+1} &\sim \frac{1}{(1-\epsilon)}[P(X_k, .) - \epsilon\nu(.)] \text{ and} \\ X'_{k+1} &\sim \frac{1}{(1-\epsilon)}[P(X'_k, .) - \epsilon\nu(.)] \\ &\Longrightarrow \mathbf{1}(X_{k+1} \neq X'_{k+1}) = 1 \end{aligned}$$

Thus,we will get

$$\begin{aligned} &\mathbb{E}[h(X_{k+1}, X'_{k+1})\mathbf{1}(X_{k+1} \neq X'_{k+1})|X_k, X'_k] = \\ &\int_\Omega \int_\Omega h(X_{k+1}, X'_{k+1}) \frac{[P(X_k, dX_{k+1}) - \epsilon\nu(.)]}{(1-\epsilon)} \\ &\qquad \frac{[P(X'_k, dX'_{k+1}) - \epsilon\nu(.)]}{(1-\epsilon)} \cdot P(\mathbf{1}(X_{k+1} \neq X'_{k+1})) \\ &= \bar{R}h(X_k, X'_k) \cdot (1-\epsilon) \\ &(\text{Since } P(\mathbf{1}(X_{k+1} \neq X'_{k+1})) = P(X_{k+1} \neq X'_{k+1}) = (1-\epsilon).) \end{aligned}$$

Hence, we have proved that $\{M_k\}$ is a Martingale.
Back to proof of **Theorem 3.2.2** :
Note that $B \geq 1$. Then, we get that

$$\begin{aligned} &P(X_k \neq X'_k, N_{k-1} < j) \\ &\leq \alpha^{-k} B^{(j-1)} \mathbb{E}\left[h(X_0, X'_0)\right] \qquad (3) \end{aligned}$$

See [GORJSR01] for the details of the proof. Combining (1), (2) and (3), we get :

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

Hence proved.
Now, this proof was for the case when $n_0 = 1$. For $n_0 > 1$, we do not want to consider completing the chain. That is, we skip constructing $X_{n+1}, ..., X_{n+n_0}$ as well as $X'_{n+1}, ..., X'_{n+n_0}$. Thus, in the proof replace N_{k-1} by N_{k-n_0} and prove that $M_{t(k)}$ is a supermartingale where

$t(k)$ is the largest time $\leq k$ and such that it is not a time of 'stay'. We have that

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

Also, we have :

$$P(X_k \neq X'_k) = P(X_k \neq X'_k, N_{k-n_0} \geq j) + P(X_k \neq X'_k, N_{k-n_0} < j).$$

Just like before, since the event that $\{X_k \neq X'_k, N_{k-n_0} \geq j\}$ is contained : first j coin flips gave tails, we get

$$P(X_k \neq X'_k, N_{k-n_0} \geq j) \leq (1-\epsilon)^j.$$

Now, let

$$M_{t(k)} = \alpha^{(k)} B_{n_0}^{-N_{k-n_0}} h(X_k, X'_k) \mathbf{1}(X_k \neq X'_k)$$

Going along the lines of the lemma before, we will get that $M_{t(k)}$ is a supermartingale. Now, proceeding like before, we get that, since $B_{n_0} \geq 1$,

$$\begin{aligned}
&P(X_k \neq X'_k, N_{k-n_0} < j) \\
&= P(X_k \neq X'_k, N_{k-n_0} \leq j-1) \\
&\leq P(X_k \neq X'_k, B_{n_0}^{-N_{k-n_0}} \geq B_{n_0}^{-(j-1)}) \\
&= P(\mathbf{1}(X_k \neq X'_k) B_{n_0}^{-N_{k-n_0}} \geq B_{n_0}^{-(j-1)}) \\
&\leq B_{n_0}^{-(j-1)} \mathbb{E}[\mathbf{1}(X_k \neq X'_k) B_{n_0}^{-N_{k-n_0}}] \qquad \text{(Markov's Inequality)} \\
&\leq B_{n_0}^{-(j-1)} \mathbb{E}[\mathbf{1}(X_k \neq X'_k) B_{n_0}^{-N_{k-n_0}} h(X_k, X'_k)] \qquad \text{(Since} h \geq 1\text{)}. \\
&= \alpha^{-k} B_{n_0}^{-(j-1)} \mathbb{E}[M_{t(k)}] \\
&\leq \alpha^{-k} B_{n_0}^{-(j-1)} \mathbb{E}[M_{t(0)}] \\
&= \alpha^{-k} B_{n_0}^{-(j-1)} \mathbb{E}[h(X_0, X'_0)] \qquad \text{(Since } N_{-n_0} = 0\text{)}
\end{aligned}$$

Hence proved. ∎

Proof of **Theorem 3.1.5** [GORJSR01]

Proof. Set $h(x,y) = \frac{1}{2}[V(x) + V(y)]$. We use the following lemma for proving this theorem.

Lemma 3.3.2. *[GORJSR01] We may assume without loss of generality that*

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

Specifically, given a small set C and a drift function V satisfying the minorisation condition

and the univariate geometric drift condition, we can find a small set $C_0 \subseteq C$ such that the minorisation condition and the univariate geometric drift condition hold (with the same n_0) and ϵ and b, but with λ replaced by some $\lambda_0 < 1$), and such that ★ also holds.

Proof. Interested readers are requested to refer [GORJSR01] for proof. ■

Now, assume that $sup_{x\in C}V(x) < \infty$. Let $sup_{x\in C}V(x) = m$. Then, from drift condition, $PV \leq \lambda V + b\mathbf{1}_C$. We have that

$$
\begin{aligned}
&sup_{(x,y)\in C\times C}\bar{R}h(x,y)\\
&= sup_{(x,y)\in C\times C}\int_\Omega\int_\Omega (P^{n_0}(x,dz)-\epsilon\nu(dz))(P^{n_0}(y,dw)-\epsilon\nu(dw))h(z,w)(1-\epsilon)^{-2}\\
&= sup_{(x,y)\in C\times C}\int_\Omega\int_\Omega (P^{n_0}(x,dz)-\epsilon\nu(dz))(P^{n_0}(y,dw)-\epsilon\nu(dw))\frac{[V(z)+V(w)]}{2(1-\epsilon)^2}\\
&\leq sup_{(x,y)\in C\times C}\int_\Omega\int_\Omega (P^{n_0}(x,dz)-\epsilon\nu(dz))(P^{n_0}(y,dw)-\epsilon\nu(dw))\frac{m}{(1-\epsilon)^2}\\
&< \infty \qquad \text{(Since } sup_{x\in C}V(x) < \infty .i.e.; m < \infty \text{ and minorisation condition holds.)}
\end{aligned}
$$

Therefore, we have $sup_{(x,y)\in C\times C}\bar{R}h(x,y) < \infty$. Thus, we get that

$$B_{n_0} = max\{1, \alpha^{n_0}(1-\epsilon)sup_{C\times C}\bar{R}h(x,y)\} < \infty.$$

Now, let $d = inf_{C^c}V$. Then, from *Proposition 3.2.1*, we have that if univariate drift condition is satisfied and for d as above, then, the bivariate drift condition is satisfied for $h(x,y) = \frac{1}{2}[V(x)+V(y)]$ and $\alpha^{-1} = \lambda + \frac{b}{(d+1)} < 1$ if $d > \frac{b}{(1-\lambda)} - 1$. Then, we have that by *Theorem 3.2.2*, if the minorisation condition and bivariate drift condition is satisfied, then,

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

Then *Theorem 3.1.5* is proved in this case. That is, if we take $\rho = \frac{1}{\alpha}$ and $M(.,.) = B^{(j-1)_{n_0}}\mathbb{E}\left[h(X_0, X_0')\right]$

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

Now, for $d \leq \frac{b}{(1-\lambda)} - 1$, we cannot use this argument because $d > \frac{b}{(1-\lambda)} - 1$ ensures aperiodicity of the chain and without this condition, we will have to assume the aperiodicity of the chain in the proof. So, we try to make C large so that we improve the chance of $d = inf_{C^c}V > \frac{b}{(1-\lambda)} - 1$. That is, we enlarge C so that d satisfies this condition and we then use aperiodicity to show that C remains a small set which means that the minorisation condition holds, perhaps for very large n_0 and smaller $\epsilon > 0$. Then, we will have that *Theorem 3.1.5* follows from *Proposition 3.2.1* and *Theorem 3.2.2* like before.
Now, choose some $d' > \frac{b}{1-\lambda} - 1$. Let $S = \{x \in \Omega : V(x) \leq d'\}$ and we also consider the set

$C' = C \cup S$. This will give us :

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

We also have that V is bounded on S by definition. Thus, $sup_{x \in C'} V(x) < \infty$. Thus, we get $sup_{(x,y) \in C' \times C'} \bar{R}h(x,y) < \infty$ and hence also, $B_{n_0} < \infty$.
So, we have that *Theorem 3.1.5* follows from *Proposition 3.2.1* and *Theorem 3.2.2* if we can prove that C' is small.

Lemma 3.3.3. *[GORJSR01] C' is a small set.*

In order to prove this lemma, we look at the following definition.

Definition 18. *[GORJSR01] A subset $C \subseteq \Omega$ is petite (or, (n_0, ϵ, ν)- petite), relative to a small set C, if there exists a positive integer $n_0, \epsilon > 0$, and a probability measure $\nu(.)$ on Ω such that*

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

The petite set is similar to the small set except that here, the different states in C are allowed to cover the minorisation measure $\epsilon\nu(.)$ at different times i.
Any small set is a petite set. However, the converse is not true for all cases. This is because the condition for being petite does not rule out the periodicity of the chain.
For an aperiodic, ϕ-irreducible Markov chain, we have the following lemma.

Lemma 3.3.4. *[GORJSR01] For an aperiodic, ϕ- irreducible Markov chain, all petite sets are small sets.*

We need the following lemma in order to use *Lemma 3.3.4*

Lemma 3.3.5. *[GORJSR01] Let $C' = C \cup \mathcal{S}$ where $\mathcal{S} = \{x \in \Omega; V(x) \leq d\}$ for some $d < \infty$, as above. Then, C' is petite.*

Proof. Refer [GORJSR01] for proof. ■

Now, we have that C' is petite and since the chain is aperiodic and ϕ-irreducible, by *Lemma 3.3.4*, we have that C' is small. Hence proving *Theorem 3.1.5*. ■

Proof of **Theorem 3.0.5*Asymptotic Convergence Theorem*** [GORJSR01]

Proof. In order to prove this, assume the following theorem :

Theorem 3.3.6. *[GORJSR01] Every ϕ-irreducible Markov chain, on a state space with countably generated σ-algebra, contains a small set $C \subseteq \Omega$ with $\phi(C) > 0$. (In fact, each $B \subseteq \Omega$ with $\phi(B) > 0$ in turn contains a small set $C \subseteq B$ with $\phi(C) > 0$.) Furthermore, the minorisation measure $\nu(.)$ may be taken to satisfy $\nu(C) > 0$.*

Proof. See [EN84],p. 16. ■

Lemma 3.3.7. *[GORJSR01] Consider a Markov chain on a state space Ω, having stationary distribution $\pi(.)$. Suppose that for some $A \subseteq \Omega$, we have $P_x(\tau_A < \infty) > 0$ for all $x \in \Omega$. Then, for π-almost-every $x \in \Omega$, $P_x(\tau_A < \infty) = 1$.*

Proof. Refer [GORJSR01]. ■

The rest of the proof follows as in [GORJSR01] ■

Lemma 3.3.8.

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

Proof. Refer [GORJSR01]. ■

Chapter 4

Applications

4.1 Introduction

In this chapter, we illustrate an application of Markov chain Monte Carlo method by applying it to a covariance realisation problem for a discrete random process.

Consider a random vector $\sigma \in \mathbb{R}^{\kappa}$ where each entry takes values in $\{-k, -k+1, ..., 1, 0, -1, ..., k-1, k\}$. There are $(2k+1)^n$ such vectors. By σ we denote a generic n-tuple of this type. Let C denote the variance-covariance matrix of this random vector. It is not clear whether any symmetric positive semi-definite matrix C, can manifest as a variance-covariance matrix for such discrete random vector. This problem has been studied in detail for spin random variables. We do not ponder on this question and throughout this chapter, we assume that we are given a matrix C which can be realized as a variance covariance matrix of the given random vector by some probability distribution. In other words, we assume that there exists a probability distribution on the space $\{-k, -k+1, ..., 1, 0, -1, ..., k-1, k\}^n$ such that the given matrix C is its covariance matrix.

Note that a probability distribution realizing C as the covariance matrix need not be unique. The idea is to find one such probability distribution, namely, the one that maximizes the entropy. Given a matrix C, assuming that it is realizable, we first obtain the explicit form of the maximum entropy probability distribution and then use MCMC to explicitly find it.

4.2 Maximum entropy problem for discrete random variables on a general state space

Let $\Omega = \{\sigma = (\sigma_i) : \sigma_i \in \{-k, -k+1, ..., 1, 0, -1, ..., k-1, k\}\}$ for some fixed $k \in \mathbb{N}$. The length of each σ in Ω is fixed to be n.
The entropy of the system described above is given by

$$\mathcal{S}(P) = -\sum_{\sigma} P(\sigma)\log P(\sigma)$$

***Consider the following problem*:** Find a probability distribution P^* on Ω such that $P^* = argmin\mathcal{S}(P)$ subject to the following constraints.

$$\begin{aligned} c_{h,k} &= \sum_{\sigma} \sigma_h \sigma_k P(\sigma) \\ 1 &= \sum_{\sigma} P(\sigma) \\ e_i &= \sum_{\sigma} \sigma_i P(\sigma) \end{aligned}$$

where $c_{h,k}$ denotes the (h, k)th entry of the variance-covariance matrix.
The problem at hand is to find the right probability distribution such that the entropy of the system is maximized. We refer to this problem as $\mathbb{P}_0$. This has been done previously in the paper [PDPMPNS13], for spin systems using the Lagrange multiplier method.

4.3 Lagrange Multiplier Method

Here, we briefly discuss the Lagrange Multiplier method. Suppose we have to maximize or minimize a function $f(x, y, z)$ subject to the constraint $g(x, y, z) = k$ for some constant k. Then, we write an equation of the form :

$$L(x, y, z, \lambda) = f(x, y, z) - \lambda(g(x, y, z) - k)$$

where L is called the ***Lagrangian*** and λ is a new variable called the ***Lagrange multiplier***.

The solution :
Take the partial derivatives of the Lagrangian with respect to the variables x, y, z and λ, equate all of it to zero and the put the values obtained back into the original function in order to get the desired answer.
That is, solve for $L_x = 0; L_y = 0; L_z = 0; L_\lambda = 0$ and substitute the solutions back into $f(x, y, z)$.

Now, we try to solve $\mathbb{P}_0$. Observe that according to the Lagrangian Multiplier method stated earlier in first chapter, we get that the Lagrangian function is given by

$$\mathcal{L}(P(\sigma)) = \Lambda(P(\sigma)) + \mathcal{S}(P(\sigma))$$

where the Lagrangian functional $\Lambda(P(\sigma))$ is

$$\begin{aligned}\Lambda(P) = \sum_{h,k} \lambda_{hk} \left(c_{hk} - \sum_{\sigma} \sigma_h \sigma_k P(\sigma) \right) + \mu \left(1 - \sum_{\sigma} P(\sigma) \right) \\ + \sum_{i} \eta_i \left(e_i - \sum_{\sigma} \sigma_i P(\sigma) \right)\end{aligned}$$

where $\lambda_{hk}, \mu, \eta_i$ are Lagrange Multipliers for all h, k, i. Thus, now we have

$$\begin{aligned}\mathcal{L}(P) = \sum_{h,k} \lambda_{hk} \left(c_{hk} - \sum_{\sigma} \sigma_h \sigma_k P(\sigma) \right) + \mu \left(1 - \sum_{\sigma} P(\sigma) \right) \\ + \sum_{i} \eta_i \left(e_i - \sum_{\sigma} \sigma_i P(\sigma) \right) - \sum_{\sigma} P(\sigma) \log P(\sigma)\end{aligned}$$

Note that we can avoid all the constant terms from $\mathcal{L}(P)$ and write

$$\begin{aligned}\mathcal{I}(P) = - \sum_{\sigma} P(\sigma) \log P(\sigma) - \sum_{h,k} \lambda_{hk} \sum_{\sigma} \sigma_h \sigma_k P(\sigma) \\ - \mu \sum_{\sigma} P(\sigma) - \sum_{i} \eta_i \sum_{\sigma} \sigma_i P(\sigma)\end{aligned}$$

Taking the partial derivative of $\mathcal{I}$ with respect to $P(\sigma)$, we get

$$\begin{aligned}\partial_P \mathcal{I}(P(\sigma)) &= -(1 + \log P(\sigma)) - \sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \mu - \sum_i \eta_i\sigma_i \\ &= -\log P(\sigma) - \sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i - (\mu + 1)\end{aligned}$$

Solving for $P(\sigma)$ we get that the gradient of $\mathcal{I}$ is 0 at

$$P^*(\sigma) = \exp\left\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\right\} e^{-(\mu+1)}$$

Thus, we arrive at

$$P^*(\sigma) = \frac{1}{Z}\exp\left\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\right\} \tag{A}$$

$$\text{where } Z = e^{-(\mu+1)}. \tag{B}$$

Now, note that

$$\begin{aligned}P^*(-\sigma) &= \frac{1}{Z}\exp\left\{-\sum_{h,k} \lambda_{hk}(-\sigma_h)(-\sigma_k) - \sum_i \eta_i(-\sigma_i)\right\} \\ &= \frac{1}{Z}\exp\left\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k + \sum_i \eta_i\sigma_i\right\} \\ &\neq P^*(\sigma)\end{aligned}$$

Hence, we have that the mean is not 0 with respect to P^*. Note that the probability distribution depends on the Lagrange Multipliers. Thus, the problem boils down to finding the right λ, where λ is the vector of all Lagrange Multipliers.
Now, we move on to the dual problem.

4.4 Dual Problem

We have $\sum_\sigma P(\sigma) = 1$. Also, taking $P(\sigma) = \frac{1}{Z}\exp\left\{-\sum_{h,k}\lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\right\}$, we get that

$$\begin{aligned}\Lambda(P) &= \mu\left(1 - \sum_\sigma P(\sigma)\right) + \sum_i \eta_i\left(e_i - \sum_\sigma \sigma_i P(\sigma)\right) + \sum_{h,k}\lambda_{hk}\left(c_{hk} - \sum_\sigma \sigma_h\sigma_k P(\sigma)\right)\\ &= \sum_i \eta_i e_i - \sum_i\sum_\sigma \eta_i\sigma_i P(\sigma) + \sum_{h,k}\lambda_{hk}c_{hk} - \sum_{h,k}\lambda_{hk}\sum_\sigma \sigma_h\sigma_k P(\sigma) \qquad \text{(C)}\\ &\quad \left(\text{Since } \sum_\sigma P(\sigma) = 1\right)\end{aligned}$$

Similarly,

$$\begin{aligned}\mathcal{S}(P) &= -\sum_\sigma P(\sigma)\log P(\sigma)\\ &= -\sum_\sigma \frac{1}{Z}\exp\left\{-\sum_{h,k}\lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\right\}\\ &\quad -\left\{-\sum_{h,k}\lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i - (\mu+1)\right\}\\ &= \sum_\sigma P(\sigma)\left\{\sum_{h,k}\lambda_{hk}\sigma_h\sigma_k + \sum_i \eta_i\sigma_i\right\} + \sum_\sigma \log Z\, P(\sigma)\\ &= \sum_\sigma\sum_{h,k}\lambda_{hk}P(\sigma)\sigma_h\sigma_k + \sum_\sigma\sum_i P(\sigma)\eta_i\sigma_i + \log Z \qquad \text{(D)}\\ &\quad \left(\text{Since } \sum_\sigma P(\sigma) = 1\right)\end{aligned}$$

Adding the two, we get $\mathcal{L}(P)$ as

$$\begin{aligned}\mathcal{L}(P) &= \sum_i \eta_i e_i - \sum_i\sum_\sigma \eta_i\sigma_i P(\sigma) + \sum_{h,k}\lambda_{hk}c_{hk} - \sum_{h,k}\lambda_{hk}\sum_\sigma \sigma_h\sigma_k P(\sigma) +\\ &\quad \sum_\sigma\sum_{h,k}\lambda_{hk}P(\sigma)\sigma_h\sigma_k + \sum_\sigma\sum_i P(\sigma)\eta_i\sigma_i + \log Z\\ &= \log Z + \sum_i \eta_i e_i + \sum_{h,k}\lambda_{hk}c_{hk} \qquad \text{(E)}\end{aligned}$$

Observe that

$$1 = \sum_{\sigma} P(\sigma) = \sum_{\sigma} \frac{1}{Z} \exp\left\{ - \sum_{h,k} \lambda_{hk} \sigma_h \sigma_k - \sum_{i} \eta_i \sigma_i \right\}$$

$$\implies Z = \sum_{\sigma} \exp\left\{ - \sum_{h,k} \lambda_{hk} \sigma_h \sigma_k - \sum_{i} \eta_i \sigma_i \right\} \tag{F}$$

Thus, we have that

$$P(\sigma) = \frac{\exp\left\{ - \sum_{h,k} \lambda_{hk} \sigma_h \sigma_k - \sum_i \eta_i \sigma_i \right\}}{\sum_{\sigma} \exp\left\{ - \sum_{h,k} \lambda_{hk} \sigma_h \sigma_k - \sum_i \eta_i \sigma_i \right\}} \tag{G}$$

Writing (E) as a function of the Lagrange Multipliers,

$$\mathcal{J}(\lambda) = \sum_{h,k} \lambda_{hk} c_{hk} + \sum_{i} \eta_i e_i + \log\left\{ \sum_{\sigma} \exp\left\{ - \sum_{h,k} \lambda_{hk} \sigma_h \sigma_k - \sum_{i} \eta_i \sigma_i \right\} \right\} \tag{H}$$

Next, we take the gradient of this function with respect to each of the Lagrange Multipliers. The ij^{th} row here represents $\frac{\partial \mathcal{J}(\lambda)}{\partial \lambda_{ij}}$ and the l^{th} row represents $\frac{\partial \mathcal{J}(\lambda)}{\partial \eta_l}$

$$\nabla \mathcal{J}(\lambda) = \begin{bmatrix} . \\ . \\ . \\ c_{ij} + 0 + \sum_\sigma \left\{ \frac{\exp\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\}}{\sum_\sigma \exp\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\}} * (-\sigma_i\sigma_j) \right\} \\ . \\ . \\ 0 + e_l + \sum_\sigma \left\{ \frac{\exp\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\}}{\sum_\sigma \exp\{-\sum_{h,k} \lambda_{hk}\sigma_h\sigma_k - \sum_i \eta_i\sigma_i\}} * (-\sigma_l) \right\} \\ . \\ . \\ . \end{bmatrix}$$

$$= \begin{bmatrix} . \\ . \\ . \\ c_{ij} - \sum_\sigma \sigma_i\sigma_j P(\sigma) \\ . \\ . \\ e_l - \sum_\sigma \sigma_l P(\sigma) \\ . \\ . \\ . \end{bmatrix}$$

Thus, $\nabla \mathcal{J}(\lambda) = 0$ gives us that $P(\sigma)$ satisfies the constraints. This λ^*, for which the above result holds, will be the critical point of $\mathcal{J}(.)$.

Hence, the critical point of $\nabla \mathcal{J}(\lambda)$, *i.e.* λ^*, if it exists, will give us the desired probability distribution $P^*(.)$ on the system such that the constraints are satisfied and such that the entropy of the system is maximized.

The discrete time dynamical system is given by

$$\lambda_{n+1} = \lambda_n - \frac{\nabla \mathcal{J}}{K}$$

and for $K > N$, this converges to the right λ^*.

The proof of this convergence is beyond the scope of this thesis.

The next section discusses some simulations that were run based on the theory discussed above. We look at spin system as well as an extended state space for which we try to approach the Entropy Maximisation problem.

Simulations

The transition probability was taken to be

$$P(X^{n+1}|X^n) = \frac{e^{-<G(X^{n+1}),\lambda>}}{e^{-<G(X^n),\lambda>}}$$

4.5 Discrete Random variables on a given state space

Given $C = ((c_{ij}))$, a realizable covariance matrix for the process described earlier in the chapter. We want to explicitly determine the maximum entropy probability distribution. The form of that is given in equation (G). Note that the maximum entropy probability distribution is parametrized by the matrix λ.S0, the problem boils down to explicitly determining the corresponding λ^*. The following algorithm converges to the correct λ^*. The proof of which is out of the scope of this thesis.
We now look at the state space $\Omega = \{\sigma = (\sigma_1, ..., \sigma_n) : \sigma_i \in \{-k, -k+1, ..., -1, 0, 1, ...k-1, k\}\}$. $G(\sigma)$ is calculated as :

$$G(\sigma) = (1, \sigma_1\sigma_2, \sigma_1\sigma_3, .., , \sigma_{n-1}\sigma_n)$$

and, λ, a vector of length = $|G|$, is the set of all lagrangian multipliers.
R is defined as the covariance vector where it contains the elements of the variance covariance matrix.

$$R = [1, c_{12}, c_{13}, ..., c_{n-1,n}].$$

$u(\lambda)$ gives an estimate of R for a given λ.
Deterministic Algorithm

- Start with an initial λ_1
- Start with r=1
- $u(\lambda_r) = \sum_\sigma \frac{1}{Z} G e^{-G^T \lambda_r}$
- $\lambda_{r+1} = \lambda_r - \frac{R-u(\lambda_r)}{K}$

- Put $r = r + 1$ and repeat till $r = H$ for some H large enough

The output of this deterministic algorithm will give us the optimum λ^* which would in turn give us the probability distribution P^* that maximizes the entropy of the system.
This involves a lot of time and as much as $(2k + 1)^n$ calculations in each loop (for $u(\lambda_r)$ and Z). To tackle this issue, we give a Markov chain Monte Carlo algorithm. The Markov chain for a fixed λ is described below:

The Markov chain

The state space is $\Omega = \{\sigma = (\sigma_1, ..., \sigma_n) : \sigma_i \in \{-k, -k + 1, ..., -1, 0, 1, ...k - 1, k\}\}$. and $G(\sigma)$ is defined as above.

- Start with an initial $X_0 = \sigma'$ for some $\sigma' \in \Omega$.
- Generate a proposal $X_{n+1} = \tilde{\sigma}$ for a given $X_n = \sigma$ with transition probability as :

$$P(\sigma, \tilde{\sigma}) = \frac{e^{-<G(\tilde{\sigma}),\lambda>}}{e^{-<G(\sigma),\lambda>}}$$

The stationary distribution of this Markov chain is exactly of the form of the probability distribution that satisfies the Maximum Entropy problem. This Markov chain is

- Reversible : We saw earlier in *Chapter 2* that Metropolis algorithms are reversible.
- Irreducible : Clearly, by construction of the Markov chain, one can see that we can go from any state to any other state.
- Recurrent
- Ergodic. The chain is both uniformly as well as geometrically ergodic. This can be proved using the theory discussed in *Chapter3*

MCMC algorithm

- Start with an initial λ_1
- Put $r = 1$
- Put $u = 0$
- Set an initial value $\sigma^{(0)}$for the Markov Chain (Y)
- Run the Markov chain for about 100 runs to get a $Y_t = \sigma^{(t)}$ in each run, take the transition probability to be $\tilde{P}(\lambda_r) = \frac{e^{G(\sigma^{(t)})^T \lambda_{r-1}}}{e^{G(\tilde{\sigma}^{(t)})^T \lambda_{r-1}}}$
- After each run of the MC, compute $u(\lambda_r) = \sum_{\sigma^{(t)}} \frac{1}{Z} G e^{-G(\sigma^{(t)})^T \lambda_r}$
  with $Z = \sum_{\sigma^{(t)}} e^{(-G(\sigma^{(t)})^T)\lambda}$

- $\lambda_{r+1} = \lambda_r - \frac{R-u(\lambda_r)}{K}$
- Put $r = r + 1$ and repeat till $r = H$ for some H large enough

We simulate the above algorithm for $n = 5, 6, 7$ are given below with the following parameters fixed: R, λ. The state space is $\Omega = \{\sigma = (\sigma_1, ..., \sigma_n) : \sigma_i \in \{-k, -k + 1, ..., -1, 0, 1, ...k - 1, k\}\}$.
The number of iterations for deterministic algorithm is (d) : 5000
The number of iterations for the Markov chain (MC) : we have considered two cases : 10000; 100000
The number of iterations for the convergence of λ (outer loop) in MCMC (l) : 1000

Table showing the final u value for different algorithms for $n = 5$

R	[1, 0.10,0.20,0.20,0.30,0.25, 0.40,0.20,0.30,0.40,0.20]	Error = $\|\|R - u\|\|$	Timelapsed
Deterministic	[1,0.10,0.20,0.20,0.30,0.25, 0.40,0.20,0.30,0.40,0.20]	1.11E-15	248.6124
MC : 10000; l : 1000	[1,0.07,0.15,0.21,0.31,0.29, 0.44,0.17,0.30,0.40,0.26]	0.2365	59.1452
MC : 100000; l : 1000	[1,0.07,0.19,0.22,0.31,0.23, 0.39,0.18,0.31,0.41,0.20]	0.0574	599.3587

Table showing the final u value for different algorithms for $n = 7$

R	[1,0.10,0.20,0.20,0.30,0.25,0.15,0.25,0.40,0.20,0.30, 0.10,0.30,0.40,0.20,0.25,0.20,0.35,0.10,0.15,0.35,0.20]	Error = $\|\|R - u\|\|$	Timelapsed
Deterministic	[1,0.10,0.20,0.20,0.30,0.25,0.15,0.25,0.40,0.20,0.30, 0.10,0.30,0.40,0.20,0.25,0.20,0.35,0.10,0.15,0.35,0.20]	7.70E-16	1224.1
MC : 10000 l : 1000	[1,0.05,0.20,0.18,0.28,0.28,0.17,0.27,0.48,0.13,0.24, 0.11,0.34,0.30,0.17,0.21,0.17,0.37,0.11,0.13,0.30,0.26]	0.2013	61.8768
MC : 100000; l : 1000	[1,0.13,0.20,0.21,0.30,0.25,0.14,0.21,0.42,0.21,0.29, 0.10,0.30,0.41,0.18,0.26,0.23,0.31,0.12,0.11,0.32,0.20]	0.0903	619.6289

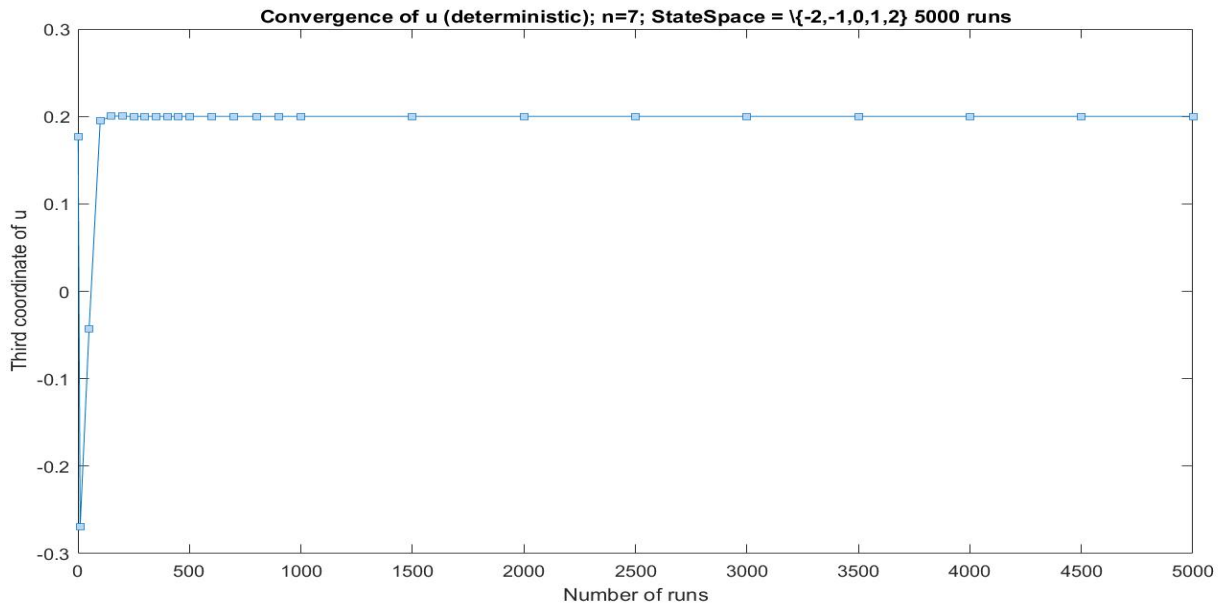

Figure 4.1: Convergence of u, n=7,{-2,-1,0,1,2} d:5000 deterministic

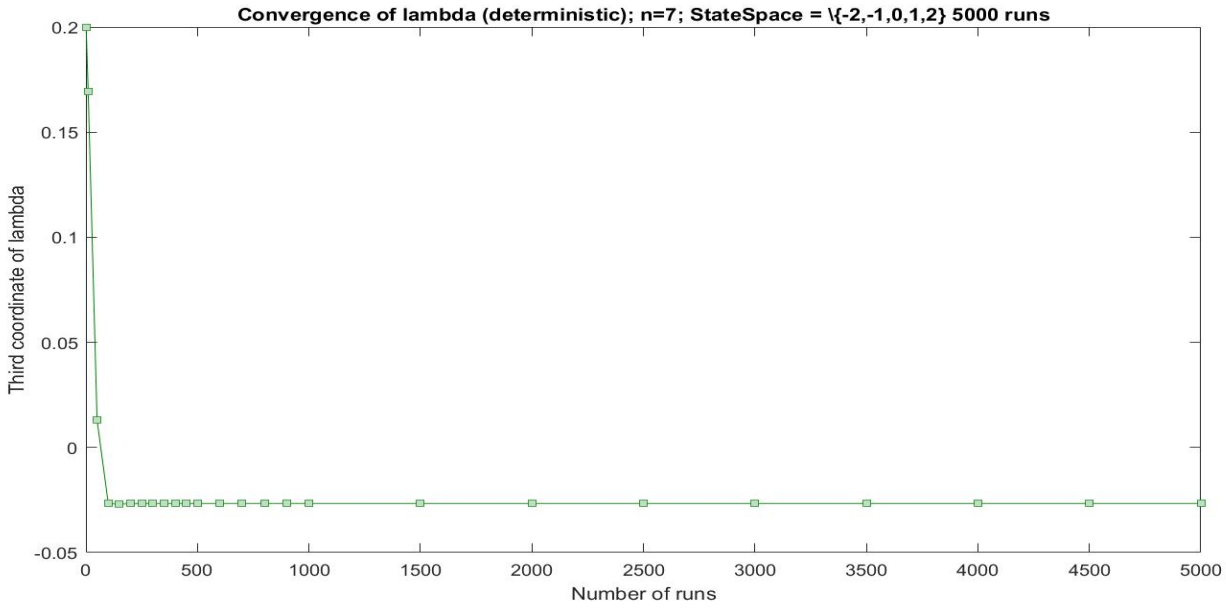

Figure 4.2: Convergence of λ, n=7, {-2,-1,0,1,2} d:5000 deterministic

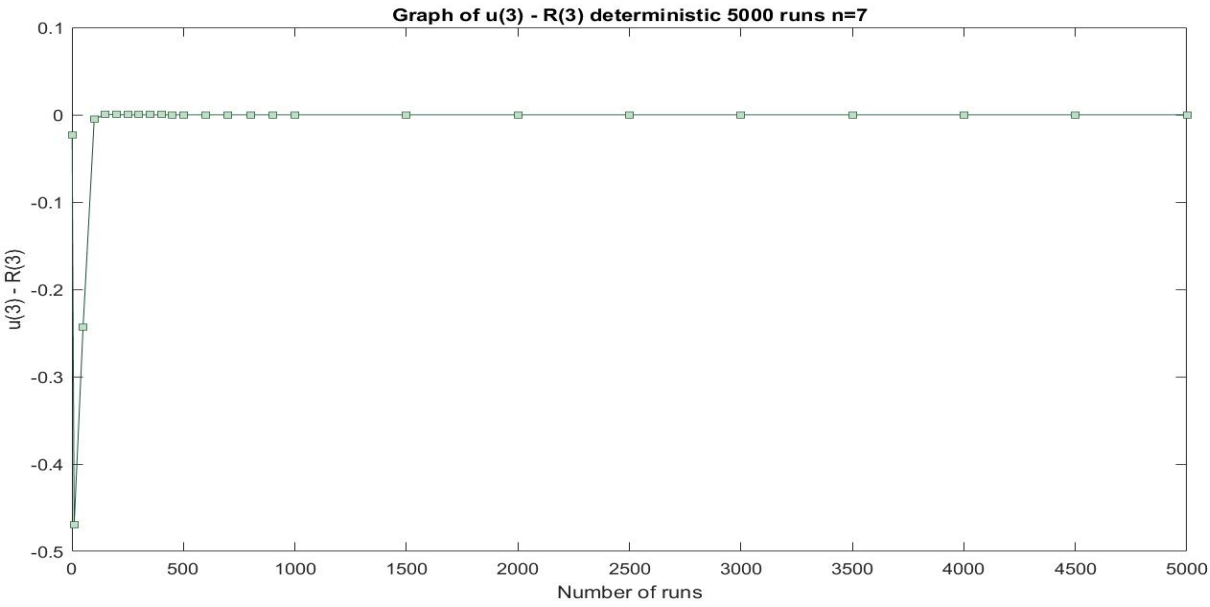

Figure 4.3: Graph of u(3) - R(3), n=7, {-2,-1,0,1,2} d:5000 deterministic

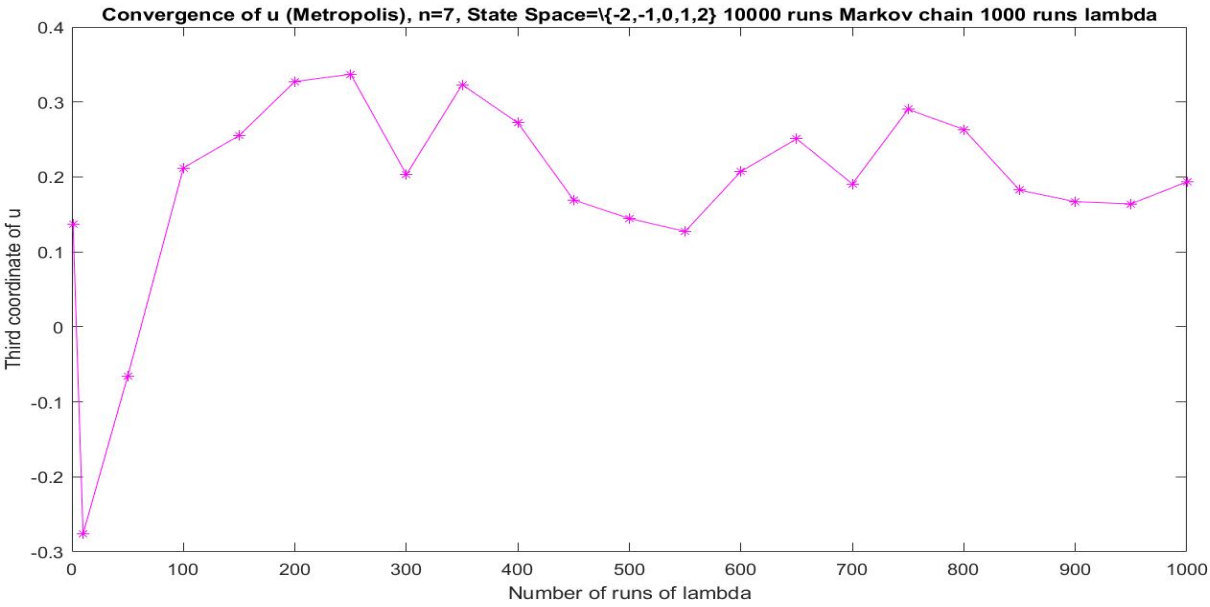

Figure 4.4: Convergence of u (MCMC), n=7,{-2,-1,0,1,2} MC:10000 l:1000

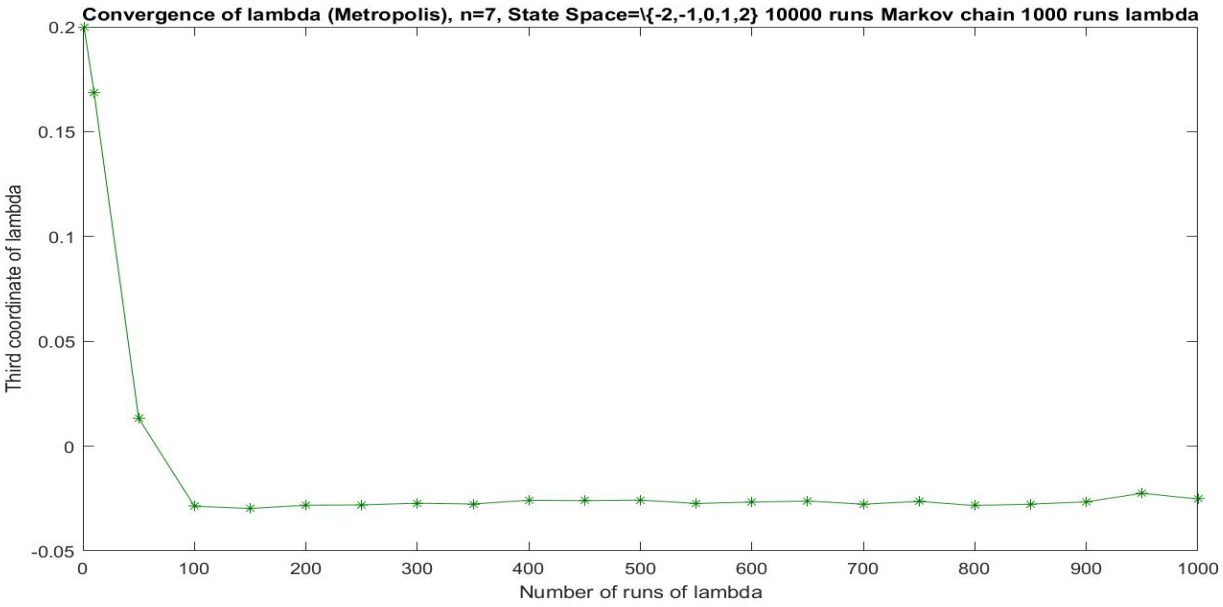

Figure 4.5: Convergence of λ (MCMC), n=7, {-2,-1,0,1,2} MC:10000 l:1000

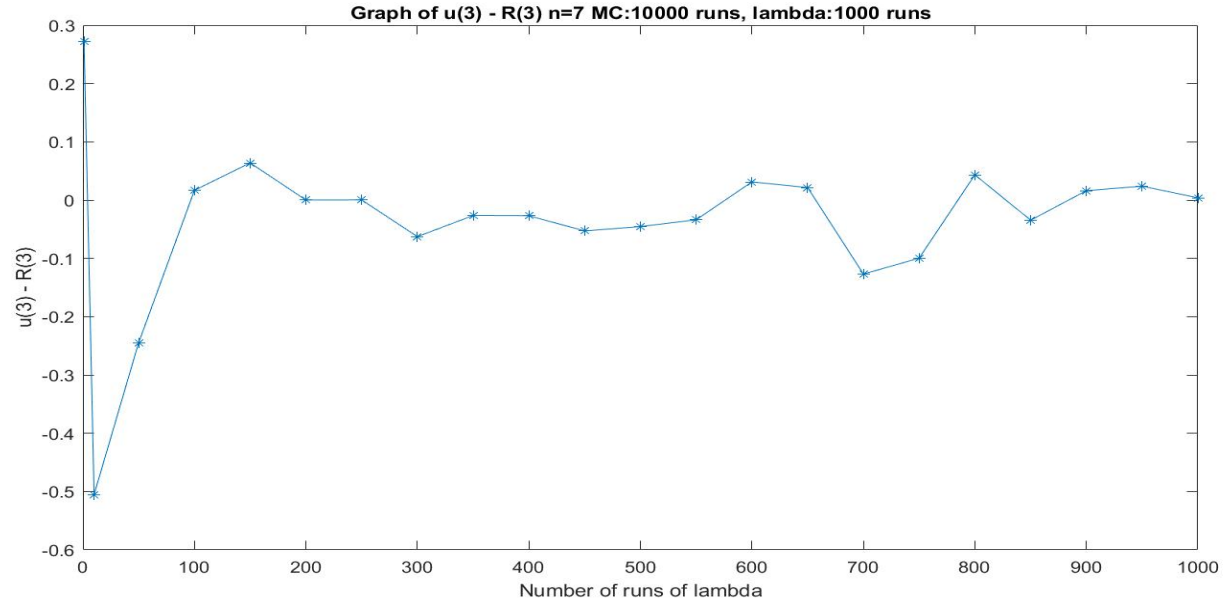

Figure 4.6: Graph of u(3) - R(3) n=7 MC:10000 l:1000

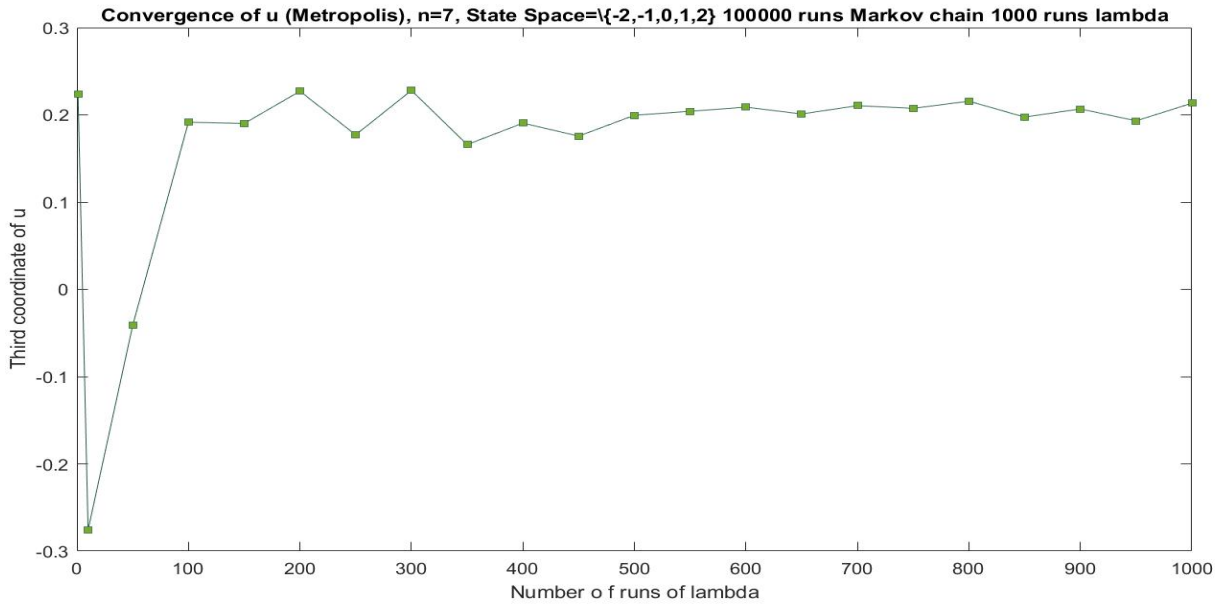

Figure 4.7: Convergence of u (MCMC), n=7,{-2,-1,0,1,2} MC:100000 l:1000

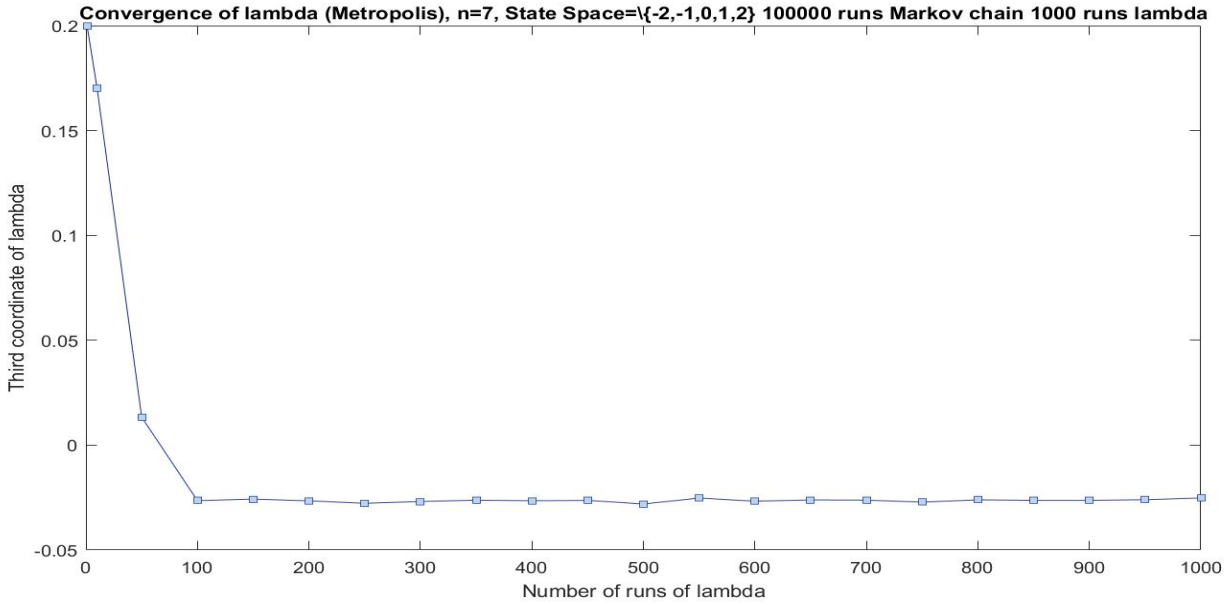

Figure 4.8: Convergence of λ (MCMC), n=7, {-2,-1,0,1,2} MC:100000 l:1000

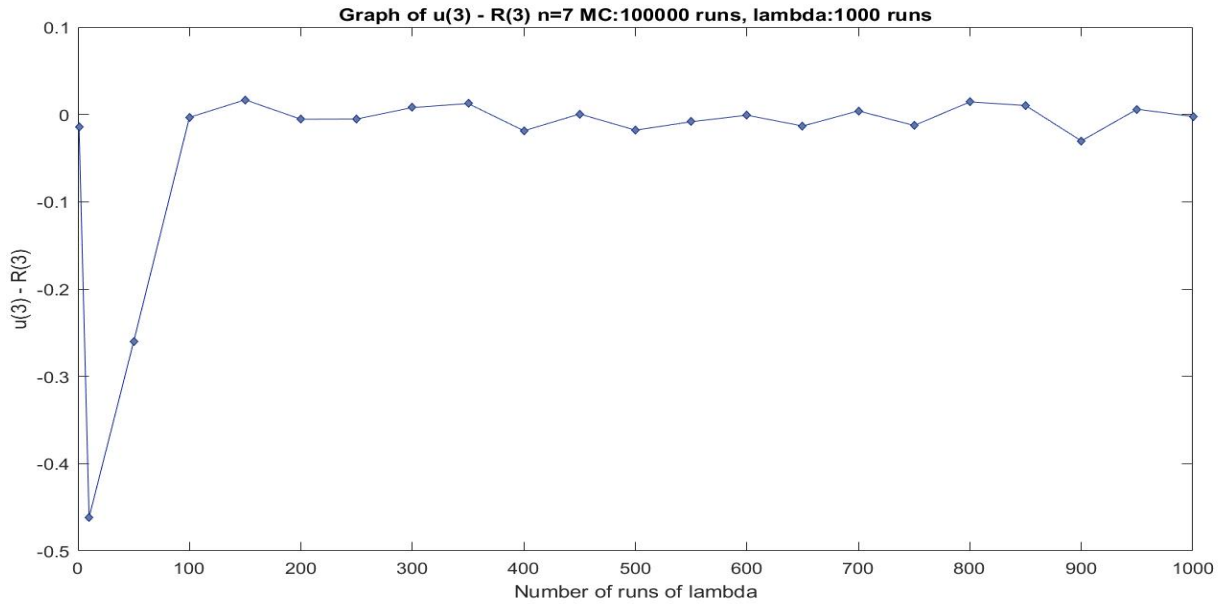

Figure 4.9: Graph of u(3) - R(3) n=7 MC:100000 l:1000

We also looked at an adaptive Markov chain Monte Carlo algorithm where the Markov chain Monte Carlo loop is allowed to run only for 1000 iterations. Clearly, for a state space so large, this does not give enough time for the convergence of the Markov chain. We run the λ-convergence (outer loop) for 10000 iterations.
Since the Markov chain is not allowed to converge, the errors are larger. However, the convergence is much faster. This is illustrated below for n=7.

Table comparing Timelapsed and Error for different n, same runs			
		Timelapsed	**Error = ||R-u||**
n=5	**Deterministic**	47.2104	1.06E-15
	MC : 10000; I : 1000	59.9652	0.1104
	MC : 100000; I : 1000	580.3025	0.0499
	MC : 1000; I : 10000	62.5889	0.4764
n=6	**Deterministic**	248.6124	1.11E-15
	MC : 10000; I : 1000	59.1452	0.2365
	MC : 100000; I : 1000	599.3587	0.0574
	MC : 1000; I : 10000	62.7262	0.425
n=7	**Deterministic**	1224.1	7.70E-16
	MC : 10000; I : 1000	61.8768	0.2013
	MC : 100000; I : 1000	619.6289	0.0903
	MC : 1000; I : 10000	62.6716	0.7586

Figure 4.10: Comparing Timelapsed and error for different n (n = 5,6,7) for same runs

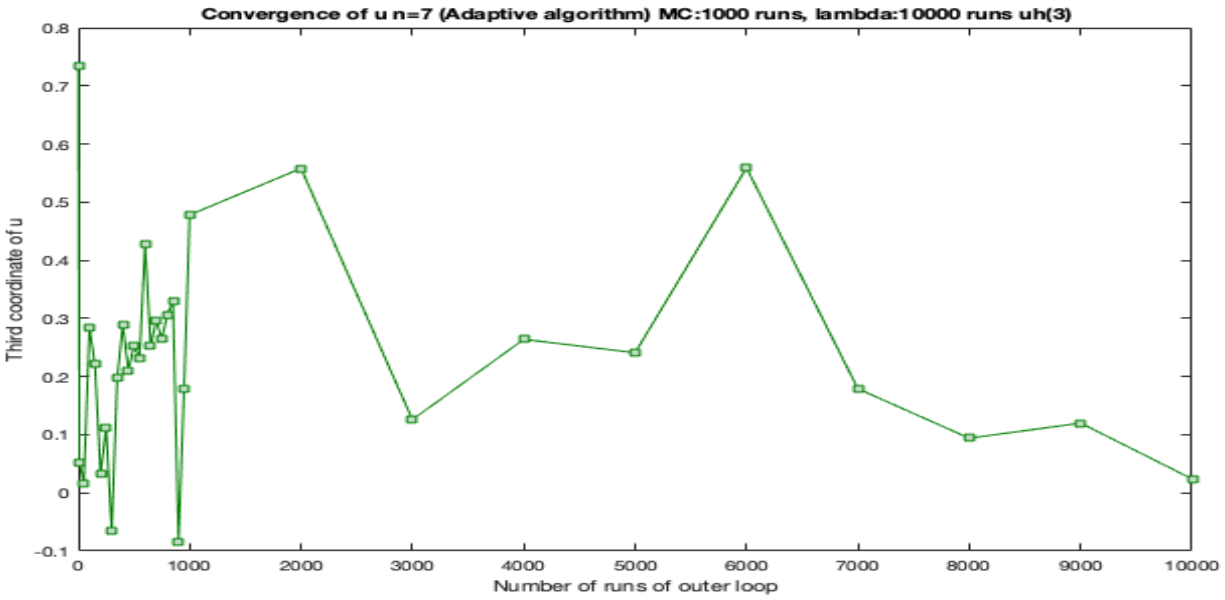

Figure 4.11: Convergence of u (Adaptive), n=7,{-2,-1,0,1,2} MC:1000 l:10000

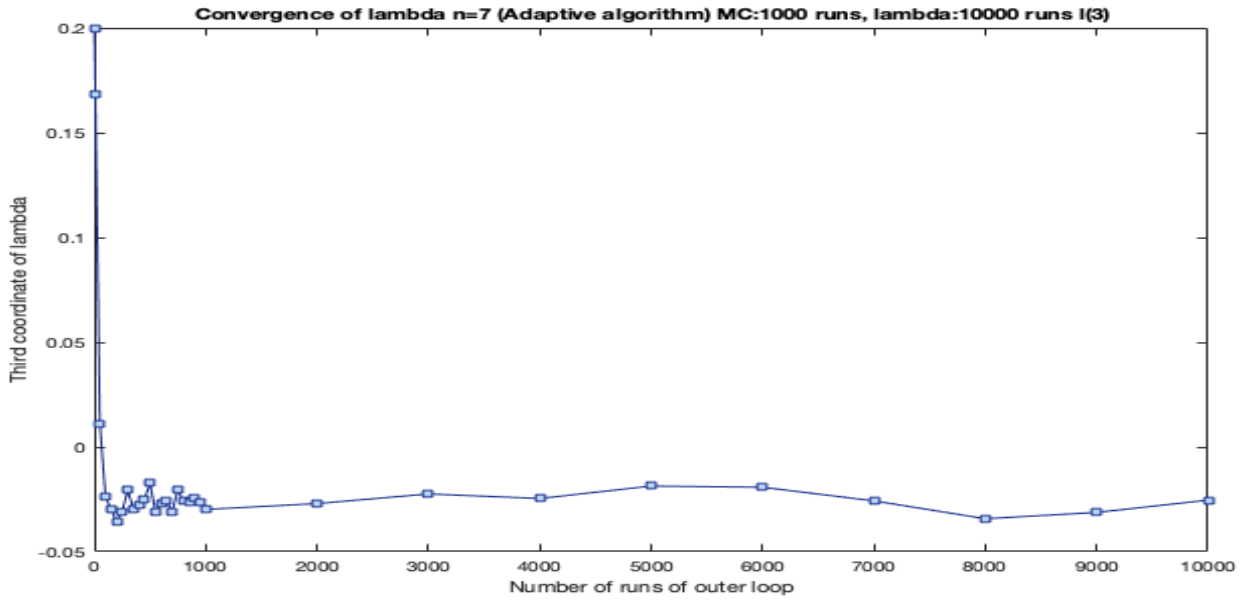

Figure 4.12: Convergence of λ (Adaptive), n=7, {-2,-1,0,1,2} MC:1000 l:10000

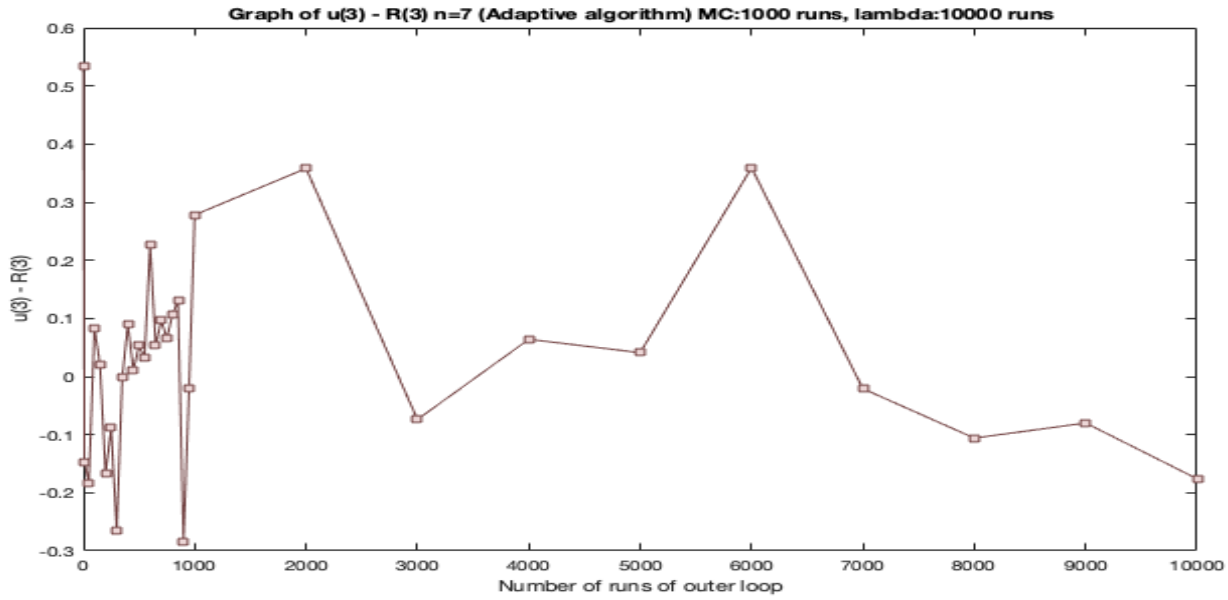

Figure 4.13: Graph of u(3) - R(3) (Adaptive), n=7, {-2,-1,0,1,2} MC:1000 l:10000

Clearly, from the above graphs, we can see that we do not get good enough convergence but looking at the table, *Figure 4.10* we can see that the time taken has considerably reduced.

Note that while we assume that the covariance matrix is realizable, the lagrange multiplier method works even if we are not given all the entries of the matrix. The simulations done here assume that we have c_{ij}'s for all i, j's. In other words, the interaction graphs of this n-random variables σ^i's is assumed to be complete.

Similar simulations can be carried out for the case when the underlying graph is not complete.

Bibliography

[GORJSR01] Gareth O. Roberts and Jeffrey S. Rosenthal, *General state space Markov chains and MCMC algorithms*, Probability Surveys, Vol. 1 20–71, 2004

[PDPMPNS13] Paolo Dai Pra, Michele Pavon and Neeraja Sahasrabudhe, *A Maximum Entropy Approach to the Realizability of Spin Correlation Matrices*, Entropy, 15, 2448-2463, 2013

[YFAJSR05] Yves F. Atchade and Jeffrey S. Rosenthal, *On adaptive Markov chain Monte Carlo algorithms*,Bernoulli, **11** , no.5, 815-828, 2005.

[AC02] Arnab Chakraborty, *Markov Chain Monte Carlo 1.Examples*, Resonance, 25-34, March 2002.

[CD11] Lecturer: Constantinos Daskalakis, Scribe: Alessandro Chiesa and Zeyuan Zhu; *6.896 Probability and Computation : Lecture 3*, 3-1 - 3-6, February 9, 2011.
Link : http://people.csail.mit.edu/costis/6896sp11/lec3s.pdf

[NM02] Neal Madras, *Fields Institute Monographs Lectures on Monte Carlo Methods*, American Mathematical Society, January 1, 2002.

[IY12] Ilker Yildirim, *Bayesian Inference: Gibbs Sampling*, August 2012.
Link: http://www.mit.edu/ ilkery/papers/GibbsSampling.pdf

[KBAPN78] K. B. Athreya and P. Ney, *A new approach to the limit theory of recurrent Markov chains*,Trans.Amer.Math.Soc. **245**, 493-501, 1978.

[SPMRLT93] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer-Verlang, London, 1993.
Link: http://decision.csl.uiuc.edu/ meyn/pages/TOC.html.

[EN78] E. Nummelin, *Uniform and ratio limit theorems for Markov renewal and semi-regenerative processes on a general state space*, Ann. Inst. Henri Poincare Series B **14** , 119-143, 1978.

[EN84] E. Nummelin, *General Irreducible Markov chains and non-negative operators*, Cambridge University Press, 1984.