On finite state ergodic Markov chains

Report of the project submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

Subin Pulari (B140100CS)

Under the guidance of

Dr. Murali Krishnan K

Department of Computer Science and Engineering

National Institute of Technology Calicut

Calicut, Kerala, India – 673 601

May 2018

Acknowledgements

I sincerely thank Dr.Murali Krishnan K for his guidance and supervision throughout the course of this project. I am indebted to Murali sir for introducing me to the world of theoretical computer science. I admire Dr.Murali Krishnan's constant efforts in developing simplified yet powerful pedagogical tools in various domains of computer science. This study is the final product of his original aim of presenting certain important results from the theory of Markov chains in a simplified form accessible to typical computer science graduates without compromising the mathematical rigour.
I express my gratitude towards Dr.Sanjay PK, Assistant professor, Mathematics Department, NIT Calicut for reviewing several important arguments presented in this study and also for the valuable technical support he provided throughout the duration of this project. Some of the simplifications we achieved in this study are contributions by Sanjay sir.
Finally, I would like to thank Dr.Vinod Pathari for his constructive advices and criticism during the project evaluation sessions.

Subin Pulari

May, 2018

National Institute of Technology Calicut

DECLARATION

"I hereby declare that this submission is my own work, and that to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text".

Place:

Date:

Signature:

Name: Subin Pulari

Roll No.: B140100CS

CERTIFICATE

This is to certify that the project report entitled: **On finite state ergodic Markov chains** *submitted by* **Subin Pulari (B140100CS)** *to National Institute of Technology Calicut towards partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide record of the work carried out by him/her under my supervision and guidance.*

Name and Signature of Guide

Signature of Head of the Department

Place:

Date:

Office Seal

Abstract

Finite state discrete time Markov chains are encountered in various contexts in computer science and communication theory. The objective of this project is to derive an elementary proof for the ergodic theorem for Markov chains that does not depend on any non-trivial results from Lebesgue's theory of integration.

A rigorous development of the probability space that models an infinite Markov chain is given by the Kolmogorov extension theorem (KET). We give a simple proof for Kolmogorov extension theorem when the underlying spaces are finite and discrete using only standard topological facts and basic measure theory. We employ the Kolmogorov extension theorem for finite discrete spaces to formulate the probability space underlying finite state Markov chains.

This is followed by a study of Markov chains using Markov shift transformations and standard apparatus from Ergodic theory. We give an elementary proof for Birkhoff's point-wise ergodic theorem for simple functions. The ergodic theorem for Markov chains follows directly from the point-wise ergodic theorem for simple functions.

Contents

List of Figures

Chapter 1

Introduction

The objective of the thesis is to develop the theory of finite state discrete time Markov chains including the ergodic theorem for Markov chains. In this introductory chapter, we give an overview of the theory that we develop in the rest of this document and also try to motivate the role of this theory, especially in the context of theoretical computer science.

Modelling of sources of information in Data Compression is among the many situations where one require probability spaces where each individual element in the sample space is an *infinite sequence* of symbols.

Let us consider a typical situation. Consider an information source emitting symbols from a finite alphabet $Q = \{0, 1, 2, \dots q-1\}$ in discrete time. The sample space in this case is the set of all infinite sequences of alphabets of the source. For instance, $(0, 1, 2, 0, 1, 2 \dots)$ would be an element in this sample space. This source can be modelled as a discrete stochastic process - essentially an infinite sequence of jointly distributed random variables $X_0, X_1, X_2 \dots$ where each $X_n, n \in \mathbb{N}$ take values from $\{0, 1, \dots q-1\}$. This provides a mathematical framework using which certain useful properties of the source, (like compressibility, see [8], [23], [28]) can be investigated.

Let $X_0, X_1, ...$ be a sequence of random variables defining a discrete stochastic process over an alphabet $Q = \{0, 1, \ldots q-1\}$ We can think of each random variable X_n as a function that takes as input an infinite sequence $(a_0, a_1 \ldots)$ which is an element from the product space $Q^{\mathbb{N}} = Q \times Q \times Q \times ...$, and yields as value the projection to the n^{th} component, a_n. Stated formally, we can model each X_n as a projection function from $Q^{\mathbb{N}}$ to Q. Intuitively, each element in the sample space is a possible sequence of symbols that the source can emit starting from time zero. The random variables captures the symbol emitted by the source at each instant $n \in \mathbb{N}$.

Discrete time finite state Markov chains are discrete stochastic processes with some additional (simplifying) mathematical properties, one of which is that at each instance of time the next state (or the symbol emitted by the source) is dependent only the present state and independent of whatever states had occured in the past. They are widely used for modelling many practical information sources - for instance, texts in the English language. (see [21],[17]). This is in particular due to the power of Markov chains to capture correlation between the occurence of source symbols. For example, the Markov chain in figure 1.1 having English alphabets as states can capture the fact that a 'u' is highly probable to occur after a 'q'. The circular nodes represent the states/symbols and the probability of transition between two states is indicated along an arrow connecting them. By considering all possible combinations of k symbols together we can build more *powerful* source models, as seen in figure 1.2 where we show a portion of Markov chain model considering all combinations of 4 symbols together. This model can express that a 'ques' will most probably be followed by a 'tion' and very less probably followed by a 'gggg'. As suggested by figure 1.2, a k^{th} order Markov source can be *reduced* to a first order Markov source with exponential (in k)

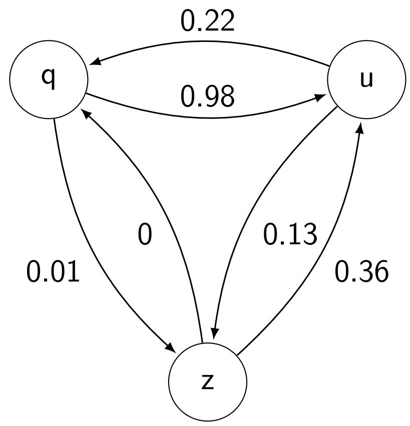

Figure 1.1: First order Markov source

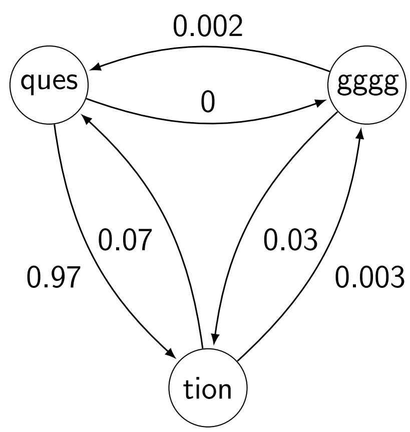

Figure 1.2: Fourth order Markov source

explosion in the state space. A nice account on modelling sources using Markov chains can be found in [21].

An initial concern is to define a probability function (on the product sample space). The existence of such a probability function is considered to be a black-box result in many texts (like [8]). More advanced texts take a traditional path through concepts in Measure Theory (as in [2],[1]) The latter approach culminates with proving Andrey Kolmogorov's Extension Theorem (also known as Kolmogorov Consistency Theorem) which proves the existence of an appropriate measure on the product sample space.

The sample space in the case of discrete time finite state Markov chains possess a simpler mathematical structure, of being a finite and discrete topological spaces. Exploiting this *simple* topological structure, we give a simple proof (in comparison with the standard method as seen in [2]) of the Kolmogorov extension theorem using standard facts from topology and measure Theory.

A Markov chain is called an ergodic Markov chain if it possess certain *homogeneous mixing* properties (see Chapters 2,7 and 8 for precise definitions). The ergodic theorem for Markov chains is an important result in the theory of Markov chains. The theorem can be proved using probabilistic arguments (as in [20] and [6]). However, the theorem can be obtained as a corollary of the Birkhoff's pointwise ergodic theorem, a standard result in ergodic theory. To obtain the ergodic theorem for finite state Markov chains, it suffice to argue Birkhoff's ergodic theorem in a restricted case, specifically when the function on which the theorem is applied is a *simple function* (see Chapter 9).

We initially obtain an elementary proof for Birkhoff's ergodic theorem for simple functions in Chapter 9. From the Birkhoff's ergodic theorem for simple functions, we easily obtain ergodic theorem for finite state Markov chains as a corollary in Chapter 10. We also demonstrate some applications of the Poincare recurrence lemma to the theory of Markov chains in Chapter 10.

1.1 Motivation

As we had noted in the previous section, Markov chains find many applications in theoretical computer science among which an important one is its use in the modelling of information sources. Ergodic processes exhibit certain *homogeneous mixing* properties, for example a Markov chain with a positive probability path between any two states defines an ergodic process (as we will show in Chapters 6 and 8).

The properties possessed by Ergodic processes often leads to optimality results in data compression. A notable one is the optimality of Lempel-Ziv 1978 compression algorithm when the underlying source is a binary stationary stochastic process (see [8]).

The ergodic theorem for Markov chains is a statement about long time behaviour of ergodic Markov chains. This result in particular, has applications in design and analysis fo Randomized algorithms (see [19]). A mathematically rigorous development of the theory of Markov chains including the ergodic theorem for Markov chains involve significant use of measure theoretic machinery, which may not be accessible to majors in communication theory and computer science.

Initially, we need to formulate the underlying probability space of finite state Markov chains. As we noted in the previous section, an essential ingredient in this process is the Kolmogorov extension theorem. However since the Markov chain has only a finite set of states, we can employ the compactness property from topology to obtain simplifications in the main argument of the standard proof for Kolmogorov extension theorem (see [2]).

The ergodic theorem for Markov chains is intuitively a generalisation of the strong law of large numbers (see [2]) to Markov processes. The ergodic theorem for Markov chains can be proved using various methods. In this thesis, we do this by venturing into ergodic theory, proving the Birkhoff's ergodic theorem and obtaining the ergodic theorem for Markov chains as a corollary of Birkhoff's ergodic theorem. We claim that this approach is more fruitful than certain traditional approaches, since we derive the results using a more general toolset of ergodic theory. This claim is rooted in our belief that arriving at certain results from a more abstract mathematical point of view could provide more insights into the results and also equip the reader with tools that are applicable for processes more general than finite state Markov chains.

Ergodic theory is the study of dynamical systems that vary over time under predetermined rules of change over time. The parts of ergodic theory that we are concerned about will deal with dynamical systems where these rules of change does not vary over time. A classic example is the system of gas particles in a container. The rules of change may specify the change of position of each particle in the next instant of time. Birkhoff's (pointwise) ergodic theorem is a seminal result in this field which was proved by G.D.Birkhoff in [5]. Simple functions are functions which take values from a finite subset of the range. We give an elementary proof for the Birkhoff's ergodic theorem for simple functions in Chapter 9. The proof uses only elementary facts from Lebesgue's theory of integration unlike most methods available in literature (like [24],[26] and [18]) which uses results like the Dominated convergence theorem and Monotone convergence theorem (see [10],[2]).

Due to the simplifications outlined above, we believe that our study can be used to introduce the theory of finite state Markov chains to graduates or advanced undergraduates in CS and communication theory in a 20 lecture course without assuming any previous knowledge in measure theory and topology. We summarize the expected prerequisities in the next section.

1.2 Prerequisities

We expect the reader to be familiar with basic notions in Real Analysis (including limits, sequence, series, convergence etc). All the necessary topics can be found in [22].

We also expect the reader has encountered topics in undergraduate probability and linear algebra. Good references for probability include [9] and the appendix on probability theory from [7]. The reader may acquire necessary linear alegbraic prerequisities from [3] or [15].

Necessary topics from topology and measure theory will be completely developed in the following chapters.

1.3 Outline of the report

In Chapter 2 we begin the development of the theory of Markov chains by developing finite run Markov chains. Here we motivate the need for the Kolmogorov extension theorem for finite discrete spaces which we will state and prove in Chapter 4. The necessary topological and measure theoretic prerequisities will be developed in Chapter 3. All the necessary results are stated in Chapter 3. But, some of the proofs are given in Appendices A and B.
In Chapter 5, we extend the theory of finite run Markov chains done in Chap-

ter 2 to infinite run Markov chains. We will also develop sufficient theory to state the ergodic theorem for Markov chains, which will proved later in Chapter 10.

In Chapter 6, we prove the Perron theorem for positive Markov chains and later prove the important convergence to stationary distribution result for positive Markov chains.

In Chapter 7, we develop basic Ergodic theory and also give a proof of the Poincare recurrence lemma. The results from Chapters 6 and 7 find application in Chapter 8 where use introduce the Markov shift transformations and ultimately prove that the shift transformation is ergodic if the underlying Markov chain is stationary.

We given an elementary proof of the Birkhoff's pointwise ergodic theorem after developing Lebesgue integration theory for simple functions in Chapter 9. The Birkhoff's ergodic theorem is applied to Markov chains in Chapter 10 to obtain the ergodic theorem for Markov chains. In this chapter we also give a simple application of Poincare recurrence lemma to Markov chains.

Appendices C to D contain certain results that were stated but left unproven in the following Chapters. The reader will be asked to refer to appendices at these points.

Chapter 2

Finite run Markov Chains and their consistency

In this chapter we first look into the process of setting up *appropriate* probability spaces for discrete stochastic process (of which Markov chains are a special case) and try to motivate the need for the Kolmogorov extension theorem in this context. In the first section we do this by considering an independent and identically distributed discrete stochastic process. In the later parts of chapter, we begin with the theory of Markov chains by developing finite run Markov chains. Extending this to infinite run Markov chains will be done in Chapter 5

2.1 Consistency and extension of finite probability spaces

Let us consider a *hypothetical machine* acting as a source of information that emits a symbol independently at random from the alphabet $Q = \{a, b, c\}$ [1] each second after being switched on. Let probability that the machine emits the symbols be $p_a = \frac{1}{5}$, $p_b = \frac{2}{5}$ and $p_c = \frac{2}{5}$.

We first consider finite time outputs of the machine. Specifically, let it be the case that the machine is switched off after 5 seconds from starting. A sample sequence emitted by the machine is shown in figure 2.1

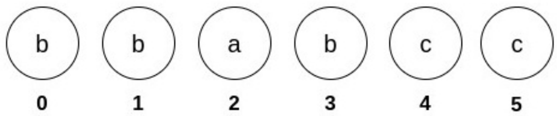

Figure 2.1: Finite sequence emitted by the machine

Let us consider the set of all sequences that can be emitted by the machine in this case, specifically $\Omega = \{(x_0, x_1, \ldots x_4) : x_i \in Q \ \forall i \in \{0, 1, 2, \ldots 4\}\} = Q^5$. Using the probabilities of the alphabets, taking the power set 2^Ω as the set of *events*, we have the usual probability function $P : 2^\Omega \mapsto [0, 1]$ such that for any element $\omega = (x_0, x_1, \ldots x_4) \in \Omega$ we have $P(\omega) = p_{x_0} p_{x_1} \ldots p_{x_4}$. For example, the probability that 3 a's occur in a sequence emitted by the machine is found to be ${}^5C_3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^2$.

Now, consider the case when the machine be switched on at time zero and emits symbols forever. A sample sequence emitted by the machine is shown in figure 2.2

[1] We use a, b and c instead of $Q = \{0, 1, 2\}$ for clarity

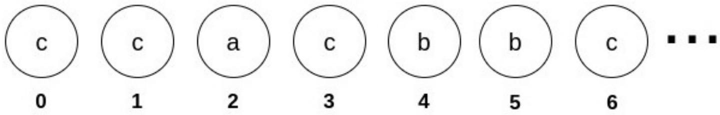

Figure 2.2: Infinite sequence emitted by the machine

Consider the set of all infinite sequences that can be emitted by the machine, let $\Omega = \{(x_0, x_1, x_2 \dots) : x_i \in Q \; \forall i \in \mathbb{N}\} = Q^{\mathbb{N}}$. Now let us try to construct a probability function P from set of events $\mathcal{E} = 2^{\Omega}$ to $[0, 1]$. Before moving further we consider some *desirable properties* of P. These are some of the possible questions we would like to answer using P:

- What is the probability that an infinite sequence begins with 'a' ?
- What is the probability that at times 10, 100 and 1000 the machine does not emit a 'c'?
- What is the probability that in the range of time from 50 to 10000, the machine emits b's only?

Intuition suggests that the answer to these questions should be the same as what we should obtain on considering the case when the machine emits a symbol *exactly* at the finite set of times considered in the above questions and remains switched off during other times. For example, consider the first question, the set of sequences we are concerned about is *intuitively* $\frac{1}{5}^{th}$ of the sample space Ω. This is exactly the probability that the machine emits an 'a' at time 0 when it is switched off at time $= 1$. We observe that P should be an *extension* of the usual probability functions we obtain on considering a finite set of times.

Let $[a, b]$ where $a < b$ denote $\{n \in \mathbb{N} : a \leq n \leq b\}$. Now, let us consider the sample spaces $\Omega_{[0,5]} = \{(x_0, x_1, \dots x_5) : x_i \in Q \; \forall i \in \{0, 1, \dots 5\}\}$ and

$\Omega_{[0,10]} = \{(x_0, x_1, \dots x_{10}) : x_i \in Q \; \forall i \in \{0, 1, \dots 10\}\}$. i.e, we consider the resulting sample spaces when the machine switched on at time 0 is switched off after times 5 and 10 respectively. We have the usual notion of probabilities over these sample spaces. Let the probability functions be $P_{[0,5]}$ and $P_{[0,10]}$ respectively.

Now consider the probability of machine emitting a 'b' at times 2,3 and 4 respectively. Both $P_{[0,5]}$ and $P_{[0,10]}$ are capable of providing the probability we need and their answers are *consistent* with each other (it can be verified that the answer is $(\frac{2}{5})^3$ in both situations).

Before making a more general statement, the following notation is defined. Let π denote the projection function from $\Omega_{[0,10]}$ to $\Omega_{[0,5]}$. That when π is applied on $(x_0, x_1, \dots x_{10}) \in \Omega_{[0,10]}$, the function outputs its coordinates from time 0 to 5, $(x_0, x_1, \dots x_5) \in \Omega_{[0,5]}$. The inverse π^{-1} when applied to any set of elements A in $\Omega_{[0,5]}$, outputs all possible elements in $\Omega_{[0,10]}$ that can be obtained by extending any $a \in A$ by inserting any possible value in coordinates 6 to 10. For example, $\pi^{-1}(\{(a, a, b, a, c, a)\}) = \{(a, a, b, a, c, a, x_6, x_7 \dots x_{10}) : x_i \in Q \; \forall i \in \{6, 7, \dots 10\}\}$

Generally, observe that for any event $A \subseteq \Omega_{[0,5]}$, it is the case that $P_{[0,10]}(\pi^{-1}(A)) = P_{[0,5]}(A)$. i.e, probability of any event $A \subseteq \Omega_{[0,5]}$ can be *consistently* answered by $P_{[0,10]}$ on being applied to the inverse projection of set A to $\Omega_{[0,10]}$.

The above argument in case of $[0, 5]$ and $[0, 10]$ can be extended to any $X, Y \subseteq \mathbb{N}$ such that $X \subseteq Y$. We will say in this case that, the set of probability functions we obtain on considering finite subsets of $\mathbb{N}$ are a *consistent family* of probability functions.

Since we have seen the notions of *extension* and *consistency*, we are in a position to see the relevance of Kolmogorov Extension Theorem. Let $\mathbb{F}$

denote the set of all probability functions that can be constructed over finite set of times (like $P_{[0,5]}$ and $P_{[0,10]}$). The existance of a probability function P on $\Omega = \{(x_0, x_1, x_2 \dots) : x_i \in Q \ \forall i \in \mathbb{N}\} = Q^{\mathbb{N}}$ which is an *extension* of every probability function in $\mathbb{F}$ is not obvious at this stage. Its existance should necessarily be argued before moving further into the study of *random processes.*

This is where Kolmogorov Extension Theorem comes into picture. Intuitively speaking, in the case considered above, when the usual probability functions over finite product sample spaces $\mathbb{F}$ is a *consistent* family of probability functions, the existance of P on $\Omega = \{(x_0, x_1, x_2 \dots) : x_i \in Q \ \forall i \in \mathbb{N}\} = Q^{\mathbb{N}}$ which is an *extension* of any probability function in $\mathbb{F}$ is guarenteed by Kolmogorov Extension Theorem.

In the next section, we begin with the theory of Markov chains and we try to further motivate the use of Kolmogorov extension theorem in this context.

2.2 Finite run Markov chains

Markov chains are discrete stochastic processes (infinite sequence of random variables $X_0, X_1, \dots$ as defined in the Introduction) with certain dependency among the random variables. In this chapter we study *finite run* Markov chains. Through this we aim to motivate the relevance of the Kolmogorov Extension Theorem in the context of Markov chains by investigating 'finite run' versions of general Markov chains. We also encounter certain *sufficient* conditions for consistency of probability functions over finite indices.

Markov chains can be seen as a random walk (a finite walk in case of *finite run* Markov chains as we define later in this chapter) over a state transition system like the *random traffic light* given in the figure below.

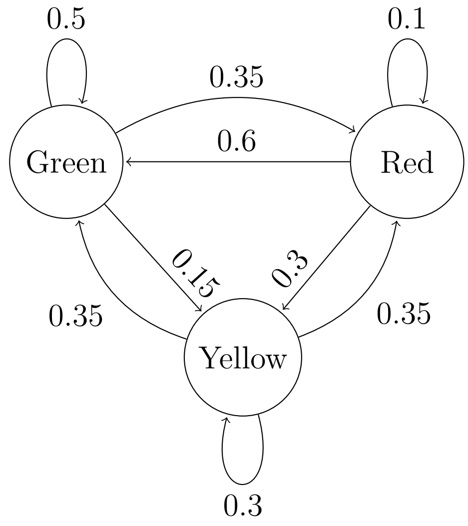

Figure 2.3: Random traffic light Markov chain

Let us examine two definitive properties of general Markov chains,

- When the process is in a state (say *green* as in the example above), the next state the system can be in is *totally* dependent on the current state and independent on the earlier states the system was in. (this is referred to as the **Markov property** of Markov chains)

- The probability of transition between any two specific states does not change with time. For example, when the traffic light Markov chain is in state *green* at any point of time, the probability of making a transition to the *yellow* state is always 0.15 (this is referred to as the **time invariance** property)

We will make use of matrices to capture the state transition model with its transition probabilities.

Definition 2.2.1. Let $Q = \{1, 2, 3, \ldots q\}$ be a finite set. A matrix $M = [p_{ij}]_{i,j \in Q}$ is defined to be a **stochastic matrix** if $0 \leq p_{ij} \leq 1$ and $\sum_{i \in Q} p_{ij} = 1$ for all $i \in Q$

We capture the essential 'ingredients' for defining a Markov chain as a *Markov system*

Definition 2.2.2 (Markov system). Let $Q = \{1, 2, 3, \dots q\}$ be a finite set. A matrix $M = [p_{i,j}]_{i,j \in Q}$ be a stochastic matrix. let $\mu_0 = [p_i]_{i \in Q}$ be a $|Q| \times 1$ column vector such that $0 \leq p_i \leq 1$ and $\sum\limits_{i \in Q} p_i = 1$. We define the triple (Q, μ, M) to be a Markov system. Q is the set of *states*, M is *transition matrix* and μ is referred to as the *initial distribution* of the Markov system (Q, μ, M).

The $(i, j)^{th}$ entry of the stochastic matrix M represent the probability that the Markov system makes a transition from state i to state j.

We will consider the case when the Markov system 'starts' at *time* 0 and 'runs' upto some $n \in \mathbb{N}$. In this chapter $[0, n]$ for any $n \in \mathbb{N}$ will represent the set $\{0, 1, \dots n\}$. Also. Q will represent a finite set $\{1, 2, 3, \dots q\}$ for some $q \in \mathbb{N}$ unless specified otherwise. Now we define a *n-Run* Markov chain.,

Definition 2.2.3. Given a Markov system (Q, μ, M), consider the sample space $\Omega_{[0,n-1]} = Q^n$ and the set of events $\mathcal{E}_{[0,n-1]} = 2^{\Omega_{[0,n-1]}}$. Let $P_{[0,n-1]}$ be any probability function on $\mathcal{E}_{[0,n-1]}$ (we specify necessary conditions on $P_{[0,n-1]}$ shortly). Let $X_0, X_1, \dots X_{n-1}$ be a sequence of random variables such that $X_i = \pi_i$ for each $i \in \{0, 1, \dots n-1\}$ where π_i's are the projection functions from $\Omega_{[0,n-1]}$ to the component spaces. $X_0, X_1, \dots X_{n-1}$ is defined to be an **n-Run Markov chain** if the following holds:

1. $P_{[0,n-1]}(X_0 = a) = p_a$ for any $a \in Q$

2. $P_{[0,n-1]}(X_i = a_i | X_{i-1} = a_{i-1}, \dots X_0 = a_0) = P_{[0,n-1]}(X_i = a_i | X_{i-1} = a_{i-1})$ for all $i \in \{1, 2, \dots n-1\}$ when each $a_i \in Q$

3. $P_{[0,n-1]}(X_i = b | X_{i-1} = a) = p_{ba}$ for all $i \in \{1, 2 \dots n-1\}$ and for all $a, b \in Q$

$(\Omega_{[0,n-1]}, \mathcal{E}_{[0,n-1]}, P_{[0,n-1]})$ is called the **n-Run probability space** for the given Markov system (Q, μ, M)

It is to be noted that the *Markov property* and *time invariance* hold *by definition* for n-run Markov chains.
The following is a direct consequence of the definition of n-Run Markov chains,

$$
\begin{aligned}
& P_{[0,n-1]}(\{(a_0, a_1, \ldots a_{n-1})\}) \\
& = P_{[0,n-1]}(\{X_0 = a_0\}) P_{[0,n-1]}(\{X_1 = a_1 | X_0 = a_0\}) \ldots \\
& \ldots P_{[0,n-1]}(\{X_n = a_n | X_0 = a_0, \ldots X_{n-1} = a_{n-1}\}) \\
& = P_{[0,n-1]}(\{X_0 = a_0\}) P_{[0,n-1]}(\{X_1 = a_1 | X_0 = a_0\}) \ldots \\
& \ldots P_{[0,n-1]}(\{X_n = a_n | X_{n-1} = a_{n-1}\}) \text{ (using Markov prop.)} \\
& = p_{a_0} p_{a_1 a_0} \ldots p_{a_{n-1} a_{n-2}}
\end{aligned}
$$

for each $(a_0, a_1, \ldots a_{n-1}) \in \Omega_{[0,n-1]}$.
We obtained that,

$$
P_{[0,n-1]}(\{(a_0, a_1, \ldots a_{n-1})\}) = p_{a_0} p_{a_1 a_0} \ldots p_{a_{n-1} a_{n-2}} \tag{2.1}
$$

for each $(a_0, a_1, \ldots a_{n-1}) \in \Omega_{[0,n-1]}$.
Conversely, given a Markov system (Q, μ, M), a probability function $P_{[0,n-1]}$ with the desirable properties specified in the definition can be obtained by defining $P_{[0,n-1]}$ as in equation 2.1. We verify the three conditions in definition 2.2.3m

1. For any $a \in Q$, $P_{[0.n-1]}(X_0 = a) = P_{[0.n-1]}\{(a)\} = p_a$.

2. If $i \in \{1, 2, \ldots n-1\}$ and if $a_i \in Q$ for all i,

$$
\begin{aligned}
&P_{[0,n-1]}(X_i = a_i | X_{i-1} = a_{i-1}, \ldots X_0 = a_0) \\
&= \frac{P_{[0,n-1]}(X_i = a_i, X_{i-1} = a_{i-1}, \ldots X_0 = a_0)}{P_{[0,n-1]}(X_{i-1} = a_i, X_{i-2} = a_{i-1}, \ldots X_0 = a_0)} \\
&= \frac{p_{a_0} p_{a_1 a_0} \cdots p_{a_i a_{i-1}}}{p_{a_0} p_{a_1 a_0} \cdots p_{a_{i-1} a_{i-2}}} \\
&= p_{a_i a_{i-1}}
\end{aligned}
$$

3. For all $i \in \{1, 2 \ldots n-1\}$ and for all $a, b \in Q$,

$$
\begin{aligned}
P_{[0,n-1]}(X_i = b | X_{i-1} = a) &= \frac{P_{[0,n-1]}(X_i = b, X_{i-1} = a)}{P_{[0,n-1]}(X_{i-1} = a)} \\
&= \frac{\sum\limits_{q_0, q_1, \ldots q_{i-2} \in Q} P_{[0,n-1]}\{(q_0, q_1, \ldots q_{i-2}, a, b)\}}{\sum\limits_{q_0, q_1, \ldots q_{i-2} \in Q} P_{[0,n-1]}\{(q_0, q_1, \ldots q_{i-2}, a)\}} \\
&= \frac{\sum\limits_{q_0, q_1, \ldots q_{i-2} \in Q} p_{a_0} p_{q_1 q_0}, \cdots p_{a q_{i-2}} p_{ba}}{\sum\limits_{q_0, q_1, \ldots q_{i-2} \in Q} p_{a_0} p_{q_1 q_0}, \cdots p_{a q_{i-2}}} \\
&= \frac{p_{ba} \times \sum\limits_{q_0, q_1, \ldots q_{i-2} \in Q} p_{a_0} p_{q_1 q_0}, \cdots p_{a q_{i-2}}}{\sum\limits_{q_0, q_1, \ldots q_{i-2} \in Q} p_{a_0} p_{q_1 q_0}, \cdots p_{a q_{i-2}}} \\
&= p_{ba}
\end{aligned}
$$

Thus, defining the probability function $P_{[0,n-1]}$ as in equation 2.1 is an equivalent way of defining n-Run Markov chains.

At this point we obtain the following fact.

Lemma 2.2.1. *Let $X_0, X_1, X_2, \ldots$ be a n-run Markov chain corresponding to a Markov system (Q, μ_0, M) ,the vector of probabilities $[P_{[0,n-1]}(X_i = b)]_{b \in Q} = M^i \mu_0$ for any $i \in [0, n-1]$*

Proof. The proof is by induction on i

The result is trivially true when $i = 0$.

Assume the result of $i = k - 1$. We will prove the result for $i = k$ (such that $0 \leq k \leq n - 1$)

For any $0 \leq i \leq n - 1$ and $a \in Q$, $[M^i \mu_0]_a$ denote the entry in a^{th} of $M^i \mu_0$. Using conditional probability, we get

$$\begin{aligned} P_{[0,n-1]}(X_k = b) &= \sum_{a \in Q} P_{[0,n-1]}(X_k = b | X_{k-1} = a) P_{[0,n-1]}(X_{k-1} = a) \\ &= \sum_{a \in Q} P_{[0,n-1]}(X_k = b | X_{k-1} = a) \left[M^{k-1} \mu_0 \right]_a \ (1) \\ &= \sum_{a \in Q} p_{ba} \left[M^{k-1} \mu_0 \right]_a \\ &= \left[M^k \mu_0 \right]_a \end{aligned}$$

We have used the induction hypothesis in obtaining equality in (1)

The above shows that $[P_{[0,n-1]}(X_k = b)]_{b \in Q} = M^k \mu_0$ and hence the result follows. $\square$

Thus, finding the set of probabilities that the Markov chain is in particular states at any time i reduces to multiplying the transition matrix M, i times with the initial distribution vector μ_0.

We extend the notion of projection functions before moving further. Let $\{B_i\}_{i \in I}$ (I arbitrary) be a any class of sets and $\prod_{i \in I} B_i$ be the product of B_i's. Let $F \subseteq I$, $\pi_F : \prod_{i \in I} B_i \mapsto \prod_{f \in F} B_f$ will denote the projection function to the product over the component spaces indexed with F. More precisely, $\pi_F((a_i)_{i \in I}) = ((a_f)_{f \in F})$ for all $(a_i)_{i \in I} \in \prod_{i \in I} B_i$. When F is a singleton,i.e, $F = \{k\}$, then we use π_k instead of $\pi_{\{k\}}$ (hence, the notation is consistent with our earlier use of π).

We now prove the main result of this section,

Lemma 2.2.2. *Given a Markov system (Q, μ_0, M), the n-Run probability spaces are* ***consistent*** *with each other. i.e, Suppose $m, n \in \mathbb{N}$ such that $n < m$. Then for any $A \in \mathcal{E}_{[0,n-1]}$, $P_{[0,m-1]}(\pi^{-1}_{[0,n-1]} A) = P_{[0,n-1]}(A)$*

Proof. $P_{[0,m-1]}(\pi^{-1}_{[0,n-1]}A)$

$$= \sum_{(a_0,a_1,\dots a_{n-1})\in A} \sum_{a_{n+1}\in Q} \sum_{a_{n+2}\in Q} \cdots \sum_{a_{m-1}\in Q} P_{[0,m-1]}\left\{(a_0,a_1,\dots a_{m-1})\right\}$$

$$= \sum_{(a_0,a_1,\dots a_{n-1})\in A} \sum_{a_{n+1}\in Q} \sum_{a_{n+2}\in Q} \cdots \sum_{a_{m-1}\in Q} p_{a_0}p_{a_1a_0}\cdots p_{a_{m-1}a_{m-2}}$$

$$= \sum_{(a_0,a_1,\dots a_{n-1})\in A} \sum_{a_{n+1}\in Q} \sum_{a_{n+2}\in Q} \cdots \sum_{a_{m-2}\in Q} p_{a_0}p_{a_1a_0}\cdots p_{a_{m-2}a_{m-3}} \sum_{a_{m-1}\in Q} p_{a_{m-1}a_{m-2}}$$

By using the property of stochastic matrices, the above sum reduces to,

$$\sum_{(a_0,a_1,\dots a_{n-1})\in A} \sum_{a_{n+1}\in Q} \sum_{a_{n+2}\in Q} \cdots \sum_{a_{m-2}\in Q} p_{a_0}p_{a_1a_0}\cdots p_{a_{m-2}a_{m-3}}$$

In successive steps, we get,

$$P_{[0,m-1]}(\pi^{-1}_{[0,n-1]}A) = \sum_{(a_0,a_1,\dots a_{n-1})\in A} p_{a_0}p_{a_0,a_1}\cdots p_{a_{n-2},a_{n-1}} = P_{[0,n-1]}(A)$$

□

We further extend the notion of projection functions to equip ourselves with notation that can capture more general projections between spaces. Let $\{B_i\}_{i\in I}$ (I arbitrary) be a any class of sets and $\prod_{i\in I} B_i$ be the product of B_i's. Let $F, G \subseteq I$ such that $F \subseteq G$, $\pi_{G\to F} : \prod_{g\in G} B_g \mapsto \prod_{f\in F} B_f$ will denote the projection function from product over the component spaces indexed with G to the product over the component spaces indexed with F. More precisely, $\pi_{G\to F}((a_i)_{g\in G}) = ((a_f)_{f\in F})$ for all $(a_g)_{g\in G} \in \prod_{g\in G} B_g$. The π_F functions we introduced before the definiton of n-Run Markov chains can be expressed in this notation as $\pi_{I\to F}$.

n-Run Markov chains intuitively are 'runs' of the general Markov chain from time 0 to $n-1$. Now we generalize this to Finite run Markov chains which are 'runs' of general Markov chains on any finite subset of time. It turns out to be the case that consistency of n-Run Markov chains is sufficient to have a well-defined notion of Finite run Markov chains. We define a finite run Markov chain below,

Definition 2.2.4. Given $F \subseteq \mathbb{N}$, $F = \{f_0, f_1, f_2, \dots f_{|F|-1}\}$ finite (without loss of generality, let $f_0 < f_1 < f_2, \dots < f_{|F|-1}$). Let (Q, μ, M) be a Markov system. Consider the sample space $\Omega_F = Q^{|F|}$ and the set of events $\mathcal{E}_F = 2^{\Omega_F}$. Let $X_{f_0}, X_{f_1}, \dots X_{f_{|F|-1}}$ be a sequence of random variables such that $X_{f_i} = \pi_{f_i}$ for each $i \in \{0, 1, \dots |F|-1\}$.
Let $n \in \mathbb{N}$ be any natural number such that $F \subseteq [0, n-1]$ (such an n exist since F is finite). Define the probability function on the space, $P_F(A) = P_{[0,n-1]}(\pi^{-1}_{[0,n-1]\to F}(A))$ for all $A \in \mathcal{E}_F$. With the above definition of P_F, $X_{f_0}, X_{f_1}, \dots X_{f_{|F|-1}}$ is defined to be a **Finite run Markov chain**.
$(\Omega_F, \mathcal{E}_F, P_F)$ is said to be a **Finite run probability space**

Now we verify that the probability function is well defined,

Lemma 2.2.3. *P_F is well defined*

Proof. Let $m, n \in \mathbb{N}$, $m < n$ be such that $F \subseteq [0, n-1]$ and $F \subseteq [0, m-1]$.
For any $A \in \mathcal{E}_F$, $P_F(A) = P_{[0,n-1]}(\pi^{-1}_{[0,n-1]\to F}(A))$
$= P_{[0,n-1]}\left(\pi^{-1}_{[0,n-1]\to[0,m-1]}\left(\pi^{-1}_{[0,m-1]\to F}(A)\right)\right)$
$= P_{[0,m-1]}\left(\pi^{-1}_{[0,m-1]\to F}(A)\right)$ (due to the lemma 2.2.2)

□

For any $n \in N$, when $F = [0, n-1]$, it is easy to see that $(\Omega_F, \mathcal{E}_F, P_F) = (\Omega_{[0,n-1]}, \mathcal{E}_{[0,n-1]}, P_{[0,n-1]})$ (the n-run probability space which we defined earlier). And thus the defintions of P_F's are consistent with the earlier definition of $P_{[0,n-1]}$'s. As we indicated, the Finite run probability spaces happens to be consistent, which we prove below,

Lemma 2.2.4. *Finite run probability spaces are **consistent** with each other. i.e, Suppose $F \subseteq G \subseteq \mathbb{N}$. Then for any $A \in \mathcal{E}_F$, $P_G\left(\pi^{-1}_{G\to F}(A)\right) = P_F(A)$*

Proof. Let $n \in \mathbb{N}$ be any natural number such that $F, G \subseteq [0, n-1]$. Now,

$$\begin{aligned} P_G\left(\pi^{-1}_{G \to F}(A)\right) &= P_{[0,n-1]}\left(\pi^{-1}_{[0,n-1]\to G}\left(\pi^{-1}_{G\to F}(A)\right)\right) \\ &= P_{[0,n-1]}\left(\pi^{-1}_{[0,n-1]\to F}(A)\right) \\ &= P_F(A) \end{aligned}$$

□

We saw that the consistency of n-Run probability spaces lead to a defintion of finite run probability spaces, such that they themselves turn out to be consistent. Given any discrete stochastic process $X_0, X_1, X_2 \ldots$, having consistency of the natural probability spaces on $[0, n-1]$ runs of the process (i.e, $X_0, X_1, \ldots X_{n-1}$), the above arguments can be extended to define and prove the consistency of probability spaces on any finite $F \subseteq \mathbb{N}$.

Now we consider *running* the Markov system from time 0 to infinity. We thus obtain random variables sequence $X_0, X_1, X_2 \ldots$. The discrete stochastic process we thus obtain is generally referred to as a Markov chain (see [8]). The sample space Ω in this case is the set of all infinite sequences of symbols from set of states Q.

The following are natural requirements on the infinite product probability space that we aim to define,

- We require that the inverse projections of events in the finite run probability spaces are events in the infinite product probability space. If $\mathcal{E}$ is the set of events in the infinite product space, then given finite $F \subseteq \mathbb{N}$, $\pi_F^{-1}(A) \in \mathcal{E}$ where $A \in \mathcal{E}_F$.
- Furthermore, we require a probability function P on $\mathcal{E}$ such that P is consistent with any finite run probability function P_F. i.e, $P\left(\pi_F^{-1}(A)\right) = P_F(A)$ for any finite $F \subseteq \mathbb{N}$ and $\forall A \in \mathcal{E}_F$ (in particular being consistent with the n-Run probability spaces $(\Omega_{[0,n-1]}, \mathcal{E}_{[0,n-1]}, P_{[0,n-1]})$)

Similar to what we stated at the end of the Introduction section, given the consistency of finite run probability spaces (as we proved in the above lemma), the existance of a probability function P with the properties mentioned above, is guarenteed by the Kolmogorov Extension theorem. In the next chapter we develop the mathematical prerequisites for proving the Kolmogorov extension theorem.

Using the Kolmogorov extension theorem, we will set up the probability spaces of infinite run Markov chains in Chapter 5

Chapter 3

Mathematical prerequisities

This chapter discusses the topological and measure theoretic prerequisities for further discussion on the Kolmogorov Extension Theorem. Some standard results from measure and topolgy are stated but not proved here. The reader may refer to the sources indicated alongside.

3.1 Topological prerequisities

Topological spaces can be considered as abstract generalizations of metric spaces (eg. $\mathbb{R}^n$) and it comprises of any non-empty set X along with a class of subsets of X (called open sets) satisfying certain closure properties.

Definition 3.1.1 (**Topological Space**). A set X along with $\mathcal{T}$ being a class of subsets of X is a topological space if the following holds:

1. $\mathcal{T}$ is closed under arbitrary unions
2. $\mathcal{T}$ is closed under finite intersections

$\mathcal{T}$ is said to be a topology on X.The subsets in $\mathcal{T}$ are called open sets and any subset of X is called a closed set if its complement is open. Topological spaces, say $(X, \mathcal{T})$ will be denoted as X when there is no confusion regarding the underlying topology $\mathcal{T}$.

A straightforward way to make any subset Y of X into a topological space is by defining the topology on that subset to be the set comprising of intersection Y with sets in $\mathcal{T}$. It is trivial to verify that this indeed is a topology on Y. This topology on $Y \subseteq X$ is termed as the *relative topology* on Y. The following definition is thus made,

Definition 3.1.2 (Relative Topology). Let $(X, \mathcal{T})$ be a topological space. Given any $Y \subseteq X$, $(Y, \{Y \cap \bar{T} : \bar{T} \in \mathcal{T}\})$ is a topological space and is called the relative topology on Y.

The notion of continuity of functions can be generalized as follows. Continuity comes handy in understanding the intuition behind *product spaces* to be defined soon.

Definition 3.1.3 (Continuous and Open Mapping). Let $(X, \mathcal{T})$ and $(Y, \mathcal{M})$ be a topological space. Let $f : X \to Y$.

- f is a continuous mapping if $f^{-1}(B) \in \mathcal{T}$ for all $B \in \mathcal{M}$
- f is an open mapping if $f(C) \in \mathcal{M}$ for all $C \in \mathcal{T}$

Intuitively, open mapping maps open sets in domain to open sets in image and inverse of open sets in image are open sets in domain in case of continuous mappings.

Open bases and subbases are to defined before defining the notion of product spaces.

Definition 3.1.4 (Open Base). Let $(X, \mathcal{T})$ be a topological space. A class of open sets $\{D_i\}_{i \in I}$ is called an open base of $(X, \mathcal{T})$ if any set in $\mathcal{T}$ can be written as union of sets in $\{D_i\}_{i \in I}$.

Definition 3.1.5 (Open Base generated by Open Subbase). Let $(X, \mathcal{T})$ be a topological space. A class of open sets $\{C_i\}_{i\in I}$ is called an open subbase if the set of all finite intersections of sets in $\{C_i\}_{i\in I}$ forms a open base for $(X, \mathcal{T})$. Here the open base is termed as the open base generated by the open subbase $\{C_i\}_{i\in I}$

The following fact is trivial and thus left unproven,

Lemma 3.1.1. *Let $\mathcal{Y} \subseteq 2^X$. Let $\mathcal{D}$ be the set of all finite intersections of sets in $\mathcal{Y}$. If $\mathcal{T}$ is the set of all unions of sets in $\mathcal{D}$, $\mathcal{T}$ is a topology on X. Or in other words, any subset of 2^X can serve as the open subbase for some topology of X. The topology is termed as the topology generated by $\mathcal{Y} \subseteq 2^X$*

In the above lemma, $\mathcal{D}$ serves as an open base for the topology on X. The following are certain terminologies regarding open bases and subbases.

Definition 3.1.6 (Subbasic and Basic Open Sets). Let $(X, \mathcal{T})$ be a topological space. Let $\{C_i\}_{i\in I}$ be an open subbase of $(X, \mathcal{T})$. Let the generated open base be $\{D_j\}_{j\in J}$. Sets in $\{C_i\}_{i\in I}$ are called subbasic open sets. Any subset of X is said to be a subbasic closed set if its complement is a subbasic open set. Sets in $\{D_j\}_{j\in J}$ are called basic open sets. Any subset of X is said to be a basic closed set if its complement is a basic open set.

If $\{(X_i, \mathcal{T}_i)\}_{i\in I}$ is a non-empty set of topological spaces. Consider the set $X = \prod\limits_{i\in I} X_i$. Among other topologies which could be defined on X, a topology of particular interest in the one in which all projections functions from X to the coordinate spaces are continuous. Lemma 3.1.1 suggests the following definition for **product topology** on X

Definition 3.1.7 (Product topology). Let $\{(X_i, \mathcal{T}_i)\}_{i\in I}$ is a non-empty set of topological spaces and $X = \prod\limits_{i\in I} X_i$. Let

$$\mathcal{Y} = \left\{\pi_i^{-1}(G_i) : G_i \in \mathcal{T}_i, i \in I\right\}$$

The topology generated by $\mathcal{Y}$ on X is called the product topology on X

We now consider a property of topological spaces called *compactness* and study the product of compact topological spaces. When it is stated that a class of subsets of a set X *cover* X it signifies that their union is equal to X. The following definitions have to be made before defining the notion of compactness,

Definition 3.1.8 (**Open cover**). An open cover in topological space $(X, \mathcal{T})$ is a set of open sets that cover X

Definition 3.1.9 (**Basic open cover**). A basic open cover in topological space $(X, \mathcal{T})$ is a set of basic open sets that cover X

Definition 3.1.10 (**Subbasic open cover**). A subbasic open cover in topological space $(X, \mathcal{T})$ is a set of subbasic open sets that cover X

Definition 3.1.11 (**Subcover**). A subcover of any class of sets that cover a set X, is a subclass which itself is a cover of X.

Now we define *compactness* of topological spaces,

Definition 3.1.12 (**Compact topological spaces**). A topological space $(X, \mathcal{T})$ is **compact** $\Leftrightarrow$ every open cover has a finite subcover.

The following definitions can simplify many arguments to follow,

Definition 3.1.13. A class of sets have **intersection property (IP)** if they have non-empty intersection

Definition 3.1.14. A class of subsets of set X has **finite intersection property (FIP)** if any finite subclass has non-empty intersection

The following fact is a consequence of the definition of compactness,

Lemma 3.1.2. *A topological space $(X, \mathcal{T})$ is compact $\Leftrightarrow$ every class of closed sets with FIP has IP*

Now we state an important theorem about compact topological spaces,

Theorem 3.1.3. *Let $(X, \mathcal{T})$ be a topological space. $(X, \mathcal{T})$ is compact $\Leftrightarrow$ Every subbasic open cover has a finite subcover $\Leftrightarrow$ Any class of subbasic closed sets with finite intersection property has intersection property.*

Now we are in a position to state an important result by Andrey Nikolayevich Tikhonov,

Theorem 3.1.4 (Tychonoff's theorem). *Product of non-empty class of compact topological spaces is compact*

A proof of the theorem can be found in Theorem A.2.2
The following definitions will be useful at a later stage,

Definition 3.1.15 (Second countable topological space). Topological space $(X, \mathcal{T})$ is *second countable* if it has a countable open base $\{B_n\}_{n \in \mathbb{N}}$

Definition 3.1.16 (Hausdorff topological space). Topological space $(X, \mathcal{T})$ is *Hausdorff* if for any $x, y \in X$ there exists $A, B \in \mathcal{T}$ such that $x \in A$, $y \in B$ and $A \cap B = \phi$

A proof for the following can be found in Chapter 5 of [25]. We do not include the proof since we do not make use of the below anywhere in the exposition.

Theorem 3.1.5. *$\{(X_i, \mathcal{T}_i)\}_{i \in I}$ is any family of Hausdorff spaces, then $\prod_{i \in I} X_i$ with the product topology is a Hausdorff space*

3.2 Measure Theoretic prerequisities

Now we introduce necessary definitions and results from Measure Theory. We begin with defining an algebra of sets.

Definition 3.2.1 (**Algebra of sets**). Let X be an arbitrary set. $\mathcal{A} \subseteq 2^X$ is defined to be an *algebra* of subsets of X if the following holds

- $X \in \mathcal{A}$
- If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
- If $A_1, A_2, \ldots A_n \in \mathcal{A}$, then $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$

An algebra by definition, is closed under finite unions. If an algebra is closed under countable unions, we define it to be a σ-algebra of sets.

Definition 3.2.2 (**Sigma algebra of sets**). Let X be an arbitrary set. $\mathcal{A} \subseteq 2^X$ is defined to be an σ-*algebra* of subsets of X if the following holds

- $X \in \mathcal{A}$
- If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
- If $A_1, A_2, A_3 \ldots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$

The following fact can be easily verified,

Lemma 3.2.1. *Let I be an arbitrary index set. If $\{\mathcal{A}_1\}_{i \in I}$ are σ-algebras of subsets of X, then $\bigcap_{i \in I} \mathcal{A}_i$ is also a σ-algebra of subsets of X*

And now we define σ-algebra generated by a collection of subsets,

Definition 3.2.3 (**Generated sigma algebra**). Let $\mathcal{A}$ be a collection of subsets of X. Consider the set $\mathcal{S} = \left\{\mathcal{H} \subseteq 2^X : \mathcal{H} \text{ is a } \sigma\text{-algebra and } \mathcal{A} \subseteq \mathcal{H}\right\}$ Then we define σ-algebra generated by $\mathcal{A}$, $\sigma(\mathcal{A}) = \bigcap_{\mathcal{H} \in \mathcal{S}} S$

By Lemma 3.2.1, $\sigma(\mathcal{A})$ is a σ-algebra. It is easy to verify that $\sigma(\mathcal{A})$ is the smallest sigma algebra containing $\mathcal{A}$ (in terms of containment). More precisely, if $\mathcal{M}$ is a σ-algebra such that $\mathcal{A} \subseteq \mathcal{M}$ then $\sigma(\mathcal{A}) \subseteq \mathcal{M}$.
We make few important definitions,

Definition 3.2.4 (**Borel sigma algebra**)**.** If $(X, \mathcal{T})$ is a topological space, we define the *Borel* σ-algebra $\mathcal{B}_X$ over X to be $\sigma(\mathcal{T})$

We are in a position to define *measurable spaces* and *measure spaces*

Definition 3.2.5 (**Measurable space**)**.** If X is an arbitrary set, $\mathcal{A}$ a σ-algebra over X then $(X, \mathcal{A})$ is called a measurable space

Definition 3.2.6 (**Measure**)**.** If X is an arbitrary set, $\mathcal{A}$ an algebra over X then $\mu : \mathcal{A} \mapsto [0, \infty]]$ is defined to be a measure on $\mathcal{A}$, if the following holds:

- $\mu(\phi) = 0$
- If $A_0, A_2, A_3 \ldots \in \mathcal{A}$ are a collection of pairwise disjoint sets in $\mathcal{A}$, i.e, $A_i \cap A_j = \phi$ if $i \neq j$ and if $\bigcup_{i=0}^{\infty} A_i \in \mathcal{A}$, then $\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \mu(A_i)$

If $\mathcal{A}$ is a σ-algebra , the tuple $(X, \mathcal{A}, \mu)$ will be referred to as a *measure space*

At this stage the above definitions can be motivated by looking back to our original intention of proving the existance of a probability measure over infinite product space with certain *desirable* properties. While approaching probability theory using measure theoretic tools, we define the set of events $\mathcal{E}$ over a sample space Ω as a *sigma*-algebra over Ω. Notice that certain properties we may require for the set of events are guarenteed by the definition of a σ-algebra (like countable union of events should be an event by itself etc). If a measure P is defined on $\mathcal{E}$ with range $[0, 1]$ such that $P(\Omega) = 1$, then P can behave similar to our intuitive notion of a probability function. Usual properties of the probability function such as the probability of disjoint set of events being the sum of their probabilities is guarenteed by the definition of a measure.

The following are some properties of measures. Proof of the following can be found in the appendix concerning Caratheodory theorem (refer lemma B.1.3 and lemma B.1.4)

Lemma 3.2.2 (Continuity from below). *$(X, \mathcal{M}, \mu)$ be a measure space. If $\{A_n\}_{n\in\mathbb{N}}$ be collection of sets from $\mathcal{M}$ such that $A_0 \subseteq A_1 \subseteq A_2 \subseteq \dots$, then $\mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$*

Lemma 3.2.3 (Continuity from above). *$(X, \mathcal{M}, \mu)$ be a measure space. If $\{A_n\}_{n\in\mathbb{N}}$ be collection of sets from $\mathcal{M}$ such that $A_0 \supseteq A_1 \supseteq A_2 \supseteq \dots$ and $\mu(X) < \infty$, then $\mu\left(\bigcap_{n=0}^{\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n)$*

The following is a major result which helps in extending a countable additive measure defined on an algebra to the generated sigma algebra

Theorem 3.2.4 (Caratheodory's Extension Theorem). *Let $\mu : \mathcal{A} \mapsto [0, \infty)$ be a finite measure on an algebra of sets $\mathcal{A}$. Then μ can be uniquely extended to a measure on the generated sigma algebra $\mathcal{B} = \sigma(\mathcal{A})$*

A proof of the above can be found in appendix on Caratheodory extension theorem (see theorem B.5.2).

Now we consider defining a σ-algebra over $\prod_{i\in I} X_i$ when we are given a family of measurable spaces $\{(X_i, \mathcal{M}_i)\}_{i\in I}$. The *product σ-algebra* over $\prod_{i\in I} X_i$ is defined below,

Definition 3.2.7 (Product sigma algebra). Given $\{(X_i, \mathcal{M}_i)\}_{i\in I}$, a family of measurable spaces. We define the *product σ -algebra* $\bigotimes\limits_{i\in I} \mathcal{B}_{X_i}$ over $\prod_{i\in I} X_i$ as $\sigma(\left\{\pi_i^{-1}(E) : E \in \mathcal{M}_i, i \in I\right\})$

For any $i \in I$ the set $\left\{\pi_i^{-1}(E) : E \in \mathcal{M}_i\right\}$ is called the set of cylinder sets with base i. Hence the product σ-algebra is the σ-algebra generated by the set of cylinder sets over all $i \in I$.

If we consider a family of measure spaces $\{(X_i, \mathcal{M}_i, \mu_i)\}_{i\in I}$ where $(X_i, \mathcal{T}_i)$ are topological spaces for each $i \in I$. Let $X = \prod_{i\in I} X_i$, in general $\bigotimes\limits_{i\in I} \mathcal{B}_{X_i} \neq \mathcal{B}_X$ (where $\mathcal{B}_X$ is the σ-algebra generated by the product topology over $X = \prod_{i\in I} X_i$).

However, it is easy to see that $\mathcal{B}_X$ will be atleast as big as $\bigotimes\limits_{i\in I} \mathcal{B}_{X_i}$ in the general case

Lemma 3.2.5. *Let $\{(X_i, \mathcal{M}_i, \mu_i)\}_{i\in I}$ be any family of measure spaces where $(X_i, \mathcal{T}_i)$ are topological spaces for each $i \in I$. Let $(X = \prod_{i\in I} X_i, T = \prod_{i\in I} T_i)$ be the product space with the product topology. We have, $\bigotimes\limits_{i\in I} \mathcal{B}_{X_i} \subseteq \mathcal{B}_X$*

But in a special situation the converse holds and these σ-algebras coincide as we prove in the next theorem,

Now we prove some necessary lemmae,

Lemma 3.2.6. *Let $(X, \mathcal{T})$ be a topological space. $(X, \mathcal{T})$ has a countable subbase $\Leftrightarrow$ $(X, \mathcal{T})$ has a countable base*

Proof. The converse is trivial since a countable base is also a countable subbase.
Let $\mathcal{C}$ be a countable subbase. The base $\mathcal{D}$ generated by $\mathcal{C}$ is countable, since it is the set of all finite intersections of sets in $\mathcal{C}$. Hence the result follows. □

We immediately obtain the following result,

Corollary 3.2.7. *Let $(X, \mathcal{T})$ be a topological space. $(X, \mathcal{T})$ is second countable $\Leftrightarrow$ $(X, \mathcal{T})$ has a countable subbase*

Lemma 3.2.8. *Let $(X, \mathcal{T})$ be a second countable topological space. If $\mathcal{C}$ is a countable subbase, then $\mathcal{B}_X = \sigma(\mathcal{C})$*

Proof. It is trivial to see that $\sigma(\mathcal{C}) \subseteq \mathcal{B}_X$. Let $\mathcal{D}$ be the countable base generated by $\mathcal{C}$. Since $\sigma(\mathcal{C})$ is closed under finite intersections, we have $\mathcal{D} \subseteq \sigma(\mathcal{C})$. Let G be any open set in $(X, \mathcal{T})$. $G = \bigcup_{i \in I} D_i$ where $D_i \in \mathcal{D}_i$. Since $\mathcal{D}$ is countable, I is countable. Since $\sigma(\mathcal{C})$ is closed under countable unions, $G \in \sigma(\mathcal{C})$. We obtain $\mathcal{T} \subseteq \sigma(\mathcal{C})$, proving that $\mathcal{B}_X \subseteq \sigma(\mathcal{C})$. Hence the result follows. $\square$

Now we prove the coincidence of $\mathcal{B}_X$ and $\bigotimes\limits_{i \in I} \mathcal{B}_{X_i}$ when the component spaces are second countable,

Theorem 3.2.9. *If $\{(X_i, \mathcal{T}_i)\}_{i \in I}$ are family of measurable spaces, I is countable and $(X_i, \mathcal{T}_i)$ are second countable topological spaces, then $\bigotimes\limits_{i \in I} \mathcal{T}_i = \mathcal{B}_X$ where $X = \prod_{i \in I} X_i$*

Proof. Since $\bigotimes\limits_{i \in I} \mathcal{B}_{X_i} \subseteq \mathcal{B}_X$ by Lemma 3.2.5, it is enough to show that $\mathcal{B}_X \subseteq \bigotimes\limits_{i \in I} \mathcal{B}_{X_i}$. Let C_i denote the countable subbase for $(X_i, \mathcal{T}_i)$. Consider the set

$$\mathcal{C} = \left\{\pi_i^{-1}(A) : i \in I, A \in \mathcal{C}_i\right\}$$

. Since each C_i is a subbase for $(X_i, \mathcal{T}_i)$ and using Lemma 3.1.1, it is easy to see that $\mathcal{C}$ is a subbase for the product topology, $\prod_{i \in I} \mathcal{T}_i$. Furthermore since each C_i is countable, $\mathcal{C}$ is a countable subbase for the product topology.

By definition, $\mathcal{C} \subseteq \bigotimes\limits_{i \in I} \sigma(\mathcal{C}_i)$. Now by Lemma 3.2.8 we have $\mathcal{C}_i = \mathcal{B}_{X_i}$ for each $i \in I$. i.e, $\mathcal{C} \subseteq \bigotimes\limits_{i \in I} \mathcal{B}_{X_i}$.

We observe that a countable subbase for the product topology (specifically $\mathcal{C}$) happens to be inside the product sigma algebra. Let $\mathcal{D}$ be the open base generated by $\mathcal{C}$. Since $\bigotimes\limits_{i \in I} \mathcal{B}_{X_i}$ is closed under finite intersections, $\mathcal{D} \subseteq \bigotimes\limits_{i \in I} \mathcal{B}_{X_i}$. Take any open set $G \in \mathcal{T}$. We have $G = \bigcup_{j \in J} D_j$ where $D_j \in \mathcal{D}$ for each $j \in J$. Since $\mathcal{D}$ is countable, J is countable. Since $\bigotimes_{i\in I} \mathcal{B}_{X_i}$ is closed under countable intersections, $G \in \bigotimes_{i\in I} \mathcal{B}_{X_i}$. This proves $\mathcal{T} \subseteq \bigotimes_{i\in I} \mathcal{B}_{X_i}$. Since $\mathcal{B}_X = \sigma(\mathcal{T})$, the theorem follows. □

There exist counter examples to the above theorem when the underlying spaces are not second countable. One of them is the Nedoma's pathology, which can be found in appendix E.

The following fact is directly employed in the proof for Kolmogorov Extension Theorem.

Theorem 3.2.10 (Continuity from above at the empty set). *Let $\mathcal{A}$ be an algebra of subsets of X, and $\mu : \mathcal{A} \mapsto \mathbb{R}^+$ a finite valued set function. And if $\{A_0, A_1, \ldots A_k\}$ is collection of pairwise disjoint sets from $\mathcal{A}$ implies $\mu(\bigcup_{i=0}^{k} A_i) = \sum_{i=0}^{k} \mu(A_i)$. Then the following are equivalent*

- *μ is countable additive on $\mathcal{A}$*
- *$\{A_n\}_{n\in\mathbb{N}}$ is collection of sets from $\mathcal{A}$ such that $A_0 \supseteq A_1 \supseteq A_2 \supseteq \ldots$ and $\bigcap_{n=1}^{\infty} A_n = \phi$ then $\lim_{n\to\infty} \mu(A_n) = 0$*

Proof. The forward implication is direct from the 'Continuity from above property' of finite measures (Theorem B.1.4).
Let $B_0, B_1, B_2, \ldots$ be a sequence of sets where $B_n \in \mathcal{A}$ for all $n \in \mathbb{N}$, such that $B_i \cap B_j = \phi$ for all $i, j \in \mathbb{N}$. Let $B = \bigcup_{n=0}^{\infty} B_n$. The backward implication follows if we show that $\mu(B) = \sum_{n=0}^{\infty} \mu(B_n)$.
Let $C_n = \bigcup_{i=0}^{n} B_i$. And let $A_n = B - C_n$. Observe that $A_0 = B \supseteq A_1 \supseteq A_2 \supseteq \ldots$ and $\bigcap_{n=1}^{\infty} A_n = \phi$. Hence from our assumption, $\lim_{n\to\infty} \mu(A_n) = 0$. By finite additivity we have $\mu(C_n) = \sum_{i=0}^{n} \mu(B_i)$. Also finite additivity implies, $\mu(B) = \mu(A_n) + \mu(C_n)$ for all $n \in \mathbb{N}$. Thus, $\lim_{n\to\infty} \mu(B) = \mu(B) = \lim_{n\to\infty} \mu(A_n) + \mu(C_n) = \lim_{n\to\infty} \sum_{i=0}^{n} \mu(B_i) = \sum_{i=0}^{\infty} \mu(B_i)$. □

Chapter 4

Kolmogorov Extension Theorem for finite discrete spaces

Before taking up the case of finite discrete spaces, we look into the Kolmogorov Extension Theorem for compact, Hausdorff and second countable spaces. Let $\{(X_i, \mathcal{T}_i)\}_{i\in I}$ be a family of topological spaces which are compact,second countable and Hausdorff. The general version of the theorem is attributed to the Russian mathematician Andrey Kolmogorov.

4.1 KET for product of compact, Hausdorff, second countable spaces

Let $\{(X_i, \mathcal{B}_i)\}_{i\in I}$ be a family of measurable spaces on X_i, where each $B_i = B_{X_i}$, the Borel sigma algebra over $(X_i, \mathcal{T}_i)$. If $F \subseteq I$ is finite, then $\pi_F : \prod_{i\in I} X_i \mapsto \prod_{f\in F} X_f$ will denote the usual projection function from $\prod_{i\in I} X_i$

to $\prod_{f\in F} X_f$. More precisely, $\pi_F((a_i)_{i\in I}) = ((a_f)_{f\in F})$ for all $(a_i)_{i\in I} \in \prod_{i\in I} X_i$. When F is a singleton,i.e, $F = \{k\}$, then we use π_k instead of $\pi_{\{k\}}$ (hence, the notation is consistent with our earlier use of π). Let $\mathcal{F}_I$ be the set of all finite subsets of I.

We make a few definitions before proceeding further. Let $\mathcal{B}$ denote the product σ-algebra over $\{(X_i, \mathcal{B}_i)\}_{i\in I}$. i.e, $\mathcal{B} = \bigotimes\limits_{i\in I} \mathcal{B}_i$

Definition 4.1.1. For each $F \in \mathcal{F}_I$, define $B_F = \{\pi_F^{-1}(E) : E \in \bigotimes\limits_{i\in F} \mathcal{B}_i\}$

B_F thus consist of all cylinder sets on base $F \in \mathcal{F}_I$. An immediate observation is,

Lemma 4.1.1. *For each $F \in \mathcal{F}_I$, B_F is a σ-algebra*

Now consider the set obtained by collecting all possible finite base cylinder set together, $\bigcup_{F\in\mathcal{F}_I} \mathcal{B}_F$. Let this be denoted as $\mathcal{A}$. We establish certain elementary facts about the entities defined above,

Lemma 4.1.2. *If $F, G \subseteq \mathcal{F}_I$ and $F \subseteq G$ then $\mathcal{B}_F \subseteq \mathcal{B}_G$*

Proof. Consider the set $\mathcal{P} = \{A : A \in \bigotimes\limits_{i\in F} \mathcal{B}_i$ and $\pi_F^{-1}(A) \in \mathcal{B}_G\}$. Since $F \subseteq G$, all the generating rectangles of $\bigotimes\limits_{i\in F} \mathcal{B}_i$ belongs to $\mathcal{P}$. If $A \in \mathcal{P}$, $\pi_F^{-1}(A^c) = (\pi_F^{-1}(A))^c \in \mathcal{B}_G$. Hence $\mathcal{P}$ is closed under complementation. If $\{A_i\}_{i=0}^{\infty}$ is a collection of sets from $\mathcal{P}$, then $\pi_F^{-1}(\bigcup\limits_{i=0}^{\infty} A_i) = \bigcup\limits_{i=0}^{\infty} \pi_F^{-1}(A_i) \in \mathcal{B}_G$. Hence $\mathcal{P}$ is closed under countable union. This shows that $\mathcal{P}$ is a σ-algebra. We obtain, $\bigotimes\limits_{i\in F} \mathcal{B}_i \in \mathcal{P} \implies \bigotimes\limits_{i\in F} \mathcal{B}_i = \mathcal{P}$. We have shown that if $A \in \bigotimes\limits_{i\in F} \mathcal{B}_i$ then, $\pi_F^{-1}(A) \in \mathcal{B}_G$, the lemma directly follows from this statement □

Corollary 4.1.3. *$\mathcal{A}$ is an algebra of subsets of X*

Proof. Let $\{A_i\}_{i=0}^{n}$ be a finite collection of sets from $\mathcal{A}$. Let $\{E_i\}_{i=0}^{n}$ be sets from $\mathcal{F}_I$ such that $A_i \in \mathcal{B}_{E_i}$. Let $E = \bigcup\limits_{i=0}^{n} E_i$. Now, from the above lemma, $A_i \in \mathcal{B}_E$ for $i \in \{0, 1, \ldots n\}$. Since $\mathcal{B}_E$ is a σ-algebra, the lemma follows. □

Lemma 4.1.4. *For any $F \in \mathcal{F}_I$, $\mathcal{B}_F \subseteq \mathcal{B}$*

Proof. Consider the set $\mathcal{P} = \{A : A \in \bigotimes\limits_{i\in F} \mathcal{B}_i \text{ and } \pi_F^{-1}(A) \in \mathcal{B}\}$. Obviously, all the generating rectangles of $\bigotimes\limits_{i\in F} \mathcal{B}_i$ belongs to $\mathcal{P}$. If $A \in \mathcal{P}$, $\pi_F^{-1}(A^c) = (\pi_F^{-1}(A))^c \in \mathcal{B}$. Hence $\mathcal{P}$ is closed under complementation. If $\{A_i\}_{i=0}^{\infty}$ is a collection of sets from $\mathcal{P}$, then $\pi_F^{-1}(\bigcup\limits_{i=0}^{\infty} A_i) = \bigcup\limits_{i=0}^{\infty} \pi_F^{-1}(A_i) \in \mathcal{B}$. Hence $\mathcal{P}$ is closed under countable union. This shows that $\mathcal{P}$ is a σ-algebra. We obtain, $\bigotimes\limits_{i\in F} \mathcal{B}_i \in \mathcal{P} \implies \bigotimes\limits_{i\in F} \mathcal{B}_i = \mathcal{P}$. We have shown that if $A \in \bigotimes\limits_{i\in F} \mathcal{B}_i$ then, $\pi_F^{-1}(A) \in \mathcal{B}$, the lemma directly follows from this statement. □

Lemma 4.1.5. $\sigma(\mathcal{A}) = \mathcal{B}$

Proof. The generating set of $\mathcal{B}$ is $\bigcup\limits_{i\in I}\{\pi_i^{-1}(A) : A \in B_i\}$. This is trivially a subset of the generating set of $\sigma(\mathcal{A})$. Hence we easily obtain $\mathcal{B} \subseteq \sigma(A)$.
We have $\mathcal{A} = \bigcup\limits_{F\in\mathcal{F}_I} \mathcal{B}_F$. From the above lemma, $\mathcal{B}_F \subseteq \mathcal{B}$ for each $F \in \mathcal{F}_I$.
Hence we get $\mathcal{A} \subseteq \mathcal{B} \implies \sigma(\mathcal{A}) \subseteq \mathcal{B}$
We have shown that $\sigma(\mathcal{A}) = \mathcal{B}$ □

Let $X = \prod_{i\in I} X_i$. Now we state Kolmogorov Extension Theorem for product of compact,Hausdorff and second countable topological spaces. If f is any function from arbitrary set A to arbitrary set Y. If $C \subseteq A$, $f|_C$ denotes the *restriction* of f to $C \subseteq A$. i.e, $f|_C(d) = f(d)$ for $d \in C$.

Theorem 4.1.6 (KET for product of compact, Hausdorff, second countable spaces). *Let P_F be a probability measure on (X_i, B_F) for each $F \in \mathcal{F}_I$. If the family of probability measures $\{P_F\}_{F\in\mathcal{F}_I}$ are 'consistent', i.e, if $F, G \in \mathcal{F}_I$ such that $F \subseteq G$ then $P_G|_{\mathcal{B}_F} = P_F$, then there exist a unique probability measure P on $(X, \mathcal{B})$ such that $P|_{\mathcal{B}_F} = P_F$ for any $F \in \mathcal{F}_I$*

Sketch of the proof. Define $\widetilde{P}$ on $\mathcal{A}$ as $\widetilde{P}(A) = P_F(A)$ when $A \in \mathcal{A}$ and $F \in \mathcal{F}_I$. This is well defined due to consistency conditions on the family $\{P_F\}_{F\in\mathcal{F}_I}$. It is easy to argue that $\widetilde{P}$ is a finitely additive set function on algebra $\mathcal{A}$. The traditional approach for proving Kolmogorov Extension Theorem starts by considering a collection of sets from $\mathcal{A}$ such that $A_0 \supseteq A_1 \supseteq A_2 \supseteq \dots$ and $\bigcap_{n=1}^{\infty} A_n = \phi$. Then by proving $\lim_{n\to\infty} P(A_n) = 0$, Lemma 3.2.10 can be used to argue countable additivity of P on the algebra. Now Caratheodory's Extension Theorem (see theorem B.5.2) can be used to prove that $\widetilde{P}$ defined on the algebra $\mathcal{A}$ can be extended uniquely to obtain a measure P on $\mathcal{B}$ which is an *extension* of the initial family of probability measures as required by the theorem. A proof for KET for a very similar case can be found in Chapter 4 in [2]. $\square$

We are concerned about proving the theorem for formulating the underlying probability space of discrete time finite state Markov Chains, where the individual spaces are mathematically much simpler. These are in fact finite and discrete spaces.

4.2 KET for product of finite discrete spaces

Now we consider a special case of discussions in Section 4.1, i.e, when $\{(X_i, \mathcal{T}_i)\}_{i\in I}$ are finite discrete. First let us recall the notation from Section 4.1. $\{(X_i, \mathcal{B}_i)\}_{i\in I}$ denotes a family of measurable spaces on X_i, where each $B_i = B_{X_i}$, the Borel sigma algebra over $(X_i, \mathcal{T}_i)$. We will denote $\prod_{i\in I} X_i$ by Ω. Now we consider the case when $(X_i, \mathcal{T}_i)$ are finite discrete for each $i \in I$. $\mathcal{B}$ denotes the product σ-algebra over $\{(X_i, \mathcal{B}_i)\}_{i\in I}$. i.e, $\mathcal{B} = \bigotimes_{i\in I} \mathcal{B}_i$. We defined $B_F = \{\pi_F^{-1}(E) : E \in \bigotimes_{i\in F} \mathcal{B}_i\}$ for each $F \in \mathcal{F}_I$. We further defined $\mathcal{A} = \bigcup_{F\in\mathcal{F}_I} \mathcal{B}_F$. We observed (in Corollary 4.1.3) that $\mathcal{A}$ is an algebra. Furthermore due to Lemma 4.1.5, $\sigma(\mathcal{A}) = \mathcal{B}$. Let $(X, \mathcal{T})$ denote the product space of $\{(X_i, \mathcal{T}_i)\}_{i\in I}$.

For simplicity we assume that $I = \mathbb{N}$. The general case can be argued without much difficulty.

We observe some elementary facts about the generating algebra $\mathcal{A} = \sigma(\bigcup_{F\in\mathcal{F}_{\mathbb{N}}} B_F)$. By a **rectangle**, we mean a set of the form $E = \pi_F^{-1}(A)$ (where $A \in \bigotimes_{f\in F} B_f$) which is precisely the set of all elements in $\mathcal{A}$. Now we define elementary rectangles,

Definition 4.2.1 (Elementary rectangle). Any set $E = \pi_F^{-1}(A) \in \mathcal{A}$ (for any $n, m \in \mathbb{N}$ and $A = (a_f)_{f\in F}$ is a singleton such that $A \in \bigotimes_{f\in F} B_f$) is called an elementary rectangle

The following lemma is a direct consequence of the finiteness of the component spaces

Lemma 4.2.1. *Let $E \in \mathcal{A}$ be any rectangle. $E = \bigcup_{i=1}^{k} E_i$ where each E_i is an elementary rectangle*

Proof. Let $\{a_i\}_{i=1}^k$ be an arbitrary indexing of $\bigotimes_{f\in F} B_f$. We can get such a finite indexing only becuase each $B_i = \mathcal{P}(X_i)$ is a finite set. Hence we can guarentee that the number of distinct elements $k \leq \prod_{f\in F} |X_f|$. Now, notice that $E = \bigcup_{i=1}^{k} \pi_F^{-1}(a_i)$. The lemma follows by observing that each $\pi_F^{-1}(a_i)$ is an elementary rectangle. $\square$

Now, we note down a corollary of lemma 4.1.2,

Corollary 4.2.2. *Let $F \in \mathcal{F}_{\mathbb{N}}$, then $B_F \subseteq B_{[m,n]}$ for some $m \leq n$*

Proof. A direct consequence of lemma 4.1.2 since for any finite set F, there exist $m, n \in \mathbb{N}, m \leq n$ such that $F \in [m, n]$. $\square$

Now, we make few important topological observations regarding $\mathcal{A}$ before proving KET for product of finite discrete spaces,

Lemma 4.2.3. *Any elementary rectangle $E = \pi_F^{-1}(A) \in \mathcal{A}$ (for any $n, m \in \mathbb{N}$ and $A = (a_f)_{f \in F}$ is a singleton such that $A \in \bigotimes\limits_{f \in F} B_f$) is a closed set in the product topology*

Proof. Notice that that for any $n \in \mathbb{N}$ and $a_n \in X_n$, $\pi_n^{-1}(a_n)$ is a closed set (since a_n is closed in the discrete topolgy on X_n). Since $E = \bigcap\limits_{f \in F} \pi_f^{-1}(a_f)$, it immediately follows that E is closed (closed sets are closed under arbitrary intersection, we only have a finite intersection here). $\square$

We get the following as a corollary,

Corollary 4.2.4. *Any $E \in \mathcal{A}$ is a closed set.*

Proof. From lemma 4.2.1, we get that $E = \bigcup\limits_{i=1}^{k} E_i$ where each E_i is an elementary rectangle. Each E_i is closed due to lemma 4.2.3 and since finite union of closed sets are closed, the result follows. $\square$

Theorem 4.2.5 (KET for product of finite discrete spaces). *Let the Topological spaces in the family $\{(X_i, \mathcal{T}_i)\}_{i \in I}$ be finite discrete spaces. Let P_F be a probability measure on (X, B_F) for each $F \in \mathcal{F}_I$. If the family of probability measures $\{P_F\}_{F \in \mathcal{F}_I}$ are 'consistent', i.e, if $F, G \in \mathcal{F}_I$ such that $F \subseteq G$ then $P_G|_{\mathcal{B}_F} = P_F$, then there exist a unique probability measure P on $(X, \mathcal{B})$ such that $P|_{\mathcal{B}_F} = P_F$ for any $F \in \mathcal{F}_I$*

Proof. Define P on $\mathcal{A}$ as $P(A) = P_F(A)$ when $A \in \mathcal{A}$ and $F \in \mathcal{F}_I$ where F is any finite subset of I such that $A \in B_F$. This is well defined due to consistency conditions on the family $\{P_F\}_{F \in \mathcal{F}_I}$. It is easy to observe that P is a finitely additive set function on algebra $\mathcal{A}$. We will prove that P is countable additive on $\mathcal{A}$. If $\{A_n\}_{n=1}^{\infty}$ be decreasing sequence of sets in $\mathcal{A}$ such that $\bigcap_{n=1}^{\infty} A_n = \phi$, to prove countable additivity of P, by Lemma 3.2.10 it is enough to prove that if $P(A_1) \geq P(A_2) \geq \dots$ and $P(A_n) \geq \epsilon$ for some

$\epsilon > 0$, then $\bigcap_{n=1}^{\infty} A_n \neq \phi$. Since the spaces are finite and discrete, each set $A_n \in \mathcal{A}$ is a closed set (due to corollary 4.2.4). Now, we have that $\{A_n\}_{n=1}^{\infty}$ is a decreasing sequence of non-empty closed sets . Since $(X, \mathcal{T})$ is compact due to Tychonoff's theorem, we have $\bigcap_{n=1}^{\infty} A_n \neq \phi$.
Thus we obtained a countably additive probability measure on $\mathcal{A}$ which can be uniquely extended to $\mathcal{B}$ by using the Caratheodory's Extension Theorem (see theorem B.5.2) $\square$

To conclude, we will consider the example of the hypothetical machine in section 2.1 where $I = \mathbb{N}$. We have natural probability functions $\widetilde{P}_F$ on each measurable space of the form $(\prod_{i \in F} X_i, \mathcal{E})$ where $\mathcal{E}$ is the power set of $\prod_{i \in F} X_i$ and F is a finite subset of $\mathbb{N}$. Now, we can define measures on $(\prod_{i \in \mathbb{N}} X_i, \mathcal{B}_F)$ (for all F finite subset of $\mathbb{N}$) as $P_F = \widetilde{P}_F \circ \pi_F$. It can be argued that this family of probability measures satisfy the *consistency* conditions required in KET. The individual spaces are $(\{a, b, c\}, 2^{\{a,b,c\}})$, are discrete and thus they are finite discrete spaces. Now Theorem 4.2.5 proves the existance of probability measure as we claimed at the end of section 2.1.

Chapter 5

Markov chains

In this chapter we extend the study of finite run Markov chains that was done in Chapter 2 to infinite run Markov chains.

5.1 Setting up Markov chains

At this stage we have in our hands, the Kolmogorov Extension theorem for product of finite discrete spaces. Now, we are in a position to complete the defintion of general finite state Markov chains. Given a Markov system (Q, μ_0, M), we are trying to define an appropriate set of events and a probability function on the infinite product space $Q^{\mathbb{N}}$.

On any finite set of indices $F \subseteq \mathbb{N}$, we had defined finite run probability spaces $(\Omega_F, \mathcal{E}_F, P_F)$. We had also shown that the finite run probability spaces are **consistent** with each other. i.e, Suppose $F \subseteq G \subseteq \mathbb{N}$. Then for any $A \in \mathcal{E}_F$, $P_G\left(\pi_{G\to F}^{-1}(A)\right) = P_F(A)$.

Let us recall the requirements for the infinite product probability space, that we had listed out at the end of Chapter 2.

- We require that the inverse projections of events in the finite run probability spaces are events in the infinite product probability space. If $\mathcal{E}$ is the set of events in the infinite product space, then given finite $F \subseteq \mathbb{N}$, $\pi_F^{-1}(A) \in \mathcal{E}$ where $A \in \mathcal{E}_F$.

- Furthermore, we require a probability function P on $\mathcal{E}$ such that P is consistent with any finite run probability function P_F. i.e, $P\left(\pi_F^{-1}(A)\right) = P_F(A)$ for any finite $F \subseteq \mathbb{N}$ and $\forall A \in \mathcal{E}_F$ (in particular being consistent with the n-Run probability spaces $(\Omega_{[0,n-1]}, \mathcal{E}_{[0,n-1]}, P_{[0,n-1]})$)

At this point, recall the notation developed in Chapter 4, $\mathcal{F}_I$ is the set of all finite subsets of I, $\mathcal{B}_F = \{\pi_F^{-1}(E) : E \in \bigotimes_{f \in F} B_f\}$, the algebra of cylindrical sets $\mathcal{A} = \bigcup_{F \in \mathcal{F}_I} \mathcal{B}_F$ and the product σ-algebra $\mathcal{B} = \sigma(\mathcal{A})$.

Theorem 5.1.1 (KET for product of finite Discrete spaces). *Let $\{(X_i, \mathcal{T}_i)\}_{i \in I}$ be a family of finite discrete topological spaces. Let P_F be a probability measure on (X, B_F) for each $F \in \mathcal{F}_I$. If the family of probability measures $\{P_F\}_{F \in \mathcal{F}_I}$ are 'consistent', i.e, if $F, G \in \mathcal{F}_I$ such that $F \subseteq G$ then $P_G|_{\mathcal{B}_F} = P_F$, then there exist a unique probability measure P on $(X, \mathcal{B})$ such that $P|_{\mathcal{B}_F} = P_F$ for any $F \in \mathcal{F}_I$*

The first requirement prompts us to define the σ-algebra to be $\sigma(\{A : A \in \mathcal{E}_F \text{ for some finite } F \subseteq \mathbb{N}\})$. This is precisely the product sigma algebra $\mathcal{B} = \bigotimes_{i \in I} \mathcal{B}_i = \bigotimes_{i \in I} \mathcal{E}_{\{i\}} = \sigma(\{A : A \in \mathcal{E}_F \text{ for some finite } F \subseteq \mathbb{N}\})$.
Now using the the family of finite run probability functions, we will see how we can obtain a family of consistent probability functions as required by the Kolmogorov extension theorem on $(Q^{\mathbb{N}}, B_F)$ for $F \in \mathcal{F}_{\mathbb{N}}$.
Consider some $F \in \mathcal{F}_{\mathbb{N}}$, On $(X, B_F = \{\pi_F^{-1}(A) : A \in \mathcal{E}_F\})$, when $E = \pi_F^{-1}(A)$, define $\widetilde{P}_F(E) = P_F(A)$. We will immediately prove why this family of probability functions so defined on $(Q^{\mathbb{N}}, B_F)$ for $F \in \mathcal{F}_{\mathbb{N}}$ is a consistent family,

Lemma 5.1.2. $\left\{\widetilde{P}_F\right\}_{F\in\mathcal{F}_{\mathbb{N}}}$ *is a consistent family of probability functions.*

Proof. We need to show when $F \subseteq G$, $\widetilde{P}_G|_{\mathcal{B}_F} = \widetilde{P}_F$.

Let $E \in \mathcal{B}_F$ such that $E = \pi_F^{-1}(A)$ where $A \in \mathcal{E}_F$. Now,

$$\begin{aligned}
\widetilde{P}_G(E) &= \widetilde{P}_G(\pi_F^{-1}(A)) \\
&= \widetilde{P}_G(\pi_G^{-1}(\pi_{G\to F}^{-1}(A))) \\
&= P_G(\pi_{G\to F}^{-1}(A)) \\
&= P_F(A) \text{ (by lemma 2.2.4)} \\
&= \widetilde{P}_F(\pi_F^{-1}(A)) \\
&= \widetilde{P}_F(E)
\end{aligned}$$

Since choice of E was arbitrary, the lemma follows. $\square$

Now, the Kolmogorov Extension theorem for finite discrete spaces guarentees the existance of a probability function $\widetilde{P}$ on $(X, \mathcal{B} = \bigotimes_{i\in I} \mathcal{B}_i)$ such that $\widetilde{P}|_{\mathcal{B}_F} = \widetilde{P}_F$ for any $F \in \mathcal{F}_{\mathbb{N}}$. This shows that the probability function given by KET meets the requirements specified by the aims we hoped to acheive on defining a probability space for infinite 'run' of a Markov system.

We thus obtained a probability space, $(Q^{\mathbb{N}}, \mathcal{B}, \widetilde{P})$ that satisifies the two requirements specified above. The infinite sequence of random variables $X_0, X_1, X_2, \ldots$, which are infact the projection functions to each coordinate of the inifnite product (i.e, $X_i = \pi_i$) will be referred to as a **general Markov chain** or a **Markov chain** in our further discussions. The probability function $\widetilde{P}$ will be replaced with P since we will no longer be concerned about finite run probability functions in the rest of the document.

5.2 Ergodic theorem for Markov chains

A important concept in the theory of Markov chains is that of a stationary distribution,

Definition 5.2.1 (Stationary distribution). Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ and associated Markov chain $X_0, X_1, X_2, \ldots$. A probability distribution vector $\mu = [\mu_i]_{i \in Q}$ is a said to be a stationary distribution if $P\mu = \mu$ (or in linear algebraic terms, μ is an eigenvector of P with eigenvalue 1). If μ_0 is a stationary distribution, then the corresponding Markov chain is said to be a stationary Markov chain.

We obtain an equivalent way of expressing the stationarity of a Markov chain,

Lemma 5.2.1. *Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ and associated Markov chain $X_0, X_1, X_2, \ldots$. The Markov chain is stationary if and only if for any event of the form $\{\omega \in \Omega : X_n(\omega) = q_0, X_{n+m_1}(\omega) = q_2, \ldots X_{n+m_k}(\omega) = q_{k-1}\}$ (for any $n, k \in \mathbb{N}$, sequence of states $q_0, q_1, \ldots q_{k-1}$ and sequence of natural numbers $m_1 < m_2 < \cdots < m_{k-1}$),*

$$
\begin{aligned}
&P\left(\{\omega \in \Omega : X_n(\omega) = q_0, X_{n+m_1}(\omega) = q_1, \ldots X_{n+m_{k-1}}(\omega) = q_{k-1}\}\right) \\
&= P\left(\{\omega \in \Omega : X_0(\omega) = q_0, X_{m_1}(\omega) = q_1, \ldots X_{m_{k-1}}(\omega) = q_{k-1}\}\right)
\end{aligned}
$$

Proof. Consider the forward implication. Since $\mu_0 = [\mu_{ij}]$ is a stationary distribution, we get

$$
\begin{aligned}
&P\left(\{\omega \in \Omega : X_n(\omega) = q_0, X_{n+m_1}(\omega) = q_1, \ldots X_{n+m_{k-1}}(\omega) = q_{k-1}\}\right) \\
&= \mu_{q_0} P^{m_1}[q_1, q_0] P^{m_2 - m_1}[q_2, q_1] \ldots P^{m_{k-1} - m_{k-2}}[q_{k-1}, q_{k-2}]
\end{aligned}
$$

Similarly,

$$P\left(\{\omega\in\Omega : X_0(\omega)=q_0, X_{m_1}(\omega)=q_1,\dots X_{m_{k-1}}(\omega)=q_{k-1}\}\right)$$
$$=\mu_{q_0}P^{m_1}[q_1,q_0]P^{m_2-m_1}[q_2,q_1]\dots P^{m_{k-1}-m_{k-2}}[q_{k-1},q_{k-2}]$$

Hence we obtain,

$$P\left(\{\omega\in\Omega : X_n(\omega)=q_0, X_{n+m_1}(\omega)=q_1,\dots X_{n+m_{k-1}}(\omega)=q_{k-1}\}\right)$$
$$=P\left(\{\omega\in\Omega : X_0(\omega)=q_0, X_{m_1}(\omega)=q_1,\dots X_{m_{k-1}}(\omega)=q_{k-1}\}\right)$$

Consider the backward implication. Applying the conditions in hypothesis to events of the form $P(\omega\in\Omega : X_i=q)$ (for any state q and $i\in\mathbb{N}$), we get

$$P\left(\{\omega\in\Omega : X_i(\omega)=q\}\right)=P\left(\{\omega\in\Omega : X_0(\omega)=q\}\right)$$

The above in particular implies that,

$$P\left(\{\omega\in\Omega : X_1(\omega)=q\}\right)=P\left(\{\omega\in\Omega : X_0(\omega)=q\}\right)$$

for any state $q\in Q$. This implies $P\mu_0=\mu_0$, or in other words the chain is a stationary Markov chain.

□

Our study will mostly focus on finite state ergodic Markov chains. We will define an ergodic Markov chain at this point.

Definition 5.2.2. Let $(Q,\mu_0,P=[p_{ij}])$ be a given Markov system. We use the notation $P^n[i,j]$ to refer to the entry in i^{th} row, j^{th} column of P^n

We also set up notation for the *path probability* over some finite sequence of states $q_1q_2q_3\dots q_k$ as,

Definition 5.2.3. Let $(Q,\mu_0,P=[p_{ij}])$ be a given Markov system.
We define $P[q_1q_2q_3\dots q_k]=p_{q_2q_1}p_{q_3q_2}p_{q_4q_3}\dots p_{q_kq_{k-1}}$

$P[q_1q_2q_3\ldots q_k]$ gives the possibility that the Markov chain moves along the *path* $q_1 \to q_2 \to q_3 \ldots q_{k-1} \to q_k$ in the successive steps.

Definition 5.2.4 (Ergodic Markov chain). Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ and associated Markov chain $X_0, X_1, X_2, \ldots$. $X_0, X_1, X_2, \ldots$ is an ergodic Markov chain if there exist a probability distribution vector $p = [p_i]$ such that $P^n[i,j] \to p_i$ as $n \to \infty$

The following lemma demonstrates the important *convergence to stationary distribution* property of ergodic Markov chains,

Lemma 5.2.2. *Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with stationary distribution $\mu = [\mu_i]_{i \in Q}$. Then,*

- *If $P^n[i,j] \to \mu_i$ as $n \to \infty$ then $P^n\mu_0 \to \mu$ as $n \to \infty$*
- *For all μ_0 if $P^n\mu_0 \to \mu$ as $n \to \infty$ then $P^n[i,j] \to \mu_i$ as $n \to \infty$*

Proof. First implication is immediate on taking μ_0 having 1 in its j^{th} row and 0 in i^{th} row for all $i \neq j$.
Now let us consider the second implication. Let $P^n[i,j] \to p_i$ as $n \to \infty$. This precisely says that $P^n \to Q$ where $Q = [q_{ij}]$ such that $q_{ij} = p_i$ for all $i, j \in Q$. Hence, $P^n\mu_0 \to Q\mu_0$ as $n \to \infty$. Since $q_{ij} = \mu_i$ for all $i, j \in Q$ and using the fact that μ_0 is a probability distribution, we get that $Q\mu_0 = \mu$. $\square$

The above also shows that an ergodic Markov chain has a unique stationary distribution

Lemma 5.2.3. *Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ and associated Markov chain $X_0, X_1, X_2, \ldots$.*

$X_0, X_1, X_2, \dots$ be an ergodic Markov chain (there exist a probability distribution vector $p = [p_i]$ such that $P^n[i, j] \to p_i$ as $n \to \infty$). The chain has a unique stationary distribution.

Proof. Lemma 5.2.2 shows that $p = [p_i]$ is indeed a stationary distribution. We will address the uniqueness now. Suppose $\nu = [\nu_i]$ is a stationary distribution, we have $P\nu = \nu$, furthermore $P^n\nu = \nu$. Set $\mu_0 = \nu$. From lemma 5.2.2, we get that $P^n\nu \to p$ as $n \to \infty$. We get, $\nu = p$, since $P^n\nu = \nu$. $\square$

One of the major objectives of this study would be to establish the result stated below, the ergodic theorem for Markov chains. Let $N(q, t)$ denote the number of times the Markov chain visits the state q before time t.

Definition 5.2.5 ($\mathbf{N_{q,t}}$)**.** Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$. For any $q \in Q$ and $t \in \mathbb{N}^+$, let $N_{q,t} : \Omega \mapsto \mathbb{R}$ be such that for any $\omega = (\omega_0, \omega_1, \omega_2, \omega_3, \dots) \in \Omega$, $N_{q,t}(\omega) = |\{0 \leq i < t : \omega_i = q\}|$

We are now in a position to state the ergodic theorem for stationary Markov chains,

Theorem 5.2.4 (**Ergodic theorem for stationary Markov chains**)**.** *Let $(Q, \mu_0 = [p_i], P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ defining a stationary Markov chain. Suppose the underlying Markov chain satisfies $P^n[i, j] \to p_i$ as $n \to \infty$ for all $i, j \in Q$, then for any $q \in Q$*

$$\lim_{t \to \infty} \frac{N_{q,t}}{t} = p_q \textit{ almost everywhere}$$

The ergodic theorem for Markov chains, intuitively provides a method to estimate the *fraction of time* that the stationary ergodic Markov chain has spent in a particular state q upto a particular time n. p_q - is an estimate for this fraction - $\frac{N_{q,t}}{t}$ which converges to p_q as $t \to \infty$. After developing basics of ergodic theory and Markov shift transformations, we give a proof of the ergodic theorem for Markov chains in Chapter 10

Chapter 6

Positive Markov chains

In this chapter we establish the fundamental result that positive Markov chains always have a stationary distribution and furthermore starting from any probability distribution, positive Markov chains (i.e a Markov chain having all entries in its transition matrix > 0) converge to the stationary distribution. An equivalent way to state this is that for all states i, j, $P^n[i, j] \to p_i$ as $n \to \infty$ (we will see why this is the case later).
Initially we venture into linear algebra and prove the Perron theorem for positive matrices. The above claims about positive Markov chains will easily follow from Perron theorem, as will be shown in this chapter.

6.1 Perron theorem for positive matrices

Perron's theorem for positive matrices is a specific case of the much more general Perron-Frobenius theorem which proves analogous statements for non-negative matrices. Given two real vectors $x = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$, we say $x \geq y$ if $x_i \geq y_i$ for all i. Similar conditions hold for $>, \leq$ and $<$.
By a positive (non-negative) matrix or vector, we mean a matrix or vector

with all its entries positive (non-negative).

Let us define the dominant eigenvalue of a matrix,

Definition 6.1.1 (Dominant eigenvalue). Let $A_{n\times n} = [a_{ij}]_{1\leq i,j\leq n}$ be a real matrix. $\lambda \in \mathbb{C}$ is a dominant eigenvalue of A if for any eigenvalue $\widetilde{\lambda}$, the following holds,

$$|\widetilde{\lambda}|\leq |\lambda|$$

We will state the Perron theorem for positive matrices before further discussion,

Theorem 6.1.1 (Perron theorem). *Let $A_{n\times n} = [a_{ij}]_{1\leq i,j\leq n}$ be a real positive matrix. A has an eigenvalue $\lambda(A)$ such that,*

1. *$\lambda(A)$ has a positive eigenvector $h > 0$*
2. *If $x \in \mathbb{C}^n$ is an eigenvector with eigenvalue $\lambda(A)$, then $x = ch$ for $c \in \mathbb{C}$. Or in other words, the dimension of the eigenspace of $\lambda(A)$ is 1.*
3. *If λ is any other eigenvalue of A, then $|\lambda|< \lambda(A)$*
4. *If $\lambda \neq \lambda(A)$ is an eigenvalue of A, then λ has no non-negative eigenvector (and thus, no non-positive eigenvectors)*

Perron theorem says that if A is a positive matrix, then A has a positive dominant eigenvalue $\lambda(A)$ and the the four conditions given in the theorem are true.

Before proving the result in the general case, we prove Perron theorem for positive symmetric matrices. In this special case, the proof is much simpler and helps in obtaining intuitions before dealing with the general case.

6.2 Perron theorem for positive symmetric matrices

We begin with the following elementary linear algebraic facts,

Lemma 6.2.1. *If M is a $n \times n$ real symmetric matrix, then all eigenvalues of M are real.*

Proof. Let λ be an eigenvalue with eigenvector x, i.e $Mx = \lambda x$. Now since $M^T = M$, $\lambda < x, x >=< \lambda x, x >=< Mx, x >=< x, M^T x >=< x, Mx >=< x, \lambda x >= \overline{\lambda} < x, x >$. We get that $< x, x > (\lambda - \overline{\lambda}) = 0$. Since $x \neq 0$, we get $\lambda = \overline{\lambda}$. This implies that λ is real. $\square$

Lemma 6.2.2. *If λ_1 and λ_2 are two non-conjugate eigenvalues (i.e $\lambda_1 \neq \overline{\lambda_2}$) of an $n \times n$ real matrix M, x_1 is an eigenvector of M corresponding to λ_1 and x_2 is an eigenvector of M^T corresponding to λ_2, then $< x_1, x_2 >= 0$*

Proof. Since M is real, $< Mx_1, x_2 >=< x_1, M^T x_2 >$. Since λ_1 and λ_2 are real, we obtain $\lambda_1 < x_1, x_2 >= \overline{\lambda_2} < x_1, x_2 >$. i.e, $(\lambda_1 - \overline{\lambda_2}) < x_1, x_2 >= 0$. Since $\lambda_1 \neq \overline{\lambda_2}$, the lemma follows $\square$

We also define doubly stochastic matrices,

Definition 6.2.1 (Doubly stochastic matrix). A real matrix $A_{n \times n} = [a_{ij}]$ is defined to be a doubly stochastic matrix if $0 \leq p_{ij} \leq 1$ for all $i, j \in \{1, 2, \ldots n\}$, $\sum_{1 \leq i \leq n} a_{ij} = 1$ for all $1 \leq j \leq n$ and $\sum_{1 \leq j \leq n} a_{ij} = 1$ for all $1 \leq i \leq n$

Doubly stochastic matrices are thus stochastic matrices with row sums $= 1$. Now we state and prove the Perron theorem for positive symmetric stochastic matrices,

Theorem 6.2.3 (Perron theorem for positive symmetric stochastic matrices). *Let P be a positive symmetric stochastic $n \times n$ matrix. Then,*

1. *1 is an eigenvalue for P with positive eigenvector $h = [1, 1, 1, \ldots 1] \in \mathbb{R}^n$*
2. *The dimension of eigenspace of 1 is 1. i.e if y is any eigenvector with eigenvalue 1, then $y = ch$ where $c \in \mathbb{C}$*
3. *If λ is any other eigenvalue of P, then $|\lambda| < 1$*
4. *No other eigenvalue has a non-negative eigenvector*

Proof. We consider each of the conclusions one-by-one

1. Since symmetric stochastic matrices are doubly stochastic, we have $\sum\limits_j p_{ij} = 1$ for all j. This precisely says that $h = [1, 1, 1, \ldots 1]$ is an eigenvector with eigenvalue $= 1$.
2. Let y be a real eigenvector with eigenvalue 1. Let i be such that y_i is the coordinate in y having the largest absolute value. We have $\sum\limits_j p_{ij} y_j = y_i$. Now $\left|\sum\limits_j p_{ij} y_j\right| \leq \sum\limits_j |p_{ij} y_j| \leq \sum\limits_j |p_{ij}| \, |y_i| = |y_i| \sum\limits_j p_{ij} = |y_i|$. Since $\sum\limits_j p_{ij} y_j = y_i$, the triangle inequality holds with equality. This implies all y_i are either non-negative or non-positive. Now from $\sum\limits_j p_{ij} y_j = y_i$ and $\sum\limits_j p_{ij} = 1$ we conclude that $y_j = y_i$ for all j since y_i has the maximum absolute value among all the coordinates. Hence y is a real multiple of $h = [1, 1, 1, \ldots 1]$. If y was a complex eigenvector with eigenvalue 1. We can write $y = a + ib$ where a and b are real eigenvectors. From the previous argument we get $a = ch$ and $b = dh$ where $c, d \in \mathbb{R}$. Hence $y = (c + id)h$. The proof for this part is complete.

3. Let y be be an eigenvector with eigenvalue λ. Let i be such that y_i is the coordinate in y having the largest absolute value. We have $\sum\limits_j p_{ij}y_j = \lambda y_i$. Now $|\lambda|\,|y_i| = \left|\sum\limits_j p_{ij}y_j\right| \leq \sum\limits_j |p_{ij}y_j| \leq \sum\limits_j |p_{ij}|\,|y_i| = |y_i|\sum\limits_j p_{ij} = |y_i|$. Hence we get that $|\lambda| \leq 1$. Since, all the eigenvalues of P are real (by lemma 6.2.1), all that is left is to rule out the possibility that -1 is an eigenvalue. Let y be be an eigenvector with eigenvalue -1. Let i be such that y_i is the coordinate in y having the largest absolute value. Without loss of generality we can assume $y_i > 0$ (else, we will consider $-y$ instead) We have $\sum\limits_j p_{ij}y_j = -y_i$. We have for all j, $-y_i \leq y_j \leq y_i$. Since $y_i, p_{ii} > 0$, $\sum\limits_j p_{ij}y_j \geq \sum\limits_{j\neq i} p_{ij}y_j \geq (-y_i)\sum\limits_{j\neq i} p_{ij} = (-y_i)(1-p_{ii})$. We get $-y_i \geq (-y_i)(1-p_{ii})$. Since $y_i > 0$, we get $(1-p_{ii}) \geq 1$ which implies $p_{ii} \leq 0$ which is a contradiction.

4. Since P^T is also a positive symmetric stochastic matrix, 1 is an eigenvalue of P^T with eigenvector $h = [1, 1, 1, \ldots 1]$. Let $\lambda \neq 1$ be an eigenvalue of P with eigenvector y having non-negative entries. By lemma 6.2.2, we have $< h, y >= \sum\limits_i y_i = 0$. But since y is non-negative, this is impossible and hence if $\lambda \neq 1$ is an eigenvalue, then it cannot have a non-negative eigenvector.

$\square$

Note: It should noticed that the above proof does not use the positivity of all the matrix entries, but only of those along the diagonal. Hence the theorem remains true if P has positive diagonal entires.

6.3 Proof of Perron theorem for positive matrices

The following function is used in the proof of the Perron theorem in the general case,

Definition 6.3.1 (Coordinate sum function). Define coordinate sum function $\sigma : \mathbb{R}^n \mapsto \mathbb{R}$ to be the function that sums up the coordinates of any vector in $\mathbb{R}^n$. i.e, $\sigma(\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}) = \sum_{i=1}^{n} x_i$.

We will begin with the proof of Perron theorem for positive matrices. We restate the theorem here,

Theorem 6.3.1 (Perron theorem). *Let $A_{n\times n} = [a_{ij}]_{1\leq i,j\leq n}$ be a real positive matrix. A has an eigenvalue $\lambda(A)$ such that,*

1. *$\lambda(A)$ has a positive eigenvector $h > 0$*
2. *If $x \in \mathbb{C}^n$ is an eigenvector with eigenvalue $\lambda(A)$, then $x = ch$ for $c \in \mathbb{C}$. Or in other words, the dimension of the eigenspace of $\lambda(A)$ is 1.*
3. *If λ is any other eigenvalue of A, then $|\lambda| < \lambda(A)$*
4. *If $\lambda \neq \lambda(A)$ is an eigenvalue of A, then λ has no non-negative eigenvector (and thus, no non-positive eigenvectors)*

The following proof is taken from [15]

Proof. We prove the four statements one by one,

1. We define the set $E = \{\lambda \in \mathbb{R} :$ there exist $x \in \mathbb{R}^n, x \geq 0$ such that $Ax \geq \lambda x\}$. Since A is positive, for any $x \geq 0$, $Ax > 0$ and hence $Ax \geq \lambda x$

holds for sufficiently small $\lambda > 0$. Hence E is non-empty and contains positive elements. We also show that E is a closed and bounded subset of $\mathbb{R}$.

Let $\{\lambda_i\}_{i=0}^{\infty}$ be a sequence from E such that $\lambda_i \to \lambda$. There exist a sequence $\{y_i\}_{i=0}^{\infty}$ such that $y_i \in \mathbb{R}^n, y_i \geq 0$ and $Ay_i \geq \lambda_i y_i$. By dividing $Ay_i \geq \lambda_i y_i$ by the sum of its coordinates, say s_i, we see that the condition $\frac{1}{s_i}Ay_i \geq \frac{1}{s_i}\lambda y_i$ holds. Hence, we can without loss of generality assume that for each y_i, the sum of its coordinates, s_i is equal to 1. Under the usual Euclidian norm, $C = \{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1\}$ is a closed and bounded set subset of $\mathbb{R}^n$. Now by using the Heine Borel theorem for R^n (theorem D.1.2), we get that C is a compact subset of $\mathbb{R}^n$. Hence there exist a convergent subsequence for $\{y_i\}_{i=0}^{\infty}$, $\{y_k\}_{k \subseteq \mathbb{N}}$ such that $y_k \to \widetilde{y} \in C$ as $k \to \infty$. For each k, we have $Ay_k \geq \lambda_k y_k$. Applying limit as $k \to \infty$ on both sides, we get $A\widetilde{y} \geq \lambda\widetilde{y}$. Hence $\widetilde{y} \in E$.

Let $\lambda \in E$ and let x be the corresponding non-negative vector such that $Ax \geq \lambda x$. Let k be such that x_k is the largest vector of x. Let $a_{max} = \max_{1 \leq i,j \leq n} \{a_{ij}\}$. Now, the k^{th} coordinate of Ax is $\sum_{j=1}^{n} a_{kj}x_j \leq \sum_{j=1}^{n} a_{max}x_j \sum_{j=1}^{n} a_{max}x_k = na_{max}x_k$. Since $Ax \geq \lambda x$, we get $\lambda x_k \leq na_{max}x_k \implies \lambda \leq na_{max}$. Since λ was arbitrary, we have shown that E is a bounded set.

Since, closed and bounded subsets of $\mathbb{R}$ has a maximum, we have shown that E has a maximum. Let this be denoted by $\lambda(A)$. Since E has positive elements, we get that $\lambda(A) > 0$

Next, we have to show that $\lambda(A)$ is an eigenvalue of A. Let $h = \begin{bmatrix} h_1 \\ h_2 \\ \cdots \\ h_n \end{bmatrix} \in \mathbb{R}^n$ be such that $Ah \geq \lambda(A)h$. We have to show that for no $k \in \{1, 2, \ldots n\}$, $\sum_{j=1}^{n} a_{kj}h_j > \lambda(A)h_k$ is a possibility. If so we obtain

$\sum_{j=1}^{n} a_{ij}h_j = \lambda(A)h_i$ for all $i \in \{1, 2, \ldots n\}$, hence proving that $\lambda(A)$ is an eigenvalue. Assume the contrary for some k, i.e $\sum_{j=1}^{n} a_{kj}h_j > \lambda(A)h_k$. Let $e_k \in \mathbb{R}^n$ such that the k^{th} coordinate of e_k is 1 and every other coordinate is 0. For some real $\epsilon > 0$, consider $y = h + \epsilon e_k$. Considering Ay, We get $\sum_{j=1}^{n} a_{ij}y_j > \lambda(A)y_i$ for every $i \neq k$. Furthermore, by choosing ϵ small enough we get $\sum_{j=1}^{n} a_{kj}y_j > \lambda(A)y_i$. i.e $Ay > \lambda(A)y$. But this means we can choose $\delta > 0$ small enough such that $Ay > (\lambda(A) + \delta)y$. Since y is a non-negative vector, this contradicts the maximality of $\lambda(A)$.

Hence we have shown that $\lambda(A)$ is an eigenvalue of A

Finally, we will show that $h > 0$. Since A is positive and h is non-negative, we get that $Ah > 0$. Since $Ah = \lambda(A)h$ and $\lambda(A) > 0$, we get $h > 0$. The proof of the first assertion is complete.

2. We argue that it is enough to show the second assertion when x is a real vector with atleast one positive entry, i.e if x is an eigenvector corresponding to $\lambda(A)$ having atleast one coordinate positive, then $x = rh$ for $r \in \mathbb{R}$. This trivially shows $x = rh$ when x has all positive entries. When x has all negative entries, we get $-x = rh$ for some $r \in \mathbb{R}$ and hence $x = -rh$. When x is a complex eigenvector $x = a + ib$ (where $a, b \in \mathbb{R}^n$), since $\lambda(A)$ is real, we get that a and b are real eigenvectors with eigenvalue $\lambda(A)$. Applying the previous argument to a and b, we get $a = r_1 h$ and $b = r_2 h$. Hence, we obtain $x = (r_1 + ir_2)h$. Hence, in general we get that if x is an eigenvector with eigenvalue $\lambda(A)$, then $x = ch$ for $c \in \mathbb{C}$.

   All that is left is to show that when x is an eigenvector corresponding to $\lambda(A)$ having atleast one coordinate positive, then $x = rh$ for $r \in \mathbb{R}$.

Assume the contrary. Now, let $r = \min_{1 \le i \le n} \left\{ \frac{h_i}{x_i} : x_i > 0 \right\}$. Now by our assumtion, $h + rx$ is non-zero and hence is an eigenvector with eigenvalue $\lambda(A)$ and atleast one entry $= 0$. But, we have earlier shown that any non-negative eigenvector corresponding to $\lambda(A)$ should in fact be a positive eigenvector. Hence our assumption must be false. This implies the existance of $r \in \mathbb{R}$ such that $x = rh$.

Now, we show that $\lambda(A)$ has no generalized eigenvector. Let x be an m^{th} order generalized eigenvector. Let $y_1, y_2, \dots y_{m-1}$ be the intermediate eigenvectors and let $m \ge 2$. From the above we get that $y_{m-1} = ch$ for $c \in \mathbb{C}$. This implies the existance of a $y \in \mathbb{C}^n$ such that $Ay = \lambda(A)y + ch$ (since $y = y_{m-2}$ works). We can without loss of generality assume that $c > 0$ (since this holds by taking either y or $-y$ in the equation). Now since $\lambda(A)$ is real, in case y is complex, we can seperate out the real parts on both sides and obtain that there exist $z \in \mathbb{R}^n$ such that $Az = \lambda(A)z + ch$. Without loss of generality we can assume that $z \ge 0$, else we can replace z with $z + rh$ for large enough real $r > 0$. This implies $Az > \lambda(A)z$ (since $c > 0$ and h is a positive vector). Now we can choose $\delta > 0$ small enough such that $Ay > (\lambda(A) + \delta)y$. Since z is non-negative, we get that $\lambda(A) + \delta \in E$ contradicting the maximality of $\lambda(A)$. Hence our assumption that there exist a generalized eigenvector must be false.

3. Let $\kappa \in \mathbb{C}$, $y \in \mathbb{C}^n$ be such that $Ay = \kappa y$. Applying triangle inequality for complex numbers, we get

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} |y_j| &\ge \left| \sum_{j=1}^{n} a_{ij} y_j \right| \\ &= |\kappa| \, |y_i| \end{aligned}$$

This says that $|\kappa| \in E$ since $\begin{bmatrix} |y_1| \\ |y_2| \\ \cdots \\ |y_n| \end{bmatrix}$ is a non-negative vector satisfying the necessary condition. This shows that $|\kappa| \leq \lambda(A)$.

All that is left is rule out the possibility that $\kappa \neq \lambda(A)$ and $|\kappa| = \lambda(A)$. Now, if $|\kappa| = \lambda(A)$, since $\sum_{j=1}^{n} a_{ij}\,|y_j| \geq |\kappa|\,|y_i|$, using a verbatim argument as in proof of assertion 1, we can show that $\begin{bmatrix} |y_1| \\ |y_2| \\ \cdots \\ |y_n| \end{bmatrix}$ is infact an eigenvector with eigenvalue $|\kappa| = \lambda(A)$. From assertion 2, we also get that $|y_i| = rh_i$ for some fixed $r \in \mathbb{R}$ This implies that equality holds in the application of triangle inequality above. Equality holds in the triangle inequality involving complex numbers if and only if all the complex numbers have the same complex argument. Hence for each i, we get $y_i = |y_i|e^{i\theta}$ for some fixed argument θ. Since $|y_i| = rh_i$, we get that $y = (re^{i\theta})h$. Now this precisely says that $\kappa = \lambda(A)$. Hence $\kappa \neq \lambda(A)$ and $|\kappa| = \lambda(A)$ is impossible.

The proof of assertion 3 is thus complete.

4. Since A is positive, we get that A^T is also positive. We get that $\lambda(A)$ is the dominant eigenvector of A^T and applying the first statement of Perron theorem, we get that A^T has a positive eigenvector $y > 0$ with eigenvalue $\lambda(A)$. Let $\lambda \neq \overline{\lambda(A)}$ since $\lambda(A)$ is real. From lemma 6.2.2, we know that eigenvectors of A and A^T corresponding to λ and $\lambda(A)$ annihilate each other. If λ has a non-negative eigenvector x, we get that $< y, x >= 0$. But since $y > 0$ and $x \geq 0$, $x \neq 0$, we get that $< y, x >> 0$. This is a contradiction. Hence if $\lambda \neq \lambda(A)$ then λ cannot have a non-negative eigenvector.

$\square$

6.4 Generalized eigenvectors and spectral theorem

Before looking at the application of Perron theorem for stochastic matrices, we develop the theory of generalized eigenvectors and the spectral theorem, both of which are critical to the proofs we do in section 6.5.

Let us recall that given a complex linear transformation matrix $A_{n\times n}$, $\lambda \in \mathbb{C}$ is said to be an eigenvalue if there exist $x \neq 0$ such that $(A-\lambda I)x = 0$ or equivalently when a non-negative vector x is a member of the null space of the transformation $A-\lambda I$. The non-negative vectors that satisfy $(A-\lambda I)x = 0$ are called eigenvectors of A corresponding to λ. Given a matrix $A_{n\times n}$, it is not always necessary that there exist an eigenvector decomposition for $\mathbb{C}^n$. For example, such a decomposition is guarenteed when the matrix is symmetric (See more in [15]). Now, we extend our usual notion of an eigenvector into *generalized eigenvectors*. By the end of the chapter we will state the spectral theorem which argues that a decomposition into *generalized eigenvectors* is always possible given any linear transformation matrix $A_{n\times n}$.

Definition 6.4.1 (**Generalized eigenvalues and eigenvectors**)**.** Let $A_{n\times n}$ be a complex matrix. $\lambda \in \mathbb{C}$ is said to be a generalized eigenvalue of A when there exist a positive $m \in \mathbb{N}$ and $x \in \mathbb{C}^n$ such that $(A-\lambda I)^m x = 0$. x is said to be a generalized eigenvector corresponding to generalized eigenvalue λ. Also, λ is said to be a m^{th} order eigenvalue and x, a m^{th} order eigenvector of A

Let us further inquire into the above definition. When x is an m^{th} order eigenvector, we have $(A-\lambda I)^m x = 0$. What happens when we take lower powers of $(A-\lambda I)$ and apply it on x?. We note that we obtain a hierarchy of generalized eigenvectors $h_1, h_2, \dots h_{m-1}$,

$$\begin{aligned}(A-\lambda I)x &= h_1\\ (A-\lambda I)^2x &= h_2\\ (A-\lambda I)^3x &= h_3\\ &\dots\\ (A-\lambda I)^ix &= h_i\\ &\dots\\ (A-\lambda I)^{m-1}x &= h_{m-1}\\ (A-\lambda I)^mx = (A-\lambda I)h_{m-1} &= 0\end{aligned}$$

From the above, we notice that h_{m-1} is infact an eigenvector of A (equivalently a first order generalized eigenvector) and $h_1, h_2 \dots h_{m-2}$ are generalized eigenvectors of order $m-1$, $m-2$ etc ... 2 respectively.
We prove a useful technical lemma about the behavior of a m^{th} order eigenvector x under successive applications of A. This lemma is employed in proving the convergence theorem of positive stochastic matrices in Chapter 6.

Lemma 6.4.1. *Let x be an m^{th} order generalized eigenvector. Let $h_1, h_2 \dots h_{m-1}$ be the intermediate eigenvectors. Then $A^nx = \lambda^n x + \binom{n}{1}\lambda^{n-1}h_1 + \binom{n}{2}\lambda^{n-2}h_2 + \dots \binom{n}{m-1}\lambda^{n-(m-1)}h_{m-1}$*

Proof. Proof proceeds by induction on n
When $n = 1$, $A^1x = \lambda x + h_1$. This proves the base case.
Assume the result for n, we will prove it for $n+1$.
We have $A^nx = \lambda^n x + \binom{n}{1}\lambda^{n-1}h_1 + \binom{n}{2}\lambda^{n-2}h_2 + \dots \binom{n}{m-1}\lambda^{n-(m-1)}h_{m-1}$

Now,

$$\begin{aligned}
A^{n+1}x &= \lambda^n Ax + \binom{n}{1}\lambda^{n-1}Ah_1 + \binom{n}{2}\lambda^{n-2}Ah_2 + \dots \binom{n}{m-1}\lambda^{n-(m-1)}Ah_{m-1} \\
&= \lambda^n(\lambda x + h_1) + \binom{n}{1}\lambda^{n-1}(\lambda h_1 + h_2) + \binom{n}{2}\lambda^{n-2}(\lambda h_2 + h_3) + \dots \\
&\dots + \binom{n}{i}\lambda^{n-i}(\lambda h_i + h_{i+1}) + \dots \binom{n}{m-1}\lambda^{n-(m-1)}(\lambda h_{m-1}) \\
&= \lambda^{n+1}x + (n+1)\lambda^n h_1 + \dots + \left(\binom{n}{i-1} + \binom{n}{i}\right)\lambda^{n+1-i}h_{m-i} + \dots \\
&+ \left(\binom{n}{m-2} + \binom{n}{m-1}\right)\lambda^{n+1-(m-1)}h_{m-1} \\
&= \lambda^{n+1}x + \binom{n+1}{1}\lambda^{n+1-1}h_1 + \dots + \binom{n+1}{i}\lambda^{n+1-i}h_{m-i} + \dots \\
&+ \binom{n+1}{m-1}\lambda^{n+1-(m-1)}h_{m-1}
\end{aligned}$$

Above we have used the fact that $\binom{n}{i-1} + \binom{n}{i} = \binom{n+1}{i}$. The lemma follows. $\square$

We have developed sufficient theory to state the spectral theorem

Theorem 6.4.2 (Spectral Theorem). *Let $A_{n\times n}$ be a complex matrix. Every vector $x \in \mathbb{C}^n$ has a decomposition into generalized eigenvectors of A. i.e, $x = x_1 + x_2 + \dots + x_m$ (for some $m \in \mathbb{N}, m \leq n$) where x_i are generalized eigenvectors of A.*

We do not prove the spectral theorem here. A proof of the spectral theorem can be found in appendix F (see theorem F.1.1).

In the next section, we proceed to a major objective of the chapter, the application of Perron theorem for positive matrices to prove the convergence theorem for stochastic matrices.

6.5 Stochastic matrices

Recall the definition of stochastic matrices, a matrix $S = [s_{ij}]_{1\leq i,j\leq n}$ is said to be a stochastic matrix if $\sum\limits_{i=1}^{n} s_{ij} = 1$ for all $1 \leq j \leq n$.
In earlier chapters, the transition matrix of a Markov chain was found to be a stochastic matrix. We prove an important result about the convergence of $S^n x$ when $x \in \mathbb{R}^n$ is a non-negative vector.
The following lemma will be employed in the proof of the convergence theorem. But the reader may defer this lemma till the convergence theorem refers back to it.

Lemma 6.5.1. *Let x be an m^{th} order generalized eigenvector. Then $A^n f \to 0$ as $n \to \infty$ if the corresponding eigenvalue λ is such that $|\lambda| < 1$*

Proof. From the lemma 6.4.1,
$A^n x = \lambda^n x + \binom{n}{1}\lambda^{n-1}h_1 + \binom{n}{2}\lambda^{n-2}h_2 + \dots \binom{n}{m-1}\lambda^{n-(m-1)}h_{m-1}$.
The lemma follows by taking limit as $n \to \infty$. □

We prove a fundamental fact about stochastic matrices,

Lemma 6.5.2. *Let $S = [s_{ij}]_{1\leq i,j\leq n}$ be a stochastic matrix . Then, $\lambda(S) = 1$*

Proof. Since $\sum\limits_{i=1}^{n} s_{ij} = 1$, we observe that $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ is an eigenvector of S. Since, 1 has a non-negative eigenvector, by assertion 4 of theorem 6.3.1, we get that $\lambda(S) = 1$. □

Now, we prove the $S^n x$ convergence theorem.

Theorem 6.5.3. *Let $S = [s_{ij}]_{1\leq i,j\leq n}$ be a stochastic matrix and $x \in \mathbb{R}^n$ be any non-negative vector.*
Then, $S^n(x) \to ch$ as $n \to \infty$ where c is a positive real number and h is the eigenvector corresponding to $\lambda(S)$

Proof. Using the spectral theorem (theorem F.1.1), we get a decomposition of x as sum of generalized eigenvectors. i.e $x = \sum_{i=1}^{n} c_i h_i$ (without loss of generality, $c_i \in \mathbb{R}$ for all i since x is a real vector). Let λ_i denote the corresponding eigenvalues. From lemma 6.5.2, $\lambda_1 = 1$. We will assume that the dominant eigenvector is h_1. We will refer to c_1 as c . Now since the dimension of the eigenspace of $\lambda(S)$ is 1 and its has no generalized eigenvectors, we get that $|\lambda_i| < \lambda_1 = \lambda(S) = 1$ (all conclusions made using theorem 6.3.1).
Now

$$\begin{aligned} S^n x &= cS^n h + \sum_{i=2}^{n} c_i S^n h_i \\ &= ch + \sum_{i=2}^{n} c_i S^n h_i \end{aligned} \tag{6.1}$$

If $i > 1$ and h_i is an ordinary eigenvector, then $S^n h_i = \lambda_i^n h_i \to 0$ as $n \to \infty$ since $|\lambda_i| < 1$.
If $i > 1$ and h_i is a generalized eigenvector, then $S^n h_i = 0$ as $n \to \infty$ since $|\lambda_i| < 1$ and using lemma 6.5.1.
Hence, we obtain that $S^n x \to ch$ as $n \to \infty$.
Now we will show that $c > 0$.

$$Sx = \begin{bmatrix} s_{11}x_1 + s_{12}x_2 + s_{13}x_3 + \cdots + s_{1n}x_n \\ s_{21}x_1 + s_{22}x_2 + s_{13}x_3 + \cdots + s_{2n}x_n \\ s_{31}x_1 + s_{32}x_2 + s_{33}x_3 + \cdots + s_{3n}x_n \\ \cdots \\ s_{n1}x_1 + s_{n2}x_2 + s_{n3}x_3 + \cdots + s_{nn}x_n \end{bmatrix}$$

Since $\sum_{i=1}^{n} s_{ij} = 1$, observe that $\sigma(Sx) = \sigma(x)$. This in turn implies $\sigma(S^n(x)) = \sigma(x)$.
Now, as $n \to \infty$, $\sigma(x) = \sigma(S^n x) \to \sigma(ch) = c\sigma(h)$. Since $\sigma(x) > 0$ and

$\sigma(h) > 0$ (x being non-negative non-zero and h being positive), we get $c = \frac{\sigma(x)}{\sigma(h)} > 0$. The proof is complete. $\square$

6.6 Application to positive Markov chains

We will conclude the chapter by proving the fundmental results about positive Markov chains as we said in the beginning. Recall that if $(Q, \mu_0, P = [p_{ij}])$ is a Markov system, we denote by $P^n[i,j]$, the $(i,j)^{\text{th}}$ entry in P^n. We intially define a positive Markov chain,

Definition 6.6.1 (Positive Markov system/chain). Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ and associated Markov chain $X_0, X_1, X_2, \ldots$. The Markov system (chain) is said to be positive, if $p_{ij} > 0$ for all $i, j \in Q$

We assemble the results in the following lemma,

Lemma 6.6.1. *Let $(Q, \mu_0, P = [p_{ij}])$ be a positive Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$ and associated positive Markov chain $X_0, X_1, X_2, \ldots$. The following are true,*

1. *$(Q, \mu_0, P = [p_{ij}])$ has a unique stationary distribution $p = [p_{ij}]$*

2. *$P^n \mu_0 \to p$ as $n \to \infty$*

Proof. We will prove the assertions one by one,

1. Applying Perron theorem for positive matrices (theorem 6.3.1) and lemma 6.5.2 on the transition matrix (which is positive and stochastic by definition) P, we get that P has a 1 as the dominant eigenvalue. From Perron theorem, we also conclude that 1 has a positive eigenvector. This implies that there exist $p \in \mathbb{R}^{|Q|}$ such that $Pp = p$. Without

loss of generality, we can assume that the $\sigma(p) = 1$ (see definition 6.3.1), otherwise replace p by $\frac{1}{\sigma(p)}p$. This indeed shows that p is a stationary distribution. Uniquness follows directly from assertion 4 of theorem 6.3.1

2. Applying theorem 6.5.3 to P and initial distribution μ_0, we get that $P^n\mu_0 \to cp$ as $n \to \infty$ where $c > 0$. From the proof of theorem 6.5.3, we know that $c = \frac{\sigma(\mu_0)}{\sigma(p)}$. Since p and μ_0 are probability distributions, we get that $\sigma(p) = 1$ and $\sigma(\mu_0) = 1$ and hence $c = 1$. The claim is thus proved.

$\square$

In lemma 5.2.2, we had already shown that $P^n[i, j] \to p_i$ as $n \to \infty$ is an equivalent to the second claim in lemma 6.6.1.

Chapter 7

Introduction to ergodic theory

7.1 Introduction to Ergodic Theory

By transformations we mean functions mapping a measure space to itself. Hereafter whenever we refer to a measure space $(X, \mathcal{M}, \mu)$, we assume it is a finite measure space. i.e, $\mu(X) < \infty$. We begin with defining measurable transformations.

Definition 7.1.1 (Measureable transformation). Let $T : (X, \mathcal{M}, \mu) \rightarrow (X, \mathcal{M}, \mu)$ be a transformation. T is measureable transformation if $\forall A \in \mathcal{M}$, $T^{-1}(A) \in \mathcal{M}$

Any transformation that we use in the following sections is assumed to be a measurable transformation unless specified otherwise. In the parts of ergodic theory discussed here, we primarily discuss measure preserving transformations.

Definition 7.1.2 (Measure preserving transformation). Let $T : (X, \mathcal{M}, \mu) \rightarrow (X, \mathcal{M}, \mu)$ be a measurable transformation. T is measure preserving if $\forall A \in \mathcal{M}$, $\mu(T^{-1}(A)) = \mu(A)$

On successive application of the transformation T, a point $x \in X$ gets mapped to a sequence of points in X. We define the orbit of a point $x \in X$ as $\{T^n(x)\}_{n=0}^{\infty}$ (where we assume $T^0(x) = x$). Now, we define few terms that deals with the way sets and functions from the space 'interact' with the transformation under consideration

Definition 7.1.3 (T-invariant sets and T-invariant functions). Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a transformation. $A \subseteq X$ is said to be a **T-invariant set** if $T^{-1}(A) = A$. A function $f : X \to Y$ where Y is an arbitrary set is a **T-invariant function** if $f \circ T(x) = f(x)$ for all $x \in X$

We next discuss the various levels of 'mixing' activity a transformation can perform. We are interested in Mixing transformations and Ergodic transformations. In the rest of the section, we define these notions and eventually prove that mixing transformations are ergodic transformations too.
Ergodic transformations are those measure preserving transformations in which the only T-invariant sets are of measure 0 or $\mu(X)$

Definition 7.1.4 (Ergodic transformation). Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measure preserving transformation. T is an ergodic transformation if $A \subseteq X$ and $T^{-1}(A) = A$ implies $\mu(A) = 0$ or $\mu(A) = \mu(X)$

Now we define mixing transformations,

Definition 7.1.5 (Mixing transformation). Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measure preserving transformation. T is an mixing transformation if for any $A, B \subseteq X$, $\lim_{n\to\infty} \mu(T^{-n}(A) \cap B)\mu(A) = \frac{\mu(A)\mu(B)}{\mu(X)}$

When $\mu(A) > 0$, an equivalent formulation is $\lim_{n\to\infty} \frac{\mu(T^{-n}(A)\cap B)}{\mu(A)} = \frac{\mu(B)}{\mu(X)}$
Intuitively, mixing transformations are those in the 'local propotion' of points in A that reach B after n applications of T (i.e, $\frac{\mu(T^{-n}(A)\cap B)}{\mu(A)}$) converges to

the 'global propotion' of B (i.e, $\frac{\mu(B)}{\mu(X)}$). This suggests that the transformation 'mixes' the points in a 'homogenous way'. We note that, when the underlying measure space is in fact a probability space, then the condition for mixing translates to $\lim_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A)\mu(B)$ for all $A, B \in \mathcal{M}$. Now we finally note that mixing implies ergodicity when T is any measure preserving transformation.

Lemma 7.1.1. *Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measure preserving transformation. If T is a mixing transformation then T is an ergodic transformation.*

Proof. Let $A \subseteq X$ satisfy $T^{-1}(A) = A$. It is enough to show that $\mu(A) \in \{0, \mu(X)\}$.
Since T is mixing, $\lim_{n\to\infty} \mu(T^{-n}(A) \cap A) = \frac{\mu(A)^2}{\mu(X)}$. Since $T^{-n}(A) = A$, we get $\lim_{n\to\infty} \mu(A \cap A) = \lim_{n\to\infty} \mu(A) = \mu(A) = \frac{\mu(A)^2}{\mu(X)}$. We obtain the equation $\mu(A)(\mu(A) - \mu(X)) = 0$. This precisely implies $\mu(A) \in \{0, \mu(X)\}$ $\square$

7.2 Recurrence

A measure preserving transformation is recurrent if for any positive measure set in the σ-algebra, almost all the points in it returns to the set after some applications of the transformation function. We define a recurrent transformation,

Definition 7.2.1. Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measure preserving transformation. If for all positive measure set $A \in \mathcal{M}$, there exist $N \subseteq A$, $\mu(N) = 0$ such that $\forall x \in A - N$, there exist $n(x)$ such that $T^{n(x)}(x) \in A$, then T is a **recurrent** transformation.

We note that the following is an equivalent way to define recurrence,
A measure preserving transformation T is recurrent if and only if for $A \in \mathcal{M}$

such that $\mu(A) > 0$,

$$\mu(A - \bigcup_{n=1}^{\infty} T^{-n}(A)) = 0$$

The set $\bigcup_{n=1}^{\infty} T^{-n}(A)$ is the set of all points in the space which can reach A at some point of time on application of T. Now the condition $\mu(A - \bigcup_{n=1}^{\infty} T^{-n}(A)) = 0$ precisely says that outside a null set $N = A - \bigcup_{n=1}^{\infty} T^{-n}(A)$, all remaining points in A return to A at some point of time.

Now we give another equivant formulation for recurrence,

Lemma 7.2.1. *Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measure preserving transformation.T is recurrent if and only if for all positive measure set $A \in \mathcal{M}$, there exist $n \in \mathbb{N}$ such that $\mu(T^{-n}(A) \cap A) > 0$*

Proof. Consider the forward implication. Since, $\mu(A - \bigcup_{n=1}^{\infty} T^{-n}(A)) = \mu(A - \bigcup_{n=1}^{\infty} (T^{-n}(A) \cap A)) = 0$ for any $A \in \mathcal{M}$ such that $\mu(A) > 0$, we get $\mu(\bigcup_{n=1}^{\infty} (T^{-n}(A) \cap A)) > 0$. This implies there exits $n \in \mathbb{N}$ such that $\mu((T^{-n}(A) \cap A)) > 0$.
Now we prove the backward implication. Consider the set $N = A - \bigcup_{n=1}^{\infty} T^{-n}(A)$. If $\mu(N) > 0$, there exist $m \in \mathbb{N}$ such that $\mu((T^{-m}(N) \cap N)) > 0$. This implies $(T^{-m}(N) \cap N) \neq \phi$. As we noted above $\bigcup_{n=1}^{\infty} T^{-n}(A)$ is the set of all points in the space which can reach A at some point of time on application of T. This implies $N = A - \bigcup_{n=1}^{\infty} T^{-n}(A)$ is the set of all points in A that can never return to A on any number of applications of T. But then since $N \subseteq A$, $(T^{-m}(N) \cap N) \neq \phi$ is a contradiction to the fact that N consists of points $x \in A$ that do not satisfy $T^m(x) \in A$ for any $m \in \mathbb{N}$. Hence our assumption that $\mu(N) > 0$ must be false. We get $\mu(N) = 0$. $\square$

Before moving to the main result of the section, we prove an useful measure theoretic fact

Lemma 7.2.2. *Let $(X, \mathcal{M}, \mu)$ be a measure space. If $\{A_i\}_{i=0}^{\infty} \subseteq \mathcal{M}$ such that $\mu(A_i \cap A_j) = 0$ if $i \neq j$ for all $i, j \in \mathbb{N}$, then $\mu(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mu(A_i)$.*

Given a collection of measurable sets $\{A_i\}_{i=0}^{\infty}$, we say $\{A_i\}_{i=0}^{\infty}$ is an almost pairwise disjoint collection of sets if $\mu(A_i \cap A_j) = 0$ for all $i \neq j$. We know that if A_i were pairwise disjoint sets then $\mu(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mu(A_i)$. Hence, the above lemma states that with respect to this summation, almost pairwise disjoint collections are identical to disjoint collection of sets

Proof. Consider the disjointification of $\{A_i\}_{i=0}^{\infty}$, $\{E_i\}_{i=0}^{\infty}$ such that $E_i = A_i - \bigcup_{j=0}^{i-1} A_j = A_i - \bigcup_{j=0}^{i-1}(A_i \cap A_j)$. Since $\mu(A_i \cap A_j) = 0$ for all $i \neq j$, we get $\mu(\bigcup_{j=0}^{i-1}(A_i \cap A_j)) = 0$. Hence $\mu(E_i) = \mu(A_i)$. Finally, $\mu(\bigcup_{i=0}^{\infty} A_i) = \mu(\bigcup_{i=0}^{\infty} E_i) = \sum_{i=0}^{\infty} \mu(E_i) = \sum_{i=0}^{\infty} \mu(A_i)$. $\square$

Now we prove the Poincare recurrence lemma which says that all measure preserving transformations on finite measure spaces are in fact recurrent transformations. The following proof is taken from [24].

Lemma 7.2.3 (Poincare recurrence lemma). *Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measure preserving transformation such that $\mu(X) < \infty$. Then, T is a recurrent transformation.*

Proof. Let $A \in \mathcal{M}$ be a set of positive measure. By lemma 7.2.1, it is enough to show that for some $n \in \mathbb{N}$, $\mu(T^{-n}(A) \cap A) > 0$. Let us assume the contrary, i.e, for all $n \in \mathbb{N}$, $\mu(T^{-n}(A) \cap A) = 0$.
Consider $T^{-i}(A)$ and $T^{-j}(A)$ for $i, j > 0$ and $i \neq j$. Without loss of generality, let $i > j$ and $i = j + k$. Let us consider the measure of $T^{-i}(A) \cap T^{-j}(A)$, $\mu(T^{-i}(A) \cap T^{-j}(A)) = \mu(T^{-j-k}(A) \cap T^{-j}(A)) = \mu(T^{-j}(T^{-k}(A) \cap A))$. Using the fact that T is measure preserving, $\mu(T^{-j}(T^{-k}(A) \cap A)) = \mu(T^{-k}(A) \cap A) = 0$. This proves that $\{T^{-i}(A_i)\}_{i=1}^{\infty}$ are almost pairwise disjoint sets. Hence,

using lemma 7.2.2, $\mu(\bigcup_{i=1}^{\infty} T^{-i}A) = \sum_{i=1}^{\infty} \mu(T^{-i}A) = \sum_{i=1}^{\infty} \mu(A) = \infty$, since A has positive measure. This contradicts the fact that the measure space is a finite measure space. Hence it must be the case that for some $n \in \mathbb{N}$, $\mu(T^{-n}(A) \cap A) > 0$. $\square$

Chapter 8

Markov shift transformation

In this chapter we define an appropriate transformation on the probability space for Markov chains that was set up in Chapter 2 with the intention to explore Markov chains using the toolset of ergodic theory.

8.1 Markov shift transformation

We define the Markov shift transformation on the probability space of Markov chains (which is the set of all infinite sequences consisting states of the Markov chains with an appropriate σ-algebra and probability measure defined on it).

Definition 8.1.1 (Markov shift transformation). Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$.
Define the **Markov shift transformation** $T : (\Omega, \mathcal{B}, P) \to (\Omega, \mathcal{B}, P)$ such that $T(q_0, q_1, q_2, \dots) = (q_1, q_2, q_3, \dots)$

The shift transformation does a 'left-shift' of any infinite sequence given as input to it. Now we start correlating properties of Markov chains with properties of transformations in ergodic theory developed in Chapter 7. First, we notice that the Markov shift transformation is measure preserving if and only if the underlying Markov chain is stationary (see definition 5.2.1).

We prove a useful lemma before proceeding further,

Lemma 8.1.1. $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ *be any measurable transformation on probability spaces and let* $\mathcal{M} = \sigma(\mathcal{A})$, *where* $\mathcal{A}$ *is an algebra. Then,* T *is measure preserving if and only if* $P(T^{-1}(A)) = P(A)$ *for all* $A \in \mathcal{A}$

Proof. The forward implication is trivial. Suppose $P(T^{-1}(A)) = P(A)$ for all $A \in \mathcal{A}$, we consider the collection of sets $\mathcal{C} \subseteq \mathcal{M}$ such that $P(T^{-1}(E)) = P(E)$ for any $E \in \mathcal{A}$. This contains the sets in the generating algebra $\mathcal{A}$. Suppose $E \in \mathcal{C}$, we have $P(T^{-1}(E^c)) = 1 - P(T^{-1}(E)) = 1 - P(E) = P(E^c)$. Hence $E^c \in \mathcal{C}$. Suppose $\{E_i\}_{i\in\mathbb{N}} \subseteq \mathcal{C}$. Assume that E_i's are disjoint. This means $T^{-1}(E_i)$'s are also disjoint. $P(T^{-1}(\bigcup\limits_{i\in\mathbb{N}} E_i)) = P(\bigcup\limits_{i\in\mathbb{N}} T^{-1}(E_i)) = \sum\limits_{i\in\mathbb{N}} P(T^{-1}(E_i)) = \sum\limits_{i\in\mathbb{N}} P(E_i) = P(\bigcup\limits_{i\in\mathbb{N}} E_i)$. The existance of the infinite series is guarenteed since the sets are disjoint and the space is having finite measure. This proves $\bigcup\limits_{i\in\mathbb{N}} E_i \in \mathcal{C}$. Now if E_i's are not disjoint, consider the 'disjointification' of $\{E_i\}_{i\in\mathbb{N}}$, $\{F_i\}_{i\in\mathbb{N}}$ where $F_i = E_i - \bigcup\limits_{j=0}^{i-1} E_j$. Since, $\bigcup\limits_{i\in\mathbb{N}} E_i = \bigcup\limits_{i\in\mathbb{N}} F_i$, using the previous argument on $\{F_i\}_{i\in\mathbb{N}}$, the result follows. □

Now we prove the main result of this section,

Theorem 8.1.2. *Let* $(Q, \mu_0, P = [p_{ij}])$ *be a Markov system with associated probability space* $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$. *Let* T *be the Markov shift operation. Then,* T *is measure preserving if and only if the Markov chain is stationary*

Proof. If T is measure preserving, we have $P(T^{-1}(A)) = P(A)$, or in general $P(T^{-n}(A)) = P(A)$, for any n when $A \in \mathcal{B}$. Consider any event of the form $X_0 = q_0, X_1 = q_1, \ldots X_k = q_k$. $\{x \in \Omega : X_0(x) = q_0, X_1(x) = q_1, \ldots X_k(x) = q_k\} = \pi_{k]}^{-1}(q_0, q_1, \ldots q_k)$. Now $T^{-n}(\pi_{k]}^{-1}(q_0, q_1, \ldots q_k)) = (\pi_{[n,n+k]}^{-1}(q_0, q_1, \ldots q_k))$.

Hence we get $P(\{x \in \Omega : X_n(x) = q_0, X_{n+1}(x) = q_1, \dots X_{n+k}(x) = q_k\} = P(\{x \in \Omega : X_0(x) = q_0, X_1(x) = q_1, \dots X_k(x) = q_k\}$ for any $n \in \mathbb{N}$. This condition implies stationarity due to 5.2.1.

This proves the forward implication.

We first verify $P(T^{-1}(A)) = P(A)$ for all $A \in \mathcal{A}$ where $\mathcal{A}$ is the generating algebra. We consider elementary rectangles in the algebra, i.e, sets of the form $A = \pi^{-1}_{[m,n]}(q_m, q_{m+1}, q_{m+2} \dots q_n)$. We have,

$$P(T^{-1}(A)) = P(\pi^{-1}_{[m+1,n+1]}(q_m, q_{m+1}, q_{m+2} \dots q_n))$$

By stationarity we have,

$$\begin{aligned} P(\pi^{-1}_{[m+1,n+1]}(q_m, q_{m+1}, q_{m+2} \dots q_n)) &= P(\pi^{-1}_{[0,n-m]}(q_m, q_{m+1}, q_{m+2} \dots q_n)) \\ &= P(\pi^{-1}_{[m,n]}(q_m, q_{m+1}, q_{m+2} \dots q_n)) \\ &= P(A) \end{aligned}$$

Consider any set $A \in \mathcal{A}$. We know by definition of $\mathcal{A}$, $A \in B_F$ for some $F \in \mathcal{F}_{\mathbb{N}}$. As a consequence of corollary 4.2.2, we can without loss of generality assume that $A \in B_{[m,n]}$ for some $[m, n]$. Now since any set in $B_{[m,n]}$ is a disjoint union of elementary rectangles (due to lemma 4.2.1), it follows that $P(T^{-1}(A)) = P(A)$ for all $A \in \mathcal{A}$. Now using lemma 8.1.1, we get that $P(T^{-1}(A)) = P(A)$ for all $A \in \mathcal{B}$. This proves the backward implication $\square$

8.2 Ergodicity of Markov shift transformation

We now turn into conditions for ergodicity of the Markov shift transformations. We proved in lemma 6.6.1 that in positive Markov chains $P^n[i, j] \to p_i$ as $n \to \infty$. Now through the following arguments we prove that if for a stationary Markov system $(Q, \mu_0, P = [p_{ij}])$, $P^n[i,j] \to p_i$ as $n \to \infty$ then the Markov shift transformation defined on its underlying probability space is a mixing transformation. Since mixing transformations are ergodic, we argue that under this condition, Markov shift is an ergodic transformation. We initially prove a useful measure theoretic fact. It says that any set in an σ-algebra can be approximated in measure by sets in the generating algebra,

Lemma 8.2.1. *Let $(X, \mathcal{M}, \mu)$ be a finite measure space such that $\mathcal{M} = \sigma(\mathcal{A})$ where $\mathcal{A}$ is an algebra. Then,given $\epsilon > 0$ for all $A \in \mathcal{M}$, there exist $\widetilde{A} \in \mathcal{A}$ such that $\mu(A\Delta\widetilde{A}) < \epsilon$*

Proof. Consider the collection $\mathcal{C}$ of all sets in $\mathcal{M}$ having the required property. Trivially, $\mathcal{A} \subseteq \mathcal{C}$. Let $A \in \mathcal{C}$ such that $\mu(A\Delta\widetilde{A}) < \epsilon$ where $\widetilde{A} \in \mathcal{A}$. Since $A^c\Delta\widetilde{A}^c = A\Delta\widetilde{A}$ we have $\mu(A^c\Delta\widetilde{A}^c) = \mu(A\Delta\widetilde{A})$. Since $\widetilde{A}^c \in \mathcal{A}$, we get $A^c \in \mathcal{C}$. Let $\{A_i\}_{i=0}^{\infty} \subseteq \mathcal{C}$. Assume $\{A_i\}_{i=0}^{\infty}$ is a disjoint collection of sets. There exists $\{A_i\}_{i=0}^{\infty} \subseteq \mathcal{A}$ such that for all i, $\mu(A_i\Delta\widetilde{A_i}) < \epsilon 2^{-i+2}$. Since the measure space is finite and A_i's are disjoint, there exists $n \in \mathbb{N}$, such that $\mu(\bigcup\limits_{i=n}^{\infty} A_i) = \sum\limits_{i=n}^{\infty} \mu(A_i) < \frac{\epsilon}{2}$. Notice, $\bigcup\limits_{i=0}^{\infty} A_i \Delta \bigcup\limits_{i=0}^{n-1} \widetilde{A_i} \subseteq \bigcup\limits_{i=0}^{n-1} (A_i\Delta\widetilde{A_i}) \bigcup \bigcup\limits_{i=n}^{\infty} A_i$ Now,

$$
\begin{aligned}
\mu(\bigcup_{i=0}^{\infty} A_i \Delta \bigcup_{i=0}^{n-1} \widetilde{A_i}) &\leq \sum_{i=0}^{n-1} P(A_i\Delta\widetilde{A_i}) + \sum_{i=n}^{\infty} A_i \\
&\leq \sum_{i=0}^{n-1} \epsilon 2^{i+2} + \frac{\epsilon}{2} \\
&\leq \epsilon \sum_{i=0}^{\infty} 2^{i+2} + \frac{\epsilon}{2} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= \epsilon
\end{aligned}
$$

This proves $\mathcal{C}$ is a σ-algebra and hence $\mathcal{C} = \mathcal{M}$. $\square$

Now we prove that given a measure preserving transformation over a measure space, the mixing condition holds for the sets in the generating algebra of the measure space if and only if the transformation is mixing. The following proof is taken from [26].

Theorem 8.2.2. *Let $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ be a measurable transformation and let $\mathcal{M} = \sigma(\mathcal{A})$, where $\mathcal{A}$ is an algebra. Then, T is a mixing transformation if and only if $\lim_{n\to\infty} \mu(T^{-n}(\widetilde{A}) \cap \widetilde{B}) = \mu(\widetilde{A})\mu(\widetilde{B})$ for all $A, B \in \mathcal{A}$*

Proof. The forward implication is trivial.
Consider the backward implication. Fix arbitrary $\epsilon > 0$. Let $A, B \in \mathcal{M}$, Let $\widetilde{A}, \widetilde{B}$ be sets in the generating algebra such that $\mu(A\Delta\widetilde{A}) < \frac{\epsilon}{5\mu(B)}$ and $\mu(B\Delta\widetilde{B}) < \frac{\epsilon}{5\mu(\widetilde{A})}$. Such sets in the algebra are guarenteed by lemma 8.2.1. It is enough to show that there exist $m \in \mathbb{N}$ such that $\forall n \geq m$, $|\mu(T^{-n}(A) \cap B) - \mu(A)\mu(B)| < \epsilon$.
Let $m \in \mathbb{N}$ be such that $|\mu(T^{-n}(\widetilde{A}) \cap \widetilde{B}) - \mu(\widetilde{A})\mu(\widetilde{B})| < \frac{\epsilon}{5}$ for all $n \geq m$, this is guarenteed by the hypothesis of the theorem.
Using the triangle inequality,

$$\begin{aligned}
|\mu(A)\mu(B) - \mu(T^{-n}(A) \cap B)| \leq{}& |\mu(A)\mu(B) - \mu(\widetilde{A})\mu(B)| \\
&+ |\mu(\widetilde{A})\mu(B) - \mu(\widetilde{A})\mu(\widetilde{B})| \\
&+ |\mu(\widetilde{A})\mu(\widetilde{B}) - \mu(T^{-n}(\widetilde{A}) \cap \widetilde{B})| \\
&+ |\mu(T^{-n}(\widetilde{A}) \cap \widetilde{B}) - \mu(T^{-n}(A) \cap B)|
\end{aligned}$$

Notice that,

$$\begin{aligned}
(T^{-n}(\widetilde{A}) \cap \widetilde{B}) - (T^{-n}(A) \cap B) &\subseteq (T^{-n}(\widetilde{A}) \cap \widetilde{B})\Delta(T^{-n}(A) \cap B) \\
&\subseteq (T^{-n}(\widetilde{A})\Delta T^{-n}(A)) \bigcup (\widetilde{B} \cap B)
\end{aligned}$$

Hence.

$$\begin{aligned}|\mu(T^{-n}(\widetilde{A})\cap\widetilde{B})-\mu(T^{-n}(A)\cap B)| &\leq \mu(T^{-n}(\widetilde{A})\Delta T^{-n}(A))+\mu(\widetilde{B}\cap B)\\ &\leq \mu(T^{-n}(\widetilde{A}\Delta A))+\mu(\widetilde{B}\cap B)\\ &= \mu(\widetilde{A}\Delta A)+\mu(\widetilde{B}\cap B)\\ &< \frac{\epsilon}{5}+\frac{\epsilon}{5}=\frac{2\epsilon}{5}\end{aligned}$$

where the second inequality follows due to the fact that T is measure preserving

Now,

$$\begin{aligned}|\mu(A)\mu(B)-\mu(\widetilde{A})\mu(B)| &\leq |\mu(B)||\mu(A)-\mu(\widetilde{A})|\\ &\leq \mu(B)\frac{\epsilon}{5\mu(B)}\\ &= \frac{\epsilon}{5}\end{aligned}$$

and,

$$\begin{aligned}|\mu(\widetilde{A})\mu(B)-\mu(\widetilde{A})\mu(\widetilde{B})| &\leq |\mu(\widetilde{A})||\mu(B)-\mu(\widetilde{B})|\\ &\leq |\mu(\widetilde{A})|\frac{\epsilon}{5\mu(\widetilde{A})}\\ &= \frac{\epsilon}{5}\end{aligned}$$

Hence we obtain, $|\mu(A)\mu(B)-\mu(T^{-n}(A)\cap B)|<\epsilon$ for all $n\geq m$. □

We have developed sufficient theory to prove that if $P^n[i,j]\to p_i$ for all states i,j in a stationary Markov system, then the corresponding Markov shift transformation is a mixing transformation. The proof of this theorem is based on the approach taken in [26].

Theorem 8.2.3. *Let $(Q,\mu_0,P=[p_{ij}])$ be a Markov system with associated probability space $(\Omega=Q^{\mathbb{N}},\mathcal{B},P)$. Let T be the Markov shift operation. If $P^n[i,j]\to p_i$ as $n\to\infty$ for all states i,j, then the Markov shift transformation is a mixing transformation.*

We first verify the mixing condition for elementary rectangles which are of the form $\pi^{-1}_{[m,n]}(q_m, q_{m+1}, \dots q_n)$, for some $m \leq n$

Lemma 8.2.4. *Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$. Let T be the Markov shift operation. If $P^n[i,j] \to p_i$ as $n \to \infty$ for all states i, j, for elementary rectangles of the form $A = \pi^{-1}_{[m,n]}(q_m, q_{m+1}, \dots q_n)$ and $B = \pi^{-1}_{[l,p]}(q_l, q_{l+1}, \dots q_p)$, the following holds*

$$\lim_{n\to\infty} P(T^{-n}(A) \cap B) = P(A)P(B)$$

Proof. Consider any $n > p - m$, since the Markov chain is stationary, we have $P(T^{-n}(A) \cap B) = p_{q_l} P[q_l q_{l+1} \dots q_p] P^{n-p+m}[q_m, q_p] P[q_m q_{m+1} \dots q_n]$. Since, $\lim\limits_{n\to\infty} P^{n-p+m}[q_m, q_p] = p_{q_m}$, we get

$$\begin{aligned}
\lim_{n\to\infty} P(T^{-n}(A) \cap B) &= p_{q_l} P[q_l q_{l+1} \dots q_p] p_{q_m} P[q_m q_{m+1} \dots q_n] \\
&= P(A)P(B)
\end{aligned}$$

The result is thus verified for elementary rectangles. □

Now we prove theorem 8.2.3,

Proof. From theorem 8.2.2, it is enough to verify the condition for mixing transformations on the sets in the generating algebra. Let $A, B \in \mathcal{A}$. From corollary 4.2.2, we know that $A \in B_{[m,n]}$ for some $m, n \in \mathbb{N}$. Hence we can assume that A is of the form $A = \prod_{i=0}^{\infty} A_i$ where $A_i = Q$ for all $i \in \mathbb{N} - \{m, m+1, \dots n\}$. Similarly, we can argue that $B = \prod_{i=0}^{\infty} B_i$ where $B_i = Q$ for all $i \in \mathbb{N} - \{l, l+1, \dots p\}$ for some $l, p \in \mathbb{N}$. Now using lemma 4.2.1, $A = \bigcup\limits_{i=1}^{k} E_i$ where E_i is an elementary rectangle of the form $E_i = \pi^{-1}_{[m,n]}(q_m, q_{m+1}, \dots q_n)$. Similarly, $B = \bigcup\limits_{j=1}^{h} F_j$ where F_j's are elementary rectangles of the form $F_j = \pi^{-1}_{[l,p]}(q_l, q_{l+1}, \dots q_p)$. Now, $\lim\limits_{n\to\infty} P(T^{-n}(A) \cap B) = \lim\limits_{n\to\infty} P(T^{-n}(\bigcup\limits_{i=1}^{k} E_i) \cap$

$\bigcup_{j=1}^{h} F_j) = \lim_{n\to\infty} P(\bigcup_{i=1}^{k} T^{-n}(E_i) \cap \bigcup_{j=1}^{h} F_j) = \lim_{n\to\infty} P(\bigcup_{i,j}(T^{-n}(E_i) \cap F_j))$ (where i ranges from 1 to k and j ranges from 1 to h). Since $(T^{-n}(E_{i_1}) \cap F_{j_1}) \cap (T^{-n}(E_{i_2}) \cap F_{j_2}) = \phi$ when $(i_1, j_1) \neq (i_2, j_2)$, we get,

$$\begin{aligned}
\lim_{n\to\infty} P(\bigcup_{i,j}(T^{-n}(E_i) \cap F_j)) &= \lim_{n\to\infty} \sum_{i,j} P(T^{-n}(E_i) \cap F_j) \\
&= \sum_{i,j} \lim_{n\to\infty} P(T^{-n}(E_i) \cap F_j) \\
&= \sum_{i,j} P(T^{-n}(E_i))P(F_j) \\
&= \sum_{i=1}^{k} P(T^{-n}(E_i)) \sum_{j=1}^{h} P(F_j) \\
&= P(A)P(B)
\end{aligned}$$

□

Along with lemma 6.6.1, we get the following corollary

Corollary 8.2.5. *Let $(Q, \mu_0, P = [p_{ij}])$ be a positive Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$. Let T be the Markov shift operation. Then the Markov shift transformation is a mixing transformation.*

Proof. Direct consequence of lemma 6.6.1 and theorem 8.2.3. □

Chapter 9

Birkhoff's pointwise ergodic theorem

Here we give an elementary proof of the Birkhoff's ergodic theorem for simple functions from a probability space $(\Omega, \mathcal{F}, P)$ to the real line which uses minimal measure theoretic machinery. In the following section we develop the basics of Lebesgue integration theory. Necessary results from integration theory for simple functions are developed. We however do not extend the theory to general class of functions like L^1 or L^p functions. Interested readers may find more on Lebesgue integration theory in [10] or [2].

In this study, we obtain the ergodic theorem for Markov chains as a consequence of the Birkhoff's (pointwise) ergodic theorem (see Chapter 10), a standard result in Ergodic theory. Ergodic theory initially grew out of research in statistical mechanics and later found applications in diverse fields including number theory, information theory etc. The pointwise ergodic theorem for L^1 spaces (see more in [10] or [2]) was proved by G.D.Birkhoff in [5]. A generalied version of the pointwise ergodic theorem is the Birkhoff-Khinchin ergodic theorem (statement and proofs of which can be found in

[14]).
Many different approaches to proving the Birkhoff's ergodic theorem were published after the publication of Birkhoff's rather long original paper. Kakutani and Yosida introduced the technique of obtaining the theorem using the maximal ergodic theorem in [27]. A proof using nostandard analysis was given by Kamae in [12]. A very short proof of the Birkhoff ergodic theorem was given by Garsia (this can be found in the book by Peter Walters [26]). The core idea of the proof that we give in this chapter is based on the approach taken by Katznelson and Weiss in [13].

9.1 Integration of simple functions

Let $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ be a simple function, $f = \sum\limits_{i=1}^{n} a_i \chi_{E_i}$ where $a_i \in \mathbb{R}, E_i \subseteq \Omega$. All functions used hereafter in this document are simple functions unless specified otherwise.
We define the **integral of f over** Ω to be $\int_\Omega f dP = \sum\limits_{i=1}^{n} a_i P(E_i)$.
If $A \subseteq \Omega$, then we define **integral of f over** A as $\int_A f dP = \int_\Omega f \chi_A dP = \sum\limits_{i=1}^{n} a_i P(A \cap E_i)$
Now we state and prove few basic laws of integration for simple functions

Lemma 9.1.1. *Let f, g be simple functions $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$. Then,*

1. *If $f \geq 0$ then $\int_\Omega f dP \geq 0$*
2. *$\int_\Omega (\alpha f + g) dP = \alpha \int_\Omega f dP + \int_\Omega g dP$ when $\alpha \in \mathbb{R}$*
3. *If $f \leq g$, then $\int_\Omega f dP \leq \int_\Omega g dP$*
4. *If A and B are disjoint, then $\int_{A \cup B} f dP = \int_A f dP + \int_B f dP$*

Proof. 1. If $f = \sum_{i=1}^{n} a_i \chi_{E_i}$, we have $a_i \geq 0$ for all i.Hence, $\int_\Omega f dP = \sum_{i=1}^{n} a_i P(E_i) \geq 0$

2. If $f = \sum_{i=1}^{n} a_i \chi_{E_i}$, we $\alpha f = \sum_{i=1}^{n} -\alpha a_i \chi_{E_i}$. Hence $\int_\Omega \alpha f dP = \alpha \sum_{i=1}^{n} a_i P(E_i) = \alpha \int_\Omega f dP$.
   When $f = \sum_{i=1}^{n} a_i \chi_{E_i}$ and $g = \sum_{j=1}^{m} b_j \chi_{F_j}$, we have $f + g = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \chi_{E_i \cap F_j}$. Now,

$$\begin{aligned}
\int_\Omega (f+g) dP &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) P(E_i \cap F_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i P(E_i \cap F_j) + \sum_{i=1}^{n} \sum_{j=1}^{m} b_j P(E_i \cap F_j) \\
&= \sum_{i=1}^{n} a_i P(E_i) + \sum_{j=1}^{m} b_j P(F_j) \\
&= \int_\Omega f dP + \int_\Omega g dP
\end{aligned}$$

   From the above two arguments we get, $\int_\Omega (\alpha f + g) dP = \alpha \int_\Omega f dP + \int_\Omega g dP$ when $\alpha \in \mathbb{R}$.

3. If $f \leq g$, we get $g - f \geq 0$. Now from (1) we get, $\int_\Omega (g - f) dP = \int_\Omega g dP - \int_\Omega f dP \geq 0$. The result follows

4.

$$
\begin{aligned}
\int_{A\cup B} f dP &= \int_\Omega f\chi_{A\cup B} dP \\
&= \sum_{i=1}^{n} a_i P((A\cup B)\cap E_i) \\
&= \sum_{i=1}^{n} a_i P(A\cap E_i) + P(B\cap E_i) \\
&= \sum_{i=1}^{n} a_i P(A\cap E_i) + \sum_{i=1}^{n} a_i P(B\cap E_i) \\
&= \int_\Omega f\chi_A dP + \int_\Omega f\chi_B dP \\
&= \int_A f dP + \int_B f dP
\end{aligned}
$$

$\square$

We end this section with a useful fact about measure preserving transformations,

Lemma 9.1.2. *Let* $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ *be a measure preserving transformation and* $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ *be a simple function. Then* $\int_\Omega f dP = \int_\Omega f \circ T dP$

Proof. It is enough to prove the lemma for characteristic functions. Since simple functions are linear combinations of characteristic functions, the result follows from lemma 9.1.1. Consider characteristic function χ_E of $E \subseteq \Omega$. Observe that $\chi_E \circ T = \chi_{T^{-1}(E)}$. Applying integral and using the fact that T is measure preserving, we get $\int_\Omega \chi_E \circ T dP = \int_\Omega \chi_{T^{-1}(E)} dP = P(T^{-1}(E)) = P(E) = \int_\Omega \chi_E dP$ $\square$

9.2 Maximal ergodic theorem

There are many formulations of the maximal ergodic theorem. Here we prove one of them which suffice to prove the Birkhoff's ergodic theorem for simple functions in the next section. If $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ is a measure preserving transformation. Let $f_n(x)$ denote the summation of function over the orbit of $x \in \Omega$ upto the the first n applications of T. i.e, $f_n(x) = f(x) + f(T(x)) + \cdots + f(T^{n-1}(x))$. $f_n(x)$ will be referred to as the n^{th} orbit sum of f on x. We intially prove a weak form of the maximal ergodic theorem. Proving the stronger version requires incremental effort only. The core idea of the proof is based on the approaches used in [18] and [13].

Theorem 9.2.1 (Weak maximal ergodic theorem). *Let $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ be a measure preserving transformation and $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ be a simple function. If for all $x \in \Omega$, there exist $n(x) \in \mathbb{N}$ such that $f_{n(x)}(x) \leq 0$, then $\int_\Omega f dP \leq 0$*

We will prove the result first in a stronger case. We state it below,

Lemma 9.2.2. *Let $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ be a measure preserving transformation and $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ be a simple function. If there exist $k \in \mathbb{N}$ such that $\forall x \in \Omega$, $f_{n(x)}(x) \leq 0$ for some $1 \leq n \leq k$, then $\int_\Omega f dP \leq 0$*

Proof. Consider the n^{th} orbit sum of f on x for arbitrary $x \in \Omega$.
$f_n(x) = f(x) + f(T(x)) + \cdots + f(T^{n-1}(x))$. Suppose $n \geq k$, then using the condition that holds across the space, a first few terms can be dropped to get $f_n(x) \leq f(T^{i_1}x) + f(T^{i_1+1}(x)) + \cdots + f(T^{n-1}(x))$ where $i_1 \leq k$. Now considering the RHS as an orbit sum of f over $T^{i_1}(x)$, we can find $i_2 > i_1$ (provided $n - (i_1) \geq k$) such that $f_n(x) \leq f(T^{i_2}x) + f(T^{i_2+1}(x)) + \cdots + f(T^{n-1}(x))$. The process continues till we get i_m such

that $f_n(x) \leq f(T^{i_m}x) + f(T^{i_m+1}(x)) + \cdots + f(T^{n-1}(x))$ where $n - (i_m) < k$. In general we can say that $f_n(x) = f(x) + f(T(x)) + \cdots + f(T^{n-1}(x)) \leq |f(T^{n-k-1}x)| + |f(T^{n-k}(x))| + |f(T^{n-k+1}(x))| + \cdots + |f(T^{n-1}(x))|$. Applying integral on the both sides and due to the third assertion of lemma 9.1.1,

$$\int_\Omega \big(f(x) + f(T(x)) + \cdots + f(T^{n-1}(x))\big)\, dP$$
$$\leq \int_\Omega \big(|f(T^{n-k-1}x)| + |f(T^{n-k}(x))| + |f(T^{n-k+1}(x))| + \cdots + |f(T^{n-1}(x))|\big)$$

From the second assertion of lemma 9.1.1,

$$\int_\Omega f(x)dP + \int_\Omega f(T(x))dP + \cdots + \int_\Omega f(T^{n-1}(x))dP$$
$$\leq \int_\Omega |f(T^{n-k-1}x)|dP + \int_\Omega |f(T^{n-k}(x))|dP + \int_\Omega |f(T^{n-k+1}(x))|dP +$$
$$\cdots + \int_\Omega |f(T^{n-1}(x))|dP$$

We rewrite the above,

$$\int_\Omega f(x)dP + \int_\Omega f(T(x))dP + \cdots + \int_\Omega f(T^{n-1}(x))dP$$
$$\leq \int_\Omega |f|\circ(T^{n-k-1}x)dP + \int_\Omega |f|\circ(T^{n-k}(x))dP + \int_\Omega |f|\circ(T^{n-k+1}(x))dP +$$
$$\cdots + \int_\Omega |f|\circ(T^{n-1}(x))dP$$

Using the fact that T is measure preserving and lemma 9.1.2 the above becomes,

$$\int_\Omega f(x)dP + \int_\Omega f(x)dP + \cdots + \int_\Omega f(x)dP$$
$$\leq \int_\Omega |f|(x)dP + \int_\Omega |f|(x)dP + \int_\Omega |f|(x)dP + \cdots + \int_\Omega |f|(x)dP$$

From the above, we get $n\int_\Omega f dP \leq k\int_\Omega |f| dP$. Dividing by n, $\int_\Omega f dP \leq \frac{k\int_\Omega |f| dP}{n}$. Applying limit as $n \to \infty$ we obtain the maximal ergodic theorem, $\int_\Omega f dP \leq 0$. $\square$

Now, using the above lemma, we prove the weak maximal ergodic theorem (theorem 9.2.1)

Proof. The idea is to define a sequence of functions ψ_k such that $\psi_k \to f$ and the stronger condition applied above holds for each ψ_k. We will obtain $\int_\Omega \psi_k dP \leq 0$. Then a limiting argument can prove the maximal ergodic theorem for f.

Define $\psi_k(x) = \begin{cases} f(x), & \text{if } \exists n \in \{1, 2, \dots k\} \text{ such that } f_n(x) \leq 0 \\ 0, & \text{otherwise} \end{cases}$

Consider ψ_k, we observe that $\forall x \in \Omega$ $(\psi_k)_n(x) \leq 0$ for some $1 \leq n \leq k$. The range of values each ψ_k can take is a subset of the set of values that f can take. Hence each ψ_k is a simple function. Using lemma 9.2.2, $\int_\Omega \psi_k dP \leq 0$ for all $k \in \mathbb{N}$.

Now let $f = \sum_{i=1}^{l} a_i \chi_{E_i}$. We note that $\psi_k = \sum_{i=1}^{l} a_i \chi_{E_{ki}}$ for each k where $E_{ki} = \{x : \psi_k(x) = a_i\} = \{x : f(x) = a_i \text{ and } \exists n(x) \in \{1, 2, \dots k\} \text{ such that } f_{n(x)}(x) \leq 0\}$. Notice $E_{1i} \subseteq E_{2i} \subseteq E_{3i} \subseteq \dots$ for all i. Since for all $x \in \Omega$, there exist $n(x) \in \mathbb{N}$ such that $f_{n(x)}(x) \leq 0$, we observe $\bigcup_{k=1}^{\infty} E_{ki} = E_i$. By continuity of measure from below (see lemma B.1.3), we get $P(E_{ki}) \to P(E_i)$ as $k \to \infty$. Now $\int_\Omega \psi_k dP = \sum_{i=1}^{l} a_i P(E_{ki})$. Applying limit, $\lim_{k\to\infty} \int_\Omega \psi_k dP = \lim_{k\to\infty} \sum_{i=1}^{l} a_i P(E_{ki}) = \sum_{i=1}^{l} a_i \lim_{k\to\infty} P(E_{ki}) = \sum_{i=1}^{l} a_i P(E_i) = \int_\Omega f dP$. Hence $\int_\Omega f dP \leq 0$. □

We finally prove a generalization of the above which we will employ in the proof of the pointwise ergodic theorem,

Theorem 9.2.3 (Maximal ergodic theorem). *Let $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ be a measure preserving transformation and $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ be a simple function. Let $A = \{x \in \Omega : \exists n(x) \in \mathbb{N}$ such that $f_{n(x)}(x) \leq 0\}$, then $\int_A f dP \leq 0$*

Proof. Consider $f\chi_A$. For all $x \in \Omega$, there exist some $n(x) \in \mathbb{N}$ such that $(f\chi_A)_{n(x)}(x) \leq 0$. Applying the previous theorem, we get $\int_\Omega f\chi_A dP \leq 0$. Since $\int_\Omega f\chi_A dP = \int_A f dP$, the theorem follows. □

9.3 Birkhoff's pointwise ergodic theorem

Here we prove the Birkhoff's ergodic theorem for the case when the function is a simple function. Birkhoff's pointwise ergodic theorem states that if $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ is a simple function and $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ an ergodic transformation, then the time average of the function $\sum_{i=0}^{n-1} f(T^i(x))$ converges to the space average $\int_\Omega f dP$ as $n \to \infty$.

Theorem 9.3.1 (Birkhoff's pointwise ergodic theorem). *Let* $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ *be an ergodic transformation and* $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ *be a simple function. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = \int_\Omega f dP \text{ almost everywhere}$$

Before we provide a proof, we explore some auxiliary facts. In the following discussion, $f_* = \liminf_{n\to\infty} \frac{f_n(x)}{n}$ and $f^* = \limsup_{n\to\infty} \frac{f_n(x)}{n}$

Lemma 9.3.2. *Let* $T : (\Omega, \mathcal{F}, P) \to (\Omega, \mathcal{F}, P)$ *be any transformation and* $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ *be a simple function. If* $\lim_{n\to\infty} \frac{f_n(x)}{n}$ *exists, it is a T-invariant function.*

Proof. If $\lim_{n\to\infty} \frac{f_n(x)}{n}$ exists, $\lim_{n\to\infty} \frac{f_n(x)}{n} = f_*(x) = f^*(x)$. Hence it is enough to show that f_* and f^* are invariant functions.
From its definition, $f_n(T(x)) = f_{n+1}(x) - f(x)$. Dividing by n, $\frac{f_n(T(x))}{n} = \frac{f_{n+1}(x)}{n+1}\frac{n+1}{n} - \frac{f(x)}{n}$. Applying $\liminf$ on both sides, we get, $f_*(x) = f_*(T(x))$. The proof goes in similar lines for f^*. □

Now we are in a position to prove the main result of the section - Birkhoff's pointwise ergodic theorem for simple functions. Certain arguments in the proof below are sourced from [24]

Proof. It is enough to prove that $f_* \geq \int_\Omega f dP$ almost everywhere. If this could be argued, since $-f$ is a simple function, we get $(-f)_* \geq \int_\Omega -f dP$. From properties of lim inf and lemma 9.1.1, we get $-f^* \geq -\int_\Omega f dP \implies f^* \leq \int_\Omega f dP$. Hence we get $\int_\Omega f dP \leq f_* \leq f^* \leq \int_\Omega f dP$. This proves the existance of the limit and we get, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i(x)) = \int_\Omega f dP$ almost everywhere.
Let $B = \{x \in \Omega : f_*(x) < \int_\Omega f dP\}$. It is enough to show that $P(B) = 0$. Assume otherwise. i.e, $P(B) > 0$
Now, $B = \bigcup_{q\in\mathbb{Q}} \{x \in \Omega : f_*(x) < q < \int_\Omega f dP\}$
Let $C_q = \{x \in \Omega : f_*(x) < q < \int_\Omega f dP\}$. Since $P(B) > 0$, we get $P(C_q) > 0$ for some $q \in \mathbb{Q}$.
Notice due to T-invariance of f_*, $T^{-1}(C_q) = \{x \in \Omega : f_*(T(x)) < q < \int_\Omega f dP\} = \{x \in \Omega : f_*(x) < q < \int_\Omega f dP\} = C_q$. Since C_q is a T-invariant set, condition of ergodicity implies $P(C_q) \in \{0, 1\}$. The only possibility is $P(C_q) = 1$. This means $P(\{x \in \Omega : f_*(x) < q < \int_\Omega f dP\}) = 1$. Let $N = \Omega - C_q$. Clearly, $P(N) = 0$.
Notice $C_q = \{x \in \Omega : f_*(x) < q\}$. This implies if $x \in C_q$, there exists $n(x) \in \mathbb{N}$ such that $\frac{f_{n(x)}(x)}{n} < q$ (in-fact an infinite sequence of such $n(x)$'s exist for each x). This happens if and only if $(f-q)_{n(x)}(x) < 0$. This implies if $x \in C_q$ there exists $n(x) \in \mathbb{N}$ such that $(f-q)_{n(x)}(x) \leq 0$.
Let $A = \{x \in \Omega : \exists n(x) \in \mathbb{N}$ such that $(f-q)_{n(x)}(x) \leq 0\}$. We observe that $C_q \subseteq A$ and $P(A) = P(C_q) = 1$. Applying the maximal ergodic theorem, we get $\int_A (f-q) dP \leq 0$. Since $P(A - C_q) = 0$ and $\int_A (f-q) dP = \int_{A-C_q} (f-q) dP + \int_{C_q} (f-q) dP$, using assertion 4 of lemma 9.1.1, we get $\int_{A-C_q} (f -$

$q)dP = 0$ and $\int_{C_q}(f-q)dP \leq 0$.

Also observe that $P(\Omega) = 1$ and $P(\Omega - C_q) = 0$. Using a similar argument, we get $\int_{\Omega}(f-q)dP = \int_{C_q}(f-q)dP \leq 0 \implies \int_{\Omega} fdP \leq qP(\Omega) = q$ which contradicts the choice of q

Hence our assumption that $P(B) > 0$ must be false. We obtain $P(B) = 0$ and the theorem follows. $\square$

Chapter 10

Application of ergodic theory to Markov chains

The results that were established in the previous chapters now enables us to attain a major objective of this exposition, to state and prove the ergodic theorem for Markov chains.

10.1 Ergodic theorem for Markov chains

We are interested in proving that the *propotion of time spend in a specific state* by a stationary Markov chain satisfying $p_{ij}^n \to p_i$ as $n \to \infty$ converges to p_i. We will formalize this intuition using the following function, (the below notation is taken from [19])

We have defined $N_{q,t}$ in Chapter 2 (see definition 5.2.5). $N(q,t)$ gives the number of times the Markov chain visits the state q before time t. With this intuition in mind, let us discuss the ergodic theorem for Markov chains

Theorem 10.1.1 (Ergodic theorem for Markov chains). *Let* $(Q, \mu_0, P = [p_{ij}])$ *be a Markov system with associated probability space* $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$*.*

Let $P = [p_{ij}]$ be the stationary distribution. Suppose the underlying Markov chain satisfies $p_{ij}^n \to p_i$ as $n \to \infty$ for all $i, j \in Q$, then,

$$\lim_{t\to\infty} \frac{N_{q,t}}{t} = p_q \textit{ almost everywhere}$$

We see that the theorem precisely captures the intuition of convergence of the *propotion of time spend in a specific state* as we mentioned at the beginning of the current section. Let $\chi_q : Q \mapsto \{0, 1\}$ be the characteristic function associated with some state $q \in Q$. i.e,

$$\chi_q(x) = \begin{cases} 1 & x = q \\ 0 & \text{otherwise} \end{cases}$$

We can *naturally* extend χ_q as a function mapping $\Omega = Q^{\mathbb{N}} \mapsto \{0, 1\}$ as $\chi_q(\omega_0, \omega_1, \omega_2, \omega_3, \dots) = \chi_q(\omega_0)$ for all $\omega = (\omega_0, \omega_1, \omega_2, \omega_3, \dots) \in \Omega$ (for simplicity we do not introduce new notation for the map from $\Omega = Q^{\mathbb{N}}$). In the defintion of $\chi_q : \Omega \mapsto \{0, 1\}$, the χ_q the function on LHS is the characteristic function defined earlier from $Q \mapsto \{0, 1\}$. The proof of the ergodic theorem for Markov chains is direct from Birkhoff's pointwise ergodic theorem (see 9.3.1) once we find an alternate way to express $N_{q,t}$ as we do below,

Proof. Observe that, $N_{q,t}(\omega) = \chi_q(\omega) + \chi_q T(\omega) + \chi_q T^2(\omega) + \cdots + \chi_q T^{t-1}(\omega)$, for all $\omega \in \Omega$.

Since $p_{ij}^n \to p_i$ as $n \to \infty$ for all $i, j \in Q$, from theorem 8.2.3, we get that the underlying Markov shift transformation T is a mixing transformation. Since mixing transformations are ergodic (by lemma 7.1.1) and χ_q being a simple function, we apply the Birkhoff's pointwise ergodic theorem (see 9.3.1) to get,

$$\begin{aligned}\lim_{t\to\infty} \frac{N_{q,t}}{t} &= \lim_{t\to\infty} \frac{\chi_q(x) + \chi_q T(x) + \chi_q T^2(x) + \cdots + +\chi_q T^{t-1}(x)}{n} \\ &= \int_\Omega \chi_q dP \text{ almost everywhere}\end{aligned}$$

By definition, $\int_\Omega \chi_q dP = 0 \times P((\pi_0^{-1}(q))^c) + 1 \times P(\pi_0^{-1}(q)) = P(\pi_0^{-1}(q)) = p_q$ (the last equality holds since the initial distribution is stationary). Hence we get,

$$\lim_{t\to\infty} \frac{N_{q,t}}{t} = p_q \text{ almost everywhere}$$

□

10.2 Application of Poincare Recurrence Lemma

In this section, we derive a simple result in the theory of stationary Markov chains using the Poincare recurrence lemma.

First we obtain an equivalent statement for recurrence of a transformation.

Lemma 10.2.1. *$T : (X, \mathcal{M}, \mu) \mapsto (X, \mathcal{M}, \mu)$ be a recurrent transformation if and only if $\forall A \in \mathcal{M}$ such that $\mu(A) > 0$,*

$$\mu\left(A \cap \bigcup_{n=1}^{\infty} T^{-n}(A)\right) = \mu(A)$$

Proof. Notice that $\bigcup_{n=1}^{\infty} T^{-n}(A)$ are those elements in $x \in A$ that can return to A after some number (say $n(x)$) applications of the transformation T. Consider the forward implication. The set of all points in A that can never return to A on any number of applications of T is $N = A \cap \left(\bigcup_{n=1}^{\infty} T^{-n}(A))\right)^c$. Since T is recurrent, $\mu(N) = 0$ (see definition 7.2.1). Now

$$\begin{aligned}
\mu(A) &= \mu\left(A \cap \left(\bigcup_{n=1}^{\infty} T^{-n}(A)\right)\right) + \mu\left(A \cap \left(\bigcup_{n=1}^{\infty} T^{-n}(A)\right)^c\right) \\
&= \mu\left(A \cap \left(\bigcup_{n=1}^{\infty} T^{-n}(A)\right)\right) + 0 \\
&= \mu\left(A \cap \left(\bigcup_{n=1}^{\infty} T^{-n}(A)\right)\right)
\end{aligned}$$

The forward implication is hence true.

Consider the backward implication. An equivalent statement is,

$$\mu\left(A - \left(A \cap \bigcup_{n=1}^{\infty} T^{-n}(A)\right)\right) = 0$$

This precisely says that there exist a null subset of A, i.e

$$N = A - \left(A \cap \bigcup_{n=1}^{\infty} T^{-n}(A)\right)$$

such that all elements in $x \in A - N = \left(A \cap \bigcup_{n=1}^{\infty} T^{-n}(A)\right)$ returns to A after some $n(x)$ applications of T. This is in fact, the definition of recurrence (see definition 7.2.1). Hence T is a recurrent transformation. $\square$

Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$. We will now apply the above theorem for Markov shift transformations when the set $A = \pi_0^{-1}(q)$ where $q \in Q$ is any state such that $p_q > 0$. Notice that for all $n \in \mathbb{N}, n > 0$, we have $T^{-n}(\pi_0^{-1}(q)) = \pi_n^{-1}(q)$.

Lemma 10.2.2. *Let $(Q, \mu_0, P = [p_{ij}])$ be a Markov system with associated probability space $(\Omega = Q^{\mathbb{N}}, \mathcal{B}, P)$. Let μ_0 be the stationary distribution. Let T be the associated Markov shift transformation. For any state $q \in Q$ such that $p_q > 0$,*

$$\frac{P\left(\pi_0^{-1}(q) \cap \bigcup_{n=1}^{\infty} \pi_n^{-1}(q)\right)}{p_q} = 1$$

The lemma intuitively says that if the Markov chain *starts* in the state q at time $= 0$, then *almost surely* the chain returns to state q at some point in the future along its run.

Proof. Since μ_0 is stationary, from lemma we know that T is a measure preserving transformation. From, Poincare recurrence lemma (lemma 7.2.3), we can conclude that T is recurrent. We will apply lemma 10.2.1 for T and $A = \pi_0^{-1}(q)$.

We get the following,

$$P\left(\pi_0^{-1}(q) \cap \bigcup_{n=1}^{\infty} \pi_n^{-1}(q)\right) = P\left(\pi_0^{-1}(q) \cap \bigcup_{n=1}^{\infty} T^{-n}(\pi_0^{-1}(q))\right)$$
$$= P(\pi_0^{-1}(q)) = p_q$$

Dividing by p_q (since $p_q > 0$), we get the desired equality. □

We have in-fact shown that all states with initial probability > 0 are *recurrent states* when the Markov chain is stationary. The standard definition of recurrent states is however different from the notion presented above (reader may find more on this in Chapter 6 of [2])

Chapter 11

Conclusion and future work

11.1 Conclusion

The existence theorem for the probability space underlying discrete time finite state Markov chains presented in this report was the Kolmogorov extension theorem, for which a simple proof could be obtained when the underlying spaces were finite and discrete. Thereafter, Markov chains were studied using the toolset of ergodic theory, where we were able to give an elementary proof for Birkhoff's ergodic theorem for simple functions. Using this theorem, one immediately obtains the ergodic theorem for Markov chains, the Borel's normal number theorem and the strong law of large numbers.
We summarize the essential idea used in simplifying the proofs of the major theorems.

- While formulating the probability space underlying finite state Markov chains, the component spaces of the infinite product measure are finite and hence are compact topological spaces. Using Tychonoff's theorem, the compactness carries over to the infinite product space. Compactness of the product space played a crutial role in simplifing certain arguments at the heart of the Kolmogorov extension theorem. The reader may wish to compare the proof we obtained in this restricted case with the proof in the general case (see [2]) for a clear picture of the simplifications involved.

- The observation that the ergodic theorem for Markov chains could essentially be obtained from Birkhoff's point-wise ergodic theorem for *simple functions* resulted in an elementary proof given in Chapter 9. Proofs of the theorem for more general classes of functions (see [26],[24]) requires application Lebesgue's dominated convergence and the monotone convergence theorems (see [10],[2]).

11.2 Future work

While simplifications to the core arguments leading to the Kolmogorov extension theorem for discrete and finite spaces could be achieved using the compactness of the underlying topology, the proof still had to make use of the Caratheodory extension theorem (see theorem B.5.2). We believe that further simplifcations could be made by proving a version of Kolmogorov extension theorem that is sufficient for the purposes of developing the ergodic theory for Markov chains without resorting to the use of Caratheodory extension theorem. This, in our opinion is the direction in which further investigation needs to be conducted.

Appendices

Appendix A

Tychonoff's theorem

We had given basic definitions in topology earlier in Chapter 3. Here, we prove some results that we had left unproven in Chapter 3 including the Tychonoff's theorem.

A.1 More on compactness

The Zorn's lemma will be employed in certain proofs to follow. We do not prove Zorn's Lemma here. The reader may find a proof in [11].

Lemma A.1.1 (**Zorn's Lemma**). *If* $(P, \leq)$ *is a partially ordered set in which every chain has an upperbound, then* $(P, \leq)$ *has atleast one maximal element*

Now, we define finite union property of class of sets (FUP),

Definition A.1.1. A class of subsets of set X has **finite union property (FUP)** if any finite subclass does not cover X

We had noted the following observation in Chapter 3,

Lemma A.1.2. *A topological space (X, T) is compact $\Leftrightarrow$ every class of closed sets with FIP has IP*

The following result is employed in two of the theorems that follows,

Theorem A.1.3. *Let (X, T) be a topological space. (X, T) is compact $\Leftrightarrow$ Every basic open cover has a finite subcover $\Leftrightarrow$ Any class of basic open sets with finite union property does not cover X.*

Proof. The equivalence of the second and third statements is trivial. All that is left to obtain the result is to prove that the second condition implies the compactness of the topological space.

Let $\{B_i\}_{i\in I}$ be an open base for (X, T). Assume that every basic open cover has a finite subcover. It is enough to show that any open cover has a finite subcover. Let $\{S_j\}_{j\in J}$ be an open cover for X. Observe that each $S_j = \bigcup_{k\in K(j)} D_k$ where each $D_k \in \{B_i\}_{i\in I}$ (by definition of open base).

Consider the class $D = \bigcup_{j\in J} \{D_k\}_{k\in K(j)}$ which is an basic open cover for X. Thus it has a finite subcover, $\{D_m\}_{m\in M}$ where $M \subseteq \bigcup_{j\in J} K(j)$ and M is finite. For each D_m let $\bar{S}_m \in \{S_j\}_{j\in J}$ such that $D_m \subseteq \bar{S}_m$. Now, $\{\bar{S}_m\}_{m\in M}$ is a fjnite subcover of $\{S_j\}_{j\in J}$. The result follows. □

The following theorem is insightful and thus being proved here, though a different result equivalent to the below result is used in proving the Tychonoff's Theorem.

Theorem A.1.4. *Let (X, T) be a topological space. (X, T) is compact $\Leftrightarrow$ Every subbasic open cover has a finite subcover $\Leftrightarrow$ Any class of subbasic open sets with finite union property does not cover X.*

Proof. The equivalence of the second and third statements is trivial. All that is left to obtain the result is to prove that the third condition implies the compactness of the topological space.

Let an open subbase of the topological space be $\{S_i\}_{i\in I}$. Let the open base generated by $\{S_i\}_{i\in I}$ be denoted as $\{B_j\}_{j\in J}$.

Let $\{C_k\}_{k\in K}$ be any class of basic open sets with finite union property. It is enough to prove that $\bigcup_{k\in K} C_k \neq X$.

First, the following fact is to be proved: $\{C_k\}_{k\in K}$ has a superclass $\{C_l\}_{l\in L}$ which is maximal with respect to having finite union property. i.e, any proper superclass of $\{C_l\}_{l\in L}$ does not have finite union property.

Consider the set of all superclasses of $\{C_k\}_{k\in K}$ having finite union property. Along with the set inclusion relation, this set is a poset. By Zorn's Lemma , it is enough to prove that any chain in this poset has a upper bound. Consider any chain in this poset. The union of all elements in the chain (say U) has finite union property (Any finite class of sets from the U does not cover X since these sets are together present in some element of the chain, and this element of the chain have finite union property by definition). Hence the chain has an upper bound. Applying Zorn's Lemma, the existence of a maximal superclass (w.r.t having FUP) $\{C_l\}_{l\in L}$ is proved. It is trivial that $\{C_l\}_{l\in L}$ is a maximal element even if the poset consisted of all classes of basic closed sets with FUP.

If $\bigcup\limits_{l\in L} C_l \neq X$ then we are done. Notice that each $C_l = \bigcap\limits_{m=1}^{N(l)} D_m$ where each D_m is a subbasic open set.

Now if it can be proved that for each C_l, atleast one D_m (where $m \in \{1, 2, \ldots N(l)\}$) is in $\{C_l\}_{l\in L}$ (let us denote such a set corresponding to each C_l by $D_{\bar{N}(l)}$), then being subbasic open sets (and from our initial assumption), $\{D_{\bar{N}(l)}\}_{l\in L}$ cannot cover X. Now observe, $\bigcup\limits_{l\in L} C_l \subseteq \bigcup\limits_{l\in L} D_{\bar{N}(l)}$ (since $C_l \subseteq \bigcup\limits_{l\in L} D_{\bar{N}(l)}$ for each l) hence proving the theorem.

All that is left is to argue that for each $C_l = \bigcap_{m=1}^{N(l)} D_m$, $D_m \notin \{C_l\}_{l\in L}$ $\forall m \in \{1, 2, \ldots N(l)\}$ cannot be the case. We will assume there exist $C_i, i \in L$ such that $D_m \notin \{C_l\}_{l\in L}$ $\forall m \in \{1, 2, \ldots N(i)\}$ and arrive at a contradiction. By maximality of $\{C_l\}_{l\in L}$, $\{C_l\}_{l\in L} \cup \{D_m\}$ does not have FUP for all $m \in \{1, 2, \ldots N(i)\}$. This implies the existence of finite class of basic open sets from $\{C_l\}_{l\in L}$, $\{E_p\}_{p\in\{1,2,\ldots P(m)\}}$ for each m such that $\bigcup_{p\in\{1,2,\ldots P(m)\}} E_p \bigcup S_m = X$.

Now, $\left(\bigcup_{m\in\{1,2,\ldots N(i)\}} \{E_p\}_{p\in\{1,2,\ldots P(m)\}} \bigcup \{C_i\} \right)$ being a finite subclass of $\{C_l\}_{l\in L}$ covers X. Hence obtaining a contradiction. $\square$

A.2 Tychonoff theorem

Another result (stated below) which is equivalent to theorem A.1.4 can be directly employed in proving the major result in this section. The following proofs are taken from [25].

Theorem A.2.1. *Let (X, T) be a topological space. (X, T) is compact $\Leftrightarrow$ Every subbasic open cover has a finite subcover $\Leftrightarrow$ Any class of subbasic closed sets with finite intersection property has intersection property.*

Proof. The equivalence of the second and third statements is trivial. All that is left to obtain the result is to prove that the third condition implies the compactness of the topological space.

Let a closed subbase of the topological space be $\{S_i\}_{i\in I}$. Let the closed base generated by $\{S_i\}_{i\in I}$ be denoted as $\{B_j\}_{j\in J}$.

Let $\{C_k\}_{k\in K}$ be any class of basic closed sets with FIP. It is enough to prove that $\bigcap_{k\in K} C_k \neq \phi$.

First, the following fact is to be proved: $\{C_k\}_{k\in K}$ has a superclass $\{C_l\}_{l\in L}$

which is maximal with respect to having finite intersection property. i.e, any proper superclass of $\{C_l\}_{l\in L}$ does not have finite intersection property.

Consider the set of all superclasses of $\{C_k\}_{k\in K}$ having finite intersection property. Along with the set inclusion relation, this set is a poset. By Zorn's Lemma , it is enough to prove that any chain in this poset has a upper bound. Consider any chain in this poset. The union of all elements in the chain (say U) has finite intersection property (Any finite class of sets from the U has IP since these sets are together present in some element of the chain, and this element of the chain have finite intersection property by definition). Hence the chain has an upper bound. Applying Zorn's Lemma, the existence of a maximal superclass (w.r.t having FIP) $\{C_l\}_{l\in L}$ is proved. It is trivial that $\{C_l\}_{l\in L}$ is a maximal element even if the poset consisted of all classes of basic closed sets with FIP.

If $\bigcap\limits_{l\in L} C_l \neq \phi$ then the result follows. Notice that each $C_l = \bigcup\limits_{m=1}^{N(l)} D_m$ where each D_m is a subbasic closed set.

Now if it can be proved that for each C_l, atleast one D_m (where $m \in \{1, 2, \ldots N(l)\}$) is in $\{C_l\}_{l\in L}$ (let us denote such a set corresponding to each C_l by $D_{\bar{N}(l)}$), then being subbasic closed sets (and from our initial assumption), $\left\{D_{\bar{N}(l)}\right\}_{l\in L}$ has IP. Now observe, $\bigcap\limits_{l\in L} D_{\bar{N}(l)} \subseteq \bigcap\limits_{l\in L} C_l$ (since $\bigcap\limits_{l\in L} D_{\bar{N}(l)} \subseteq C_l$ for each l) hence proving the theorem.

All that is left is to argue that for each $C_l = \bigcup\limits_{m=1}^{N(l)} D_m$, $D_m \notin \{C_l\}_{l\in L}$ $\forall m \in \{1, 2, \ldots N(l)\}$ cannot be the case. We will assume there exist $C_i, i \in L$ such that $D_m \notin \{C_l\}_{l\in L}$ $\forall m \in \{1, 2, \ldots N(i)\}$ and arrive at a contradiction. By maximality of $\{C_l\}_{l\in L}$, $\{C_l\}_{l\in L} \cup \{D_m\}$ does not have FIP for all $m \in \{1, 2, \ldots N(i)\}$. This implies the existence of finite class of basic closed sets from $\{C_l\}_{l\in L}$, $\{E_p\}_{p\in\{1,2,\ldots P(m)\}}$ for each m such that $\bigcap\limits_{p\in\{1,2,\ldots P(m)\}} E_p \bigcap S_m = \phi$.

Now, $\left(\bigcup_{m\in\{1,2,...N(i)\}} \{E_p\}_{p\in\{1,2,...P(m)\}} \bigcup \{C_i\}\right)$ being a finite subclass of $\{C_l\}_{l\in L}$ has empty intersection. Hence obtaining a contradiction. $\square$

Now we are in a position to state and prove an important result by Andrey Nikolayevich Tikhonov,

Theorem A.2.2. ***Tychonoff's Theorem:*** *Product of non-empty class of compact topological spaces is compact*

Proof. Let $\{(X_i, T_i)\}_{i\in I}$ be a non-empty class of compact topological spaces. The claim is that $\prod_{i\in I} X_i$ with the product topology is a compact topological space (let this space be denoted as X).
It is enough to show that if $\{F_j\}_{j\in J}$ is a class of subbasic closed sets in X with FIP in , then it has non-empty intersection (or IP).
Now, each $F_j = \prod_{i\in I} F_{ij}$ where each $F_{ij} = X_i$ except for one, say F_{kj} where F_{kj} is a closed set in X_k (from the construction of the product topology). Consider the sets $S_i = \{F_{ij}\}_{j\in J}$ obtained by fixing i and varying over j. Each S_i is a class of closed sets in X_i with finite intersection property (Since $\{F_j\}_{j\in J}$ has FIP). Each X_i being compact, each S_i has non-empty intersection ($\bigcap_{j\in J} F_{ij} \neq \phi$ for all i).
Observe, $\bigcap_{j\in J} F_j = \prod_{i\in I} \bigcap_{j\in J} F_{ij}$. By Axiom of Choice, $\bigcap_{j\in J} F_j$ is non-empty. The result follows. $\square$

Appendix B

Caratheodory extension theorem

B.1 Basic measure theoretic results

We had developed basic measure theoretic notions in Chapter 3. Here, we prove some results that we had left unproven in Chapter 3 including the Caratheodory extension theorem.

If $\{A_i\}_{j=0}^{\infty}$ is a collection of sets, then by the **disjointification** of $\{A_i\}_{i=0}^{\infty}$, we mean the countable collection of sets $\{F_i\}_{i=0}^{\infty}$ where $F_0 = A_0$ and $F_i = A_i - \bigcup_{j=0}^{i-1} A_j$. By definition, we have that $\{F_i\}_{i=0}^{\infty}$ is a pairwise disjoint collection of sets such that $\bigcup_{i=0}^{\infty} A_i = \bigcup_{i=0}^{\infty} F_i$

Lemma B.1.1 (Monotonicity of measure). *Let $(X, \mathcal{M}, \mu)$ be a measure space. If $A, B \in \mathcal{M}$ such that $A \subseteq B$, then $\mu(A) \leq \mu(B)$*

Proof. Using countable additivity of measure over disjoint sets, we get $\mu(B) = \mu(A) + \mu(B - A)$. The lemma follows. $\square$

Lemma B.1.2 (**Countable subadditivity of measure**). *Let* $(X, \mathcal{M}, \mu)$ *be a measure space. If* $\{A_i\}_{j=0}^{\infty}$ *is a collection of sets from* $\mathcal{M}$*,* $\mu\left(\bigcup_{i=0}^{\infty} A_i\right) \leq \sum_{i=0}^{\infty} \mu(A_i)$

Proof. Consider the disjointification of $\{A_i\}_{j=0}^{\infty}$, i.e $\{F_i\}_{i=0}^{\infty}$. We have $\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \mu\left(\bigcup_{i=0}^{\infty} F_i\right) = \sum_{i=0}^{\infty} \mu(F_i)$ (due to countable additivty over disjoint sets). By definition of the disjointification, we have $F_i \subseteq A_i$ for all i. Hence using lemma B.1.1, we get $\sum_{i=0}^{\infty} \mu(F_i) \leq \sum_{i=0}^{\infty} \mu(A_i)$. The lemma follows. □

Lemma B.1.3 (**Continuity of measure from below**). *Let* $(X, \mathcal{M}, \mu)$ *be a measure space. If* $\{A_i\}_{j=0}^{\infty}$ *is a collection of sets from* $\mathcal{M}$ *such that* $A_0 \subseteq A_1 \subseteq A_2 \subseteq \ldots$*, then* $\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \lim_{i \to \infty} \mu(A_i)$

Proof. Consider the disjointification of $\{A_i\}_{j=0}^{\infty}$, i.e $\{F_i\}_{i=0}^{\infty}$. We have $\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \mu\left(\bigcup_{i=0}^{\infty} F_i\right) = \sum_{i=0}^{\infty} \mu(F_i)$ (due to countable additivty over disjoint sets). From the definition of disjointification, we observe the following $\mu(A_n) = \sum_{i=0}^{n} \mu(F_i)$ for all $n \geq 0$. Now, $\sum_{i=0}^{\infty} \mu(F_i) = \lim_{n \to \infty} \sum_{i=0}^{n} \mu(F_i) = \lim_{n \to \infty} \mu(A_n) = \lim_{i \to \infty} \mu(A_i)$. □

Lemma B.1.4 (**Continuity of measure from above**). $(X, \mathcal{M}, \mu)$ *be a measure space. If* $\{A_n\}_{n \in \mathbb{N}}$ *be collection of sets from* $\mathcal{M}$ *such that* $A_0 \supseteq A_1 \supseteq A_2 \supseteq \ldots$ *and* $\mu(X) < \infty$*, then* $\mu\left(\bigcap_{n=0}^{\infty} A_n\right) = \lim_{n \to \infty} \mu(A_n)$

Proof. Since $\mu(X) < \infty$, for any $A \subseteq X$, we have $\mu(A) < \infty$. We have $A_0 \supseteq A_1 \supseteq A_2 \supseteq \ldots$, consider their complements, we obtain $A_0^c \subseteq A_1^c \subseteq A_2^c \subseteq \ldots$. Using continuity from below (lemma B.1.3), we get $\lim_{n \to \infty} \mu(A_i^c) = \mu\left(\bigcup_{n=0}^{\infty} A_n^c\right)$. For any A_n, we have $\mu(A_n) + \mu(A_n^c) = \mu(X)$. Since $\mu(A_n^c)$ is finite, it is meaningful to subtract $\mu(A_n^c)$ on both sides to obtain $\mu(A_n) = \mu(X) - \mu(A_n^c)$. Applying limits on both sides,

$$\begin{aligned}
\lim_{n\to\infty} \mu(A_n) &= \lim_{n\to\infty} \mu(X) - \lim_{n\to\infty} \mu(A_n^c) \\
&= \mu(X) - \lim_{n\to\infty} \mu(A_n^c) \\
&= \mu(X) - \mu\left(\bigcup_{n=0}^{\infty} A_n^c\right) \\
&= \mu\left(\left(\bigcup_{n=0}^{\infty} A_n^c\right)^c\right) \\
&= \mu\left(\bigcap_{n=0}^{\infty} A_n\right)
\end{aligned}$$

□

B.2 Caratheodory extension theorem

The very useful Caratheodory extension theorem addresses the question of meaningfully *extending* the measure defined on an algebra to the σ-algebra generated by the algebra. We will show that this is indeed possible and furthermore in the case of finite measures, this is an *unique* extension (analogous results for σ-finite measure spaces can be found in [10]). The extension is established through a series of steps. We do this in order, in different subsections below. The proof presented here is taken from [10].

We will restrict the proof to finite measure spaces. Let $(X, \mathcal{M}, \mu)$ be a finite measure space and $\mathcal{A}$ be an algebra such that $\mathcal{M} = \sigma(\mathcal{A})$. Let μ_0 be a measure defined on $\mathcal{A}$. Recall that this means,

- $\mu_0(\phi) = 0$
- If $\{A_i\}_{i=0}^{\infty}$ is a countable collection of pairwise disjoint sets from $\mathcal{A}$ and if $\bigcup_{i=0}^{\infty} A_i \in \mathcal{A}$, then $\mu_0(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mu_0(A_i)$

We intially go through the various steps to obtain an *unique extension* and at the end, we state the Caratheodory extension theorem

B.3 From a measure on $\mathcal{A}$ to an outer measure on $\mathcal{P}(X)$

We first define an outer measure.

A function $\mu^* : \mathcal{P}(X) \to [0, \infty]$ is an **outer measure** if the following holds,

- $\mu^*(\phi) = 0$
- If $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$
- For any countable collection of sets $\{A_i\}_{i=0}^{\infty}$ from $\mathcal{P}(X)$, $\mu^*(\bigcup\limits_{i=0}^{\infty} A_i) \leq \sum\limits_{i=0}^{\infty} \mu^*(A_i)$

Now, we define a set function on $\mathcal{P}(X)$ using μ_0 and argue that it is an outer measure. The motivation comes from the way the area of a region on a plane can be obtained as a infimum over all coverings of the region using rectangles. For $E \subseteq X$, define $\mu^*(E) = \inf \left\{ \sum\limits_{i=0}^{\infty} \mu_0(A_i) : A_i \in \mathcal{A}, E \subseteq \bigcup\limits_{i=0}^{\infty} A_i \right\}$

Claim B.3.1. *μ^* is an outer measure on $\mathcal{P}(X)$*

Proof. Since $X \in \mathcal{A}$, $\mu^*(E)$ is defined for all $E \in \mathcal{P}(X)$. $\mu^*(\phi) = 0$ since $\mu_0(\phi) = 0$ and $\phi \subseteq \bigcup\limits_{i=0}^{\infty} \phi = \phi$. When $A \subseteq B$, $\left\{ \sum\limits_{i=0}^{\infty} \mu_0(A_i) : A_i \in \mathcal{A}, B \subseteq \bigcup\limits_{i=0}^{\infty} A_i \right\} \subseteq \left\{ \sum\limits_{i=0}^{\infty} \mu_0(A_i) : A_i \in \mathcal{A}, A \subseteq \bigcup\limits_{i=0}^{\infty} A_i \right\}$, hence it follows that $\mu^*(A) \leq \mu^*(B)$.

To prove countable subadditivity property, consider countable collection of sets $\{A_i\}_{i=0}^{\infty}$ from $\mathcal{P}(X)$. Choose an arbitrary $\epsilon > 0$. Now, for each i, there exist $\{E_{i,j}\}_{j=0}^{\infty}$, $E_{i,j} \in \mathcal{A}$ such that $\sum\limits_{j=0}^{\infty} \mu_0(E_{i,j}) < \mu^*(A_i) + \epsilon 2^{i+1}$.

Observe that $\bigcup_{i=0}^{\infty} A_i \subseteq \bigcup_{i=0}^{\infty}\bigcup_{j=0}^{\infty} E_{i,j}$. Hence, $\mu^*(\bigcup_{i=0}^{\infty} A_i) \leq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \mu_0(E_{i,j}) \leq \sum_{i=0}^{\infty}(\mu^*(A_i) + \epsilon 2^{i+1}) \leq \sum_{i=0}^{\infty} \mu^*(A_i) + \epsilon$. Since ϵ is arbitrary, countable subadditivity follows. $\square$

Recapitulating what we did, we obtained an outer measure μ^* on $\mathcal{P}(X)$ using μ_0 defined on the algebra $\mathcal{A}$. It is a trivial observation that $\mu^*(A) = \mu_0(A)$ for all $A \in \mathcal{A}$.

B.4 From an outer measure on $\mathcal{P}(X)$ to a measure on $\mathcal{M}$

We first define a class of 'well-behaved' sets with respect to the outer measure μ^*. A set $A \subseteq X$ is defined to be a **μ^*-measurable set** (or Caratheodory measureable set) if $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c)$ for all $E \subseteq X$. The measurable sets intuitively are the ones which nicely 'paritions' every subset of X. We will denote the set of all μ^*-measurable sets as $\mathcal{M}^*$. It is to be noted that $\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c)$ is the only non-trivial inequality to be verified in order to guarentee the condition for set A to be measurable, because the reverse inequality holds from the definition of an outer measure and that fact that $E = (E \cap A) \cup (E \cap A^c)$. Now, we proceed with the next step in the process of extending μ_0. We straightaway prove the following theorem,

Lemma B.4.1. *$\mathcal{M}^*$ is a σ-algebra and μ^* is a measure on $\mathcal{M}^*$*

Proof. We prove both the claims in an interleaved sequence of arguments. Due to the inherent symmetry in the definition, it is trivial that $\mathcal{M}^*$ is closed under complementation.

If $A, B \in \mathcal{M}$, then for any $E \subseteq X$, $\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c) = \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B) + \mu^*(E \cap A^c \cap B^c)$. Since $(A \cup B) \subseteq (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$, we get that , $\mu^*(E \cap A) + \mu^*(E \cap A^c) = \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B) + \mu^*(E \cap A^c \cap B^c) \geq \mu^*(E \cap (A \cup B)) + \mu^*(E \cap (A \cup B)^c)$. This proves that $\mathcal{M}^*$ is closed under finite union.

If $A, B \in \mathcal{M}^*$ and $A \cap B = \phi$, $\mu^*(A \cup B) = \mu^*((A \cup B) \cap A) + \mu^*((A \cup B) \cap A^c)$ and hence μ^* is finitely additive on $\mathcal{M}^*$.

We will consider countable union of sets in $\mathcal{M}^*$. Let $\{A_i\}_{i=1}^{\infty}$ be a countable collection of sets from $\mathcal{M}^*$. We need to show that $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{M}^*$. We will prove the claim in the case when A_i's are pairwise disjoint (The general case directly follows by considering the disjointification of $\{A_i\}_{i=1}^{\infty}$, i.e $\{F_i\}_{i=1}^{\infty}$ and we had seen earlier that $\bigcup\limits_{i=1}^{\infty} A_i = \bigcup\limits_{i=1}^{\infty} F_i$). Let $B_n = \bigcup\limits_{i=1}^{n} A_i$ and $B = \bigcup\limits_{n=1}^{\infty} B_n = \bigcup\limits_{i=1}^{\infty} A_i$. Now for any $E \subseteq X$, $\mu^*(E \cap B_n) = \mu^*(E \cap B_n \cap A_n) + \mu^*(E \cap B_n \cap A_n^c) = \mu^*(E \cap A_n) + \mu^*(E \cap B_{n-1})$ (for $n > 0$). Now this proves that for any n, $\mu^*(E \cap B_n) = \sum\limits_{i=1}^{n} \mu^*(E \cap A_i)$. Now due to closure under finite union, for any $E \subseteq \ x$, $\mu^*(E) = \mu^*(E \cap B_n) + \mu^*(E \cap B_n^c) = \sum\limits_{i=1}^{n} \mu^*(E \cap A_i) + \mu^*(E \cap B_n^c) \geq \sum\limits_{i=1}^{n} \mu^*(E \cap A_i) + \mu^*(E \cap B^c)$ (the last inequality follows since $B_n \subseteq B$ for all $n > 0$). Taking limit as $n \to \infty$ on both sides, we obtain, $\mu^*(E) \geq \sum\limits_{i=1}^{\infty} \mu^*(E \cap A_i) + \mu^*(E \cap B^c)$. Now using countable subadditivity of outer measures, we get $\mu^*(E) \geq \mu^*(\bigcup\limits_{i=1}^{\infty}(E \cap A_i)) + \mu^*(E \cap B^c) = \mu^*(E \cap B) + \mu^*(E \cap B^c)$. This proves closure of $\mathcal{M}^*$ under countable union. Since $\mu^*(E) \geq \mu^*(\bigcup\limits_{i=1}^{\infty}(E \cap A_i)) + \mu^*(E \cap B^c) = \mu^*(E \cap B) + \mu^*(E \cap B^c) \geq \mu^*(E)$, we proved that all the above terms are in fact equal. Now, when $E = B$, we get that $\mu^*(E) = \sum\limits_{i=1}^{n} \mu^*(B \cap A_i) + \mu^*(B \cap B^c) = \sum\limits_{i=1}^{n} \mu^*(A_i)$. This proves countable additivty of μ^* on $\mathcal{M}^*$. The proof is complete. $\square$

We have the completed a major part in our process of extending μ_0. We prove some properties of the σ-algebra $\mathcal{M}^*$ that we obtained above and the measure μ^* defined on it.

Lemma B.4.2. $\mathcal{A} \subseteq \mathcal{M}^*$

Proof. Let $A \in \mathcal{M}^*$.It is enough to prove that for $E \subseteq X$, $\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c)$. Choose a collection of sets $\{B_i\}_{i=0}^{\infty}$ from $\mathcal{A}$ such that $\sum_{i=0}^{\infty} \mu_0(B_i) \leq \mu(E) + \epsilon$ for arbitrary $\epsilon > 0$. We get, $\mu(E) + \epsilon \geq \sum_{i=0}^{\infty} \mu_0(B_i) = \sum_{i=0}^{\infty} \mu_0(B_i \cap A) + \sum_{i=0}^{\infty} \mu_0(B_i \cap A^c)$ (This follows from the fact that μ_0 is countable additive on $\mathcal{A}$ and $B = (B \cap A) \cup (B \cap A^c)$). Since $E \cap A \subseteq \bigcup_{i=0}^{\infty} (B_i \cap A)$, $E \cap A^c \subseteq \bigcup_{i=0}^{\infty} (B_i \cap A^c)$ and also due to the fact that $\mathcal{A}$ is an algebra, we get $\sum_{i=0}^{\infty} \mu_0(B_i \cap A) + \sum_{i=0}^{\infty} \mu_0(B_i \cap A^c) \geq \mu_0(E \cap A) + \mu_0(E \cap A^c)$. The lemma follows since ϵ is arbitrary. □

Lemma B.4.3. $\mu_0(A) = \mu^*(A)$ *for all* $A \in \mathcal{A}$

Proof. Since $A_0 = A$ and $A_i = \phi$ for all $i > 0$ is a covering of A, we easily obtain $\mu^*(A) \leq \mu_0(A)$. Now we have to show $\mu_0(A) \leq \mu^*(A)$ when $A \in \mathcal{A}$. Let $\{B_i\}_{i=0}^{\infty}$ be any collection of sets from $\mathcal{A}$ such that $A \subseteq \bigcup_{i=0}^{\infty} B_i$. Let $B = \bigcup_{i=0}^{\infty} B_i$. Let $\{F_i\}_{i=0}^{\infty}$ be the disjointification of $\{B_i\}$. We have $A \subseteq \bigcup_{i=0}^{\infty} F_i$ and $F_i \subseteq B_i$ for all i. Define, $E_i = A \cap F_i$. It is easy to see that $\{E_i\}_{i=0}^{\infty} \subseteq \mathcal{A}$. Furthermore $\bigcup_{i=0}^{\infty} E_i = \bigcup_{i=0}^{\infty} (A \cap F_i) = A$. since E_i's are disjoint, using countable additivty of μ_0, we get $\mu(A) = \sum_{i=0}^{\infty} \mu_0(E_i)$. By monotonicity of outer measures, we get $\sum_{i=0}^{\infty} \mu_0(E_i) \leq \sum_{i=0}^{\infty} \mu_0(B_i)$ □

The above lemmas show that we are almost done with extension, only the *uniqueness* remains to be looked into. Now since $\mathcal{A} \subseteq \mathcal{M}^*$ and since $\mathcal{M}^*$

is a σ-algebra, we get that $\sigma(\mathcal{A}) = \mathcal{M} \subseteq \mathcal{M}^*$. Hence when restricted to $\mathcal{M}$, μ^* is a measure defined on it. i.e $\mu = \mu^*|\mathcal{M}$ is a measure on $\mathcal{M}$.

B.5 Uniqueness of the extension

Before summarizing the results of our effort till now, we prove that if ν is a measure on $\mathcal{M}$ such that $\nu(A) = \mu_0(A)$ for all $A \in \mathcal{A}$, then in fact, $\nu(E) = \mu(E)$ for all $E \in \mathcal{M}$. In this sense, μ is an unique extension of μ_0 from $\mathcal{A}$ to $\mathcal{M}$

Lemma B.5.1. *If ν is a measure on $\mathcal{M}$ such that $\nu(A) = \mu_0(A)$ for all $A \in \mathcal{A}$, then $\nu(E) = \mu(E)$ for all $E \in \mathcal{M}$*

Proof. Choose sets $\{A_i\}_{i=0}^{\infty}$ from $\mathcal{A}$, such that $E \subseteq \bigcup\limits_{i=0}^{\infty} A_i$. Since ν is a measure, we have $\nu(E) \leq \sum\limits_{i=0}^{\infty} \nu(A_i)$. Since ν and μ coincide on $\mathcal{A}$, we get $\sum\limits_{i=0}^{\infty} \nu(A_i) = \sum\limits_{i=0}^{\infty} \mu_0(A_i)$. From the definiton of $\mu = \mu^*|\mathcal{M}$, we get that $\nu(E) \leq \mu(E)$.

If $\{A_i\}_{i=0}^{\infty}$ is a collection of sets from $\mathcal{A}$ and $A = \bigcup\limits_{i=0}^{\infty} A_i$, then using continuity of measure from below, we get $\nu(A) = \lim_{n\to\infty} \nu(\bigcup\limits_{i=0}^{n} A_i)$. Since μ and ν coincide on $\mathcal{A}$, we get that $\lim_{n\to\infty} \nu(\bigcup\limits_{i=0}^{n} A_i) = \lim_{n\to\infty} \mu(\bigcup\limits_{i=0}^{n} A_i) = \mu(A)$ (The final equality again follows due to continuity of measure from below).

Employing the above observation, we prove the reverse inequality, i.e $\mu(E) \leq \nu(E)$ for all $E \in \mathcal{M}$. Choose $\{A_i\}_{i=0}^{\infty}$ from $\mathcal{A}$, such that $E \subseteq A$ and $\mu(A) \leq \mu(E) + \epsilon$ for $\epsilon > 0$ (where $A = \bigcup\limits_{i=0}^{\infty} A_i$). From the previous observation, we have $\mu(A) = \nu(A)$. Now finally, $\mu(E) \leq \mu(A) = \nu(A) = \nu(E) + \nu(A - E) \leq \nu(E) + \mu(A - E) \leq \nu(E) + \epsilon$ (where we have used $\nu(E) \leq \mu(E)$). Since ϵ is arbitrary we get $\mu(E) \leq \nu(E)$.

Hence we conclude that $\mu(E) = \nu(E)$ for all $E \in \mathcal{A}$. $\square$

We have all the work necessary to prove the Caratheodory extension theorem. The proof of the theorem is immediate from the lemmas above.

Theorem B.5.2 (Caratheodry extension theorem). *Let μ_0 be a finite measure defined on an algebra $\mathcal{A}$ and $\mathcal{M} = \sigma(\mathcal{A})$. Then, there exist a unique measure μ on $\mathcal{M}$ such that $\mu(A) = \mu_0(A)$ for all $A \in \mathcal{A}$*

Proof. Theorem follows from the lemmas above □

Appendix C

Modes of convergence

Here we introduce various modes of convergence of functions from probability spaces to real numbers.

Let $f_n : (\Omega, \mathcal{F}, P) \to \mathbb{R}$ where $n \in \mathbb{N}$ be a sequence of real valued functions on a probability space and $f : (\Omega, \mathcal{F}, P) \to \mathbb{R}$. The following are the different modes of convergence (in decreasing order of strength, the reader is expected to verify that any mode of convergence implies the ones below it)

1. We say f_n **converges uniformly** to f if, for all $\epsilon > 0$, there exist $m \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ if $n \geq m$ for all $x \in \Omega$.

2. f_n **converges pointwise** to f if, for all $x \in \Omega$ and $\epsilon > 0$, there exist $m_x \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \epsilon$ for all $n \geq m_x$.

3. f_n **converges in probability** to f if for any $\epsilon > 0$ and $\delta > 0$, there exist $m \in \mathbb{N}$ such that $P(\{x \in \Omega : |f_n(x) - f(x)| \geq \epsilon) < \delta$ for all $n \geq m$. Expressed more concisely, f_n converges in probability to f if for any $\epsilon > 0$, $\lim_{n\to\infty} P(\{x \in \Omega : |f_n(x) - f(x)| \geq \epsilon) = 0$.

Appendix D

Heine Borel Theorem

D.1 Heine Borel Theorem for $\mathbb{R}^n$

It is trivial to prove that compact subsets of $\mathbb{R}^n$ are closed and bounded (in the usual Euclidian metric). Heine Borel theorem provides the converse and hence we observe that compact subsets of $\mathbb{R}^n$ are exactly the closed and bounded subsets of $\mathbb{R}^n$. We prove it intially for $\mathbb{R}^1$ and use the Tychonoff's theorem to obtain the generalized version in $\mathbb{R}^n$. The proof below is taken from [25].

Theorem D.1.1 (**Heine Borel theorem in** $\mathbb{R}$)**.** *Let $S \subseteq \mathbb{R}$. Then, the following are equivalent,*

- *S is compact*
- *S is closed and bounded*

Proof. The forward implication is trivial.
We will prove the backward implication. If S is closed and bounded, then $S \subseteq [a, b]$ for some $a, b \in \mathbb{R}$. Since closed subsets of compact sets are compact, it is enough to show that $[a, b]$ is compact for all $a, b \in \mathbb{R}$. The set $\mathcal{C} =$

$\{[a,c]|c \in (a,b)\} \cup \{[d,b]|d \in (a,b)\}$ is a closed subbase for the relative topology on $[a,b]$ as a subset of $\mathbb{R}$. It is enough to show that any non-empty $\mathcal{A} \subseteq \mathcal{C}$ with finite intersection property has non-empty intersection. Now the compactness follows by theorem A.2.1. If $\mathcal{A}$ consists only of sets of the form $[a,c]$, then the conclusion holds since $a \in \cap_{A \in \mathcal{A}} A$. Similarly, the intersection is non-empty if all sets are of the form $[d,b]$. Suppose $\mathcal{A}$ has both types of sets. Let $\widetilde{d} = sup\{d|[d,b] \in \mathcal{A}\}$. It is enough to show that $\widetilde{d}$ belongs to all sets of the form $[a,c]$, since it will then follow that $\widetilde{d} \in \cap_{A \in \mathcal{A}} A$. Suppose there exits $[a,c]$ in $\mathcal{A}$ such that $\widetilde{d} \geq c$. Then by definition of $\widetilde{d}$, there exits $[d,b] \in \mathcal{A}$ such that $d \geq c$. But then $\{[a,c],[d,b]\}$ is a finite set such that $[a,c] \cap [d,b] = \phi$. This contradicts the finite intersection property of $\mathcal{A}$ and we obtain $\widetilde{d} \in \cap_{A \in \mathcal{A}} A \neq \phi$. □

Theorem D.1.2 (Heine Borel theorem in $\mathbb{R}^n$). *Let $S \subseteq \mathbb{R}^n$. Then, the following are equivalent,*

- *S is compact*
- *S is closed and bounded*

Proof. The forward implication is trivial. Since S is closed and bounded S is the subset of an appropriate n-dimensional rectangle $\prod_{i=1}^{n}[a_i,b_i]$. Since each $[a_i,b_i]$ is a compact subspace of $\mathbb{R}$ by D.1.1, using Tychonoff's theorem we get that $\prod_{i=1}^{n}[a_i,b_i]$ is a compact topological space where the product is the taken over the relative topologies in each $[a_i,b_i]$ (as a subset of $\mathbb{R}$). It remains to show that the product of relative topologies of each $[a_i,b_i]$ is the same as the relative topology on $\prod_{i=1}^{n}[a_i,b_i]$ as a subset of $\mathbb{R}^n$. This is done through a series of steps. The sequence of steps are given below. The verification of these are elementary and hence left to the reader,

1. Inside any n-dimensional open rectangle of the form $\prod_{i=1}^{n}(a_i, b_i)$, there is a n-dimensional open sphere (set of the form $\{x \in \mathbb{R}^n | d(x, c) < r\}$ where $c \in \mathbb{R}^n$, $r \in \mathbb{R}$ and d is the Euclidian metric in $\mathbb{R}^n$) totally contained inside the open rectangle. And conversely inside any n-dimensional open sphere, there is an n-dimensional open rectangle totally contained inside it.

2. The usual euclidian metric in $\mathbb{R}^n$ is generated by the set of all n-dimensional open spheres. And thus the above proves that the metric topology on $\mathbb{R}^n$ can be also generated by the set of all n-dimensional open rectangles in $\mathbb{R}^n$

3. The above shows that sets of the form,
$\{O_1 \times O_2 \times O_3 \times \dots O_n | O_i$ is a set of the form $[a_i, c)$ or (c, d) or $(d, b_i]\}$ is an open base for the relative topology on $\prod_{i=1}^{n}[a_i, b_i]$ as a subset of $\mathbb{R}^n$.

4. The set,
$\{O_1 \times O_2 \times O_3 \times \dots O_n | O_i$ is a set of the form $[a_i, c)$ or (c, d) or $(d, b_i]\}$ is an open base for the product topology also.

Since these two topologies on the same set $\prod_{i=1}^{n}[a_i, b_i]$ has the same open base, we can conclude that product of relative topologies of each $[a_i, b_i]$ is the same as the relative topology on $\prod_{i=1}^{n}[a_i, b_i]$ as a subset of $\mathbb{R}^n$.

$\square$

Appendix E

Nedoma's pathology

E.1 Nedoma's pathology

In theorem 3.2.9, it was proved that when the underlying spaces are second countable, then the product sigma algebra of the individual Borel sigma algebras $\bigotimes_{i \in I} \mathcal{B}_{X_i}$ coincides with the Borel sigma algebra $\mathcal{B}_X$ generated by the product topology where $X = \prod_{i \in I} X_i$. This is not true in general. We will provide a counter example in this section called Nedoma's Pathology. The following proof is based on the blog post in [16]

Before looking at the main argument we consider some simple properties of product of Hausdorff Spaces. Recall that a topological space $(X, \mathcal{T})$ is Hausdorff when every pair of points has disjoint open neighbourhoods. Consider the product $X \times X$ of any set X with itself. We define the **diagonal set** as $\Delta = \{(x, x) | x \in X\}$

Lemma E.1.1. *Let $(X, \mathcal{T})$ be a Hausdorff space. Then the diagonal set is closed in the product topology.*

Proof. It is enough to show that $X \times X - \Delta$ is open. Consider any $(x, y) \in \Delta^c$. We have $x \neq y$. Since X is Hausdorff, we have open sets A and B in $\mathcal{T}$ such that $x \in A, y \in B$ and $A \cap B = \phi$. Now $A \times B$ is an open set in the product topology and $A \times B \cap \Delta = \phi$ since A and B are disjoint. Since choice of (x, y) was arbitrary, we showed that any point in Δ^c has a open neighbourhood totally contained in Δ^c. The lemma follows. □

This shows that Δ is a measurable set in Borel sigma algebra generated by the product topology $\mathcal{B}_X$.
Now suppose Δ is expressable as the union of sets of the form $A \times B$ where $A, B \in X$,i.e $\Delta = \bigcup_{i \in I} A_i \times B_i$ where $A_i, B_i \in X$ then it is easy to see that each A_i, B_i must be singletons and hence cardinality of I should greater than or equal to that of X. Keeping the two points above in mind we state the following fact

Theorem E.1.2. *If S is a measurable set in the product sigma algebra $\mathcal{B}_X \bigotimes \mathcal{B}_X$, then S is expressable as the union of $|\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})|$ sets of the form $A \times B$ where $A, B \in X$.*

Now it should be clear how a counterexample can be produced. We consider Hausdorff space $(X, \mathcal{T})$ such that $|X| > |\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})|$. Now the diagonal set is a measurable set in $\mathcal{B}_X$. But now, Δ cannot be expressed as the union of $|\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})|$ sets of the form $A \times B$ where $A, B \in X$. Hence Δ is not measurable in $\mathcal{B}_X \bigotimes \mathcal{B}_X$. Hence we get $\mathcal{B}_X \bigotimes \mathcal{B}_X \subset \mathcal{B}_X$. We conclude this section with the proof of the above theorem.

Lemma E.1.3. *If S is a measurable set in the product sigma algebra $\mathcal{B}_X \bigotimes \mathcal{B}_X$, then there exits a countable collection of sets $\{A_i\}_{i \in \mathbb{N}}$ such that $S \in \sigma(\{A_i \times A_j | i, j \in \mathbb{N}\})$*

Proof. Consider the collection $\mathcal{C}$ of all sets in $\mathcal{B}_X \bigotimes \mathcal{B}_X$ having the property required by the lemma. $\mathcal{C}$ trivially contains all the measurable rectangles

$\{A \times B | A, B \in \mathcal{B}_X\}$. If $D \in \mathcal{C}$ then $D^c \in \mathcal{C}$ since the same countable collection that lets D into $\mathcal{C}$ lets D^c too into $\mathcal{C}$. If $\{D_i\}_{i \in \mathbb{N}} \subseteq \mathcal{C}$, then by taking the union of all the countable collections of cross product sets which lets each D_i get into $\mathcal{C}$ will again give a countable collection $\{A_i\}_{i \in \mathbb{N}}$ such that $\bigcup_{i \in \mathbb{N}} D_i \in \sigma(\{A_i \times A_j | i, j \in \mathbb{N}\})$. We have $\mathcal{C} \subseteq \mathcal{B}_X \bigotimes \mathcal{B}_X$ and now we obtain $\mathcal{C} = \mathcal{B}_X \bigotimes \mathcal{B}_X$. The lemma follows $\square$

Now we give the proof of the theorem above,

Proof. Given $S \in \mathcal{B}_X \bigotimes \mathcal{B}_X$, let $\{A_i\}_{i \in \mathbb{N}}$ be countable collection guarenteed by the preceding lemma such that $S \in \sigma(\{A_i \times A_j | i, j \in \mathbb{N}\})$.
Consider the set of all infinite length binary strings $\{0, 1\}^{\mathbb{N}}$. Let $x = (x_1, x2, x_3, \dots) \in \{0, 1\}^{\mathbb{N}}$. Define $B_x = \bigcap_{i : x_i = 1} A_i \bigcap_{i : x_i = 0} A_i^c$. Observe that $\{B_x : x \in \{0, 1\}^{\mathbb{N}}\}$ is a partition of X.
We now argue that any $A \in \sigma(\{A_i \times A_j | i, j \in \mathbb{N}\})$ can be written as an arbitrary union of sets of the form $B_x \times B_y$ where $x, y \in \{0, 1\}^{\mathbb{N}}$. Consider the collection $\mathcal{C}$ of all sets in $\sigma(\{A_i \times A_j | i, j \in \mathbb{N}\})$ that has the required property.
Any A_i can be written as arbitrary union of sets B_x where $x \in \{0, 1\}^{\mathbb{N}}$. This is easily observed by taking the union of sets B_x over all binary strings with the the i^{th} position set to 1. Since union commutes over cartesian product we observe that every $A_i \times A_j$ can be written as an arbitrary union of sets of the form $B_x \times B_y$ where $x, y \in \{0, 1\}^{\mathbb{N}}$. Let $A = \bigcup_{(x,y) \in J} B_x \times B_y$ where $J \subseteq \{0, 1\}^{\mathbb{N}} \times \{0, 1\}^{\mathbb{N}}$. Since sets of the form $B_x \times B_y$ forms a partition of $X \times X$, $A^c = \bigcup_{(x,y) \in J^c} B_x \times B_y$. Finally, it is trivial to observe that if each set in $\{A_i\}_{i \in \mathbb{N}}$ can be written as arbitrary union of sets of the form $B_x \times B_y$, so can be $\bigcup_{i \in \mathbb{N}} A_i$. Hence $\mathcal{C}$ is a $\sigma-$algebra containing $\{A_i \times A_j | i, j \in \mathbb{N}\}$ and hence is equal to $\sigma(\{A_i \times A_j | i, j \in \mathbb{N}\})$.

Hence, S can be written as an arbitrary union of sets of the form $B_x \times B_y$ where $x, y \in \{0,1\}^{\mathbb{N}}$. Since there are atmost $|\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})|$ sets of the form $B_x \times B_y$, the theorem follows. $\square$

The reader should note that the above is infact true if $|\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})|$ is replaced with $|\mathcal{P}(\mathbb{N})|$. But to prove that there are atmost $|\mathcal{P}(\mathbb{N})|$ sets of the form $B_x \times B_y$, requires some set-theoretic arguments which we do not intend to explore here.

Appendix F

Spectral theorem

Here we will establish an important result in Linear Algebra, that is the spectral decomposition theorem. The necessary notions including that of Generalized eigenvectors were developed in section 6.4. Recall that when $A_{n\times n}$ is a complex matrix. $\lambda \in \mathbb{C}$ is said to be a generalized eigenvalue of A if there exist a positive $m \in \mathbb{N}$ and $x \in \mathbb{C}^n$ such that $(A - \lambda I)^m x = 0$. x is said to be a generalized eigenvector corresponding to generalized eigenvalue λ. Also, λ is said to be a m^{th} order eigenvalue and x, a m^{th} order eigenvector of A.

F.1 Proof of the spectral theorem

Here, we restate the spectral theorem and directly proceed to a proof,

Theorem F.1.1 (**Spectral Theorem**). *Let $A_{n\times n}$ be a complex matrix. Every vector $x \in \mathbb{C}^n$ has a decomposition into generalized eigenvectors of A. i.e, $x = x_1 + x_2 + \cdots + x_m$ (for some $m \in \mathbb{N}, m \leq n$) where x_i are generalized eigenvectors of A.*

The core of the proof that we will do here (the proof is taken from [15]) is the below linear algebraic fact. Given a complex polynomial p, we will denote the null space of $p(A)$ by N_p. Also, $\oplus$ denotes the direct sum of vector spaces.

Lemma F.1.2. *Let $p_1, p_2 \dots p_m$ be polynomials with complex coefficients. If all p_i and p_j , $i \neq j$ does not have common roots, then, $N_{p_1 p_2 \dots p_n} = N_{p_1} \oplus N_{p_2} \dots \oplus N_{p_m}$.*

Proof. We will prove the theorem when $m = 2$. The validity of the statement for $m = 2$ implies $N_{p_1 p_2 \dots p_n} = N_{p_1} \oplus N_{p_2 p_3 p_4 \dots p_m}$. Now the general case follows by simple induction.
Since p_1 and p_2 have no common root, we get that $gcd(p_1, p_2) = 1$ (or any $c \in \mathbb{C}$). This implies the existance of complex polynomials r and s such that $p_1 r + p_2 s = 1$.
Applying A to the polynomials, we get $p_1(A)r(A) + p_2(A)s(A) = I$.
We will initially prove that any $x \in N_{p_1 p_2}$ has a decomposition of the desired kind. We will address the uniqueness in a moment.
Applying $p_1(A)r(A)+p_2(A)s(A) = I$ to x, we get $p_1(A)r(A)x+p_2(A)s(A)x = Ix = x$. Since $x \in N_{p_1 p_2}$, we observe that $p_1(A)r(A) \in N_{p_2}$ and $p_2(A)s(A) \in N_{p_1}$. The existance is proved.
Let us suppose there exist two decompositions of the desired kind for $x \in N_{p_1 p_2}$. i.e, $x = x_a + x_b = x_{a'} + x_{b'}$ (where $x_a, x_{a'} \in N_{p_1}$ and $x_b, x_{b'} \in N_{p_2}$). Now $y = x_a - x_{a'} = x_{b'} - x_b \in N_{p_1} \cap N_{p_2}$.
Let us apply $p_1(A)r(A) + p_2(A)s(A) = I$ onto y. We get $p_1(A)r(A)y + p_2(A)s(A)y = 0 + 0 = 0 = Iy = y$. This implies $x_a = x_{a'}$ and $x_b = x_{b'}$. The uniqueness of the decomposition is hence proved. □

Now, we prove theorem F.1.1

Proof. Let $x \in \mathbb{C}^n$. Since the $n+1$ vectors $x, Ax, A^2x, \dots A^n x$ are linearly dependent, there exist a polynomial p of degree $\leq n$ such that $p(A)x = 0$. Let $p(A) = \prod_{i=1}^{m}(A - r_i I)^{s_i}$ where m is the number of distinct roots of $p(A)$. Since x in the null space of $\prod_{i=1}^{m}(x - r_i I)^{s_i}$, using lemma F.1.2, we get that $x = x_1 + x_2 + \cdots + x_m$ where each x_i is in the null space of $(x - r_i I)^{s_i}$, equivalently $(A - r_i I)^{s_i}(x_i) = 0$ for all i. This shows that x is expressable as a sum of generalized eigenvectors of A. The theorem follows. $\square$

The reader may notice that the proof of the spectral theorem does not make use of the uniqueness of decomposition as guarenteed by lemma F.1.2.

References

[1] Robert B Ash and Catherine Doleans-Dade. *Probability and measure theory*. Academic Press, 2000.

[2] Siva Athreya and Viakalathur Shankar Sunder. *Measure & probability*. Universities Press, 2008.

[3] Sheldon Axler. *Linear algebra done right*, volume 2. Springer, 1997.

[4] Jordan Bell. Lecture note on kolmogorov extension theorem - http://individual.utoronto.ca/jordanbell/notes/kolmogorov.pdf, June 2014.

[5] George D Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[6] CHAD CASAROTTO. Markov chains and the ergodic theorem. *University of Chicago*, 8, 2007.

[7] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. Introduction to algorithms second edition, 2001.

[8] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[9] William Feller. An introduction to probability theory and its applications: Volume 1. 1968.

[10] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013.

[11] Paul R Halmos. *Naive set theory*. Courier Dover Publications, 2017.

[12] Teturo Kamae. A simple proof of the ergodic theorem using nonstandard analysis. *Israel Journal of Mathematics*, 42(4):284–290, 1982.

[13] Yitzhak Katznelson and Benjamin Weiss. A simple proof of some ergodic theorems. *Israel Journal of Mathematics*, 42(4):291–296, 1982.

[14] Andrei Nikolaevich Kolmogorov. A simplified proof of the birkhoff–khinchin ergodic theorem. *Uspekhi Matematicheskikh Nauk*, (5):52–56, 1938.

[15] Peter D Lax. Linear algebra. pure and applied mathematics. 1996.

[16] David R. MacIver. Nedoma's pathology - https://www.drmaciver.com/2006/04/journal-of-obscure-results-1-nedomas-pathology/, June 2006.

[17] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[18] John W Milner. Lecture note on birkhoff ergodic theorem - http://www.math.stonybrook.edu/ jack/dynotes/dn9.pdf.

[19] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Chapman & Hall/CRC, 2010.

[20] James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[21] W Richard. Hamming. coding and information theory, 1986.

[22] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

[23] Amir Said. Introduction to arithmetic coding-theory and practice. *Hewlett Packard Laboratories Report*, pages 1057–7149, 2004.

[24] César Ernesto Silva. *Invitation to ergodic theory*, volume 42. American Mathematical Soc., 2008.

[25] George F Simmons. *Introduction to topology and modern analysis*. Tokyo, 1963.

[26] Peter Walters. *An introduction to ergodic theory*, volume 79. Springer Science & Business Media, 2000.

[27] Kôsaku Yosida and Shizuo Kakutani. Birkhoff's ergodic theorem and the maximal ergodic theorem. *Proceedings of the Imperial Academy*, 15(6):165–168, 1939.

[28] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, 24(5):530–536, 1978.

Index

N

O

P

R

S

T