The Ergodic Theorem and Markov Chain Strong Laws

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Yan Huang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF MATHEMATICS

John Baxter

August 2015

Acknowledgements

I indebted to Professor John Baxter for his consistent help and sharp insights on this project. I also would like to thank Professor Lawrence Gray and Professor Glen Meeden for their precious time in serving as my committee members. I also want to thank my family for their unconditionally support, both financially and spiritually. However, all errors are on my own.

Dedication

I dedicate this thesis to my parents Zhiping and Juan. I hope that this achievement will complete the dream that you had for me all those many years ago when you chose to give me the best education you could.

Abstract

The purpose of this paper is to explain the pointwise Ergodic Theorem and then to apply it to stationary Markov Chains. The Ergodic Theorem is a theorem which shows that the time-averages of a stationary sequence of random variables converge almost surely, and also gives a way to evaluate the limit of these averages. In the setting of Markov chains, the Ergodic Theorem can be used to obtain an important convergence fact about Markov chains.

Table of Contents

The purpose of this paper is to explain the pointwise Ergodic Theorem and then to apply it to stationary Markov Chains. The Ergodic Theorem is a theorem which shows that the time-averages of a stationary sequence of random variables (defined below) converge almost surely, and also gives a way to evaluate the limit of these averages. In the setting of Markov chains, the Ergodic Theorem can be used to obtain an important convergence fact about Markov chains. We will define Markov chains later, in Section 5. The chains we consider will be sequences of random variables $X_0, X_1, \ldots$ taking values in a countable state space S, and having a transition function p. We usually assume that p is irreducible, that is, such that the chain can move from any point to any other point with positive probability. In this situation, it will sometimes happen that all the random variables have the same distribution π. In this case p is said to have stationary distribution π. In his textbook on Markov Chains (Durrett, 2012, Section 1.4), Durrett states the following theorem.

Strong Law for Markov Chains: Let S be a countable set, and let $X_0, X_1, \ldots$ be a sequence of S-valued random variables on a probability space $(\Omega, \mathfrak{F}, P)$, which is a Markov chain having transition function p. Suppose p is irreducible and has stationary distribution π. Let r be a real-valued function on S, where $r(x)$ can be interpreted as a reward that we earn in state x. Suppose that $\sum \pi(y)|r(y)| < \infty$. Then as $n \to \infty$,

$$\frac{1}{n}\sum_{k=0}^{n-1} r(X_k) \to \sum_{x} \pi(x) r(x)$$

almost surely. This theorem is used as one of the major tools in Durrett's textbook for solving problems and examples. A proof is given in his book using return times and the

Strong Law of Large Numbers. In my paper, I will show how to derive his theorem in a different way as an application of the Ergodic Theorem.

1 Stationary Stochastic Processes

Let G be any set, and let $\mathfrak{G}$ be a σ-algebra of subsets of G, so that $(G, \mathfrak{G})$ is a measurable space.

Let $(\Omega, \mathfrak{F}, P)$ be a probability space, and for each n let X_n be a measurable map from Ω to G, so that X_n is a G-valued random variable.

In this paper,

either (i) the space G will be the real line $\mathbb{R}$, and $\mathfrak{G}$ will be the Borel sets $\mathfrak{B}$,

or (ii) the space G will be a countable set S which is the state space of Markov Chain, and $\mathfrak{G}$ will be the collection of all subsets of S.

The key idea in our discussion of convergence is the concept of a stationary sequence.

Definition 1 (Stationary sequences)

A stochastic process $(X_0, X_1, \ldots)$ is strictly stationary if for each fixed non-negative integer k the distribution of the random vector $(X_n, X_{n+1}, \ldots, X_{n+k})$ has the same distribution for all non-negative integers n.

Although the stationary property is easy to check, when we actually give the proof of convergence in the Ergodic Theorem we will also need to work with the concept of a measure-preserving transformation.

A measurable transformation T on a probability space $(\Omega, \mathfrak{F}, P)$ is called *measure preserving* if T is a measurable map from Ω to Ω, and

$$P\left(T^{-1}(A)\right) = P(A)$$

for all $A \in \mathfrak{F}$.

Definition 2 (A sequence generated with a measure-preserving transformation)

For any measure-preserving transformation T, a sequence $X_0, X_1, \ldots$ of random variables will be said to be generated by T if

$$X_n = X_0 \circ T^n, \qquad n = 0,1,2 \ldots,$$

The concepts described in Definition 1 and Definition 2 are closely connected. Any sequence generated with a measure-preserving transformation is stationary, and, conversely, any stationary sequence has a representation in terms of a sequence generated with a measure-preserving transformation. We will prove these statements in two steps.

From Definition 2 to Definition 1

Let T be a measure-preserving transformation on some probability space $(\Omega, \mathfrak{F}, P)$.

Let Y be a random variable on this space.

Let

$$X_n = Y \circ T^n, n = 0,1, \ldots,$$

so that $X = (X_0, X_1, \ldots)$ is generated by T in the sense of Definition 2.

For any nonnegative integer k, let A be a measurable subset of G^{k+1}.

Since $X_{j+1} = X_j \circ T$ for each j,

$$P\left((X_1, \ldots, X_{k+1}) \in A\right) = P(X_0 \circ T, \ldots, X_k \circ T) \in A)$$

$$= P\left((X_0, \ldots, X_k) \circ T \in A\right)$$

$$= P\left(T^{-1}(\{(X_0, \ldots, X_k) \in A\})\right)$$

$$= P((X_0, \ldots, X_k) \in A).$$

This shows that $(X_1, \dots, X_{k+1})$ has the same distribution as $(X_0, \dots, X_k)$.

Similarly, we can show that $(X_n, \dots, X_{n+k})$ has the same distribution as $(X_0, \dots, X_k)$.

Thus, $(X_0, X_1, \dots)$ is stationary in the sense of Definition 1.

From Definition 1 to Definition 2

Let X be a sequence $(X_0, X_1, \dots)$ of G-valued random variables on a probability space $(\Omega, \mathfrak{F}, P)$, such that X is stationary in the sense of Definition 1. In the rest of this section we will show how to construct a representation $Z = (Z_0, Z_1, \dots)$ of X on $(G^\infty, \mathfrak{G}^\infty, Q)$, where $\mathfrak{G}^\infty$ is the product sigma-algebra on the sequence space G^∞, and Q is the distribution of X as a map into G^∞, so that

$$Q(B) = P(X^{-1}(B))$$

for all B in $\mathfrak{G}^\infty$. This representation has all the essential properties of the sequence X, and we will show that it is generated with a measure-preserving transformation, so Z satisfies Definition 2.

For later use we recall some definitions generated with product σ-algebras.

A measurable rectangle for G^{n+1} is a subset $A_0 \times A_1 \times \dots \times A_n$ of G^{n+1} where $A_i \in \mathfrak{G}$ for each $i = 0,1, \dots, n$. Then a measurable rectangle for G^∞ is a subset $A_0 \times A_1 \times \dots \times A_n \times G \times G \times \dots$. Thus a measurable rectangle in G^∞ is a set of the form $A \times G^\infty$ where A is a measurable rectangle in G^{n+1} for some n.

A cylinder set for G^∞ is a set of the form $B \times G^\infty$ where $B \in \mathfrak{G}^{n+1}$.

The product sigma-algebra $\mathfrak{G}^\infty$ is defined to be the σ-algebra generated by the measurable rectangles for G^∞. It is also the σ-algebra generated by the cylinder sets.

Let Z_n be defined on the probability space $(G^\infty, \mathfrak{G}^\infty, Q)$ by $Z_n\big((z_0, z_1, \dots)\big) = z_n$. Notice that Z_n reads the n-th coordinate of the sequence z. For any $D \in \mathfrak{G}$,

$Z_n^{-1}(D) = G^n \times \mathrm{D} \times G^\infty$, so $Z_n^{-1}(D)$ is a measurable rectangle. Thus Z_n is measurable.

Define the shift transformation T such that

$$T(z) = T((z_0, z_1, \dots)) = (z_1, z_2, \dots). \qquad (1)$$

From the definitions, $Z_n = Z_0 \circ T^n$.

We will show that T is a measure-preserving transformation. First, we need to prove that T is measurable from $\mathfrak{G}^\infty$ to $\mathfrak{G}^\infty$.

Let $D_n \in \mathfrak{G}$, and let $B_n = Z_n^{-1}(D_n)$ for each n, so that $B_n = G^n \times D_n \times G^\infty$.

Then

$$T^{-1}(B_n) = G^{n+1} \times D_n \times G^\infty$$

is a measurable rectangle, and hence $T^{-1}(B_n) \in \mathfrak{G}^\infty$.

Thus,

$$T^{-1}\Big(\bigcap_{t=0}^{n} B_t\Big) = \bigcap_{t=0}^{n} T^{-1}(B_t) \in \mathfrak{G}^\infty.$$

Since any measurable rectangle can be expressed as $\bigcap_{t=0}^{n} B_t$ for some choice of $D_0, D_1, \dots, D_n$, it follows that

$$T^{-1}(B) \in \mathfrak{G}^\infty$$

for any measurable rectangle B.

The collection of sets A such that $T^{-1}(A) \in \mathfrak{G}^\infty$ is a σ-algebra. Since it contains all the measurable rectangles, it contains all sets in $\mathfrak{G}^\infty$. Thus, $T^{-1}(A) \in \mathfrak{G}^\infty$ for any $A \in \mathfrak{G}^\infty$, and so T is measurable.

We would like to show that T is measure-preserving with respect to the measure Q.

Let A be a measurable rectangle, so that

$$A = \bigcap_{t=0}^{n} Z_t^{-1}(D_t) = D_0 \times D_1 \times \ldots \times D_n \times G^{\infty}.$$

for some subsets $D_0, D_1, \ldots, D_n$ of G.

Then

$$Q(A) = P(X \in A)$$

$$= P(X \in D_0 \times \ldots \times D_n \times G^{\infty})$$

$$= P((X_0, \ldots, X_n) \in D_0 \times \ldots \times D_n).$$

Since X is stationary in the sense of Definition 1,

$$Q(A) = P((X_1, \ldots, X_{n+1}) \in D_0 \times \ldots \times D_n)$$

$$= P(X \in \ G \times D_0 \times \ldots \times D_n \times G^{\infty}) = P\big(X \in T^{-1}(A)\big) = Q(T^{-1}(A)).$$

This shows that T is measure-preserving on measurable rectangles with respect to the measure Q.

We now recall Dynkin's $\pi - \lambda$ Theorem. This theorem is equivalent to one which was proved earlier by Sierpinski (see Fristedt-Gray 1997, Appendix G).

Definition Let *P* and L be collections of subsets of a set *X*. The collection *P* is called a π-system if it is closed under finite intersections; i.e. if $A, B \in P$ then $A \cap B \in P$.

The collection *L* is called a λ-system if the following hold:

1. $\emptyset \in L$;
2. If $A \in L$ then $A^c \in L$;

3. L is closed under countable disjoint unions; i.e. if $A_1, A_2, \dots \in L$ and if $A_i \cap A_j = \emptyset$ for every $i \neq j$, then $\bigcup_{j=1}^{\infty} A_j \in L$.

<u>Dynkin's $\pi - \lambda$ Theorem</u> Let P be a π-system of subsets of X, and let L be a λ-system of subsets of X. Suppose also that $P \subset L$. Then:

$$\sigma(P) \subset L.$$

It is easy to check that the measurable rectangles for $\mathfrak{B}^{\infty}$ form a π-system.

Let L be the collection of all sets $B \in \mathfrak{G}^{\infty}$ such that

$$Q(B) = Q\left(T^{-1}(B)\right).$$

It is easy to check that L is a λ-system.

By the $\pi - \lambda$ theorem, L contains the σ-algebra generated by measurable rectangles.

That is,

$$\mathfrak{G}^{\infty} \subset L.$$

Thus T is measure-preserving on all of $\mathfrak{G}^{\infty}$ with respect to Q.

Since $Z_n = Z_0 \circ T^n$, $(Z_0, Z_1, \dots)$ is a sequence generated with a measure-preserving transformation T in the sense of Definition 2.

Since $(Z_0, Z_1, \dots)$ has the same joint distribution as $(X_0, X_1, \dots)$, any physical meaningful property of$(X_0, X_1, \dots)$ can be expressed in terms of $(Z_0, Z_1, \dots)$. Thus we have shown that a good representation of the original process exists which satisfies Definition 2. We will refer to $(Z_0, Z_1, \dots)$ as the coordinate representation for $(X_0, X_1, \dots)$.

2 Functions of a Stationary Process

Theorem 1 Let $(\Omega, \mathfrak{F}, P)$ be a probability space and let $X_0, X_1, \dots$ be a sequence of G-valued random variables which is stationary in the sense of Definition 1. Let f be a real-valued function on G^∞ which is measurable with respect to $\mathfrak{G}^\infty$. Let $Y_n = f(X_n, X_{n+1}, \dots)$ for each $n = 0,1, \dots$. Then $Y_0, Y_1, \dots$ is stationary in the sense of Definition 1.

(Theorem 1 remains true if instead of a real-valued function f one considers a measurable function f with values in a general measurable space.)

Proof of Theorem 1:

Define $f_k(\boldsymbol{x}) = f(x_k, x_{k+1}, \dots)$.

Define the shift T on G^∞ as in equation (1).

Since $f_k = f \circ T^k$ and f is measurable, f_k is measurable since T is measurable.

Let $(x_0, x_1, \dots) = F(f_0(x_0, x_1, \dots), f_1(x_0, x_1, \dots), \dots)$.

If we let $\boldsymbol{x} = (x_0, x_1, \dots)$, we have

$F(\boldsymbol{x}) = (f_0(\boldsymbol{x}), f_1(\boldsymbol{x}), \dots)$

or with briefer notation,

$F = (f_0, f_1, \dots)$

Since each f_k is measurable, it is easy to check that F is measurable from $\mathfrak{G}^\infty$ to $\mathbb{R}^\infty$.

For example, let $B_0, \dots, B_m$ be Borel subsets of $\mathbb{R}$.

Let $B = B_0 \times B_1 \times \dots \times B_m \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \dots$. Then

$$F^{-1}(B) = f_0^{-1}(B_0) \cap f_1^{-1}(B_1) \cap f_m^{-1}(B_m)$$

and so $F^{-1}(B)$ is measurable.

Thus the inverse image of any measurable rectangle in $\mathbb{R}^\infty$ is measurable.

Since $\mathfrak{B}^\infty$ is generated by measurable rectangles, it is the smallest σ-algebra containing measurable rectangles.

Since the collection of all sets B such that $F^{-1}(B)$ is measurable forms a σ-algebra, it must contain $\mathfrak{B}^\infty$.

Thus, $F^{-1}(B)$ is measurable for any $B \in \mathfrak{B}^\infty$, so F is measurable.

Let $X = (X_0, X_1, \dots)$, and $X^1 = (X_1, X_2, \dots)$.

Since X is stationary in the sense of Definition 1, each finite sequence $X_0, X_1, \dots, X_n$ has the same distribution as $X_1, X_2, \dots, X_{n+1}$.

It follows that for any measurable rectangle B,

$$P(X \in B) = P(X^1 \in B).$$

The collection of sets B such that $P(X \in B) = P(X^1 \in B)$ is a λ-system, while the collection of measurable rectangles is easily seen to be a π-system.

Hence by the $\pi - \lambda$ theorem, $P(X \in B) = P(X^1 \in B)$ holds for every set B in the σ-algebra generated by measurable rectangles, that is, for every $B \in \mathfrak{B}^\infty$.

Thus, we have proved that X^1 and X has the same distribution. (The same argument shows that the distribution of any sequence X of random variables is completely determined by the finite-dimensional distributions of the sequence.)

By definition, $Y_k = f_k(X)$. Let $Y = (Y_0, Y_1, \dots)$, so that

$$Y = F(X).$$

Let $Y^1 = (Y_1, Y_2, \dots)$, and thus $Y^1 = F(X^1)$.

Since X and X^1 have the same distribution, it is immediate that Y and Y^1 have the same distribution.

Thus, $(Y_0, Y_1, \ldots)$ is stationary in the sense of Definition 1.

We can then apply this result:

Example a: $k = 0$. If *f* is a measurable real-valued function on S, then $(f(X_0), f(X_1), \ldots)$ is stationary.

Example b: $k = 1$. If *f* is a measurable real-valued function on S^2, then $(f(X_0, X_1), f(X_1, X_2), \ldots)$ is stationary

3 The Ergodic Theorem

Let T be a measure-preserving transformation on a probability space $(\Omega, \mathfrak{F}, P)$. A set $A \in \mathfrak{F}$ will be said to be *invariant* for T if $T^{-1}(A) = A$.

Lemma 1 Let S be the collection of all invariant sets for T. Then S is a σ-algebra. If X is a real-valued function which is measurable with respect to S, then $X \circ T = X$.

Proof of Lemma 1:

It follows easily from the definitions that S is a σ-algebra. Let X be measurable with respect to S.

Let $Y = X \circ T$. Then for any real number c, $\{Y = c\} = \{X \circ T = c\} = T^{-1}(\{X = c\})$.

Since X is measurable with respect to S, $\{X = c\} \in S$, and hence

$$T^{-1}(\{X = c\}) = \{X = c\}.$$

Thus $\{Y = c\} = \{X = c\}$, for *every* real number c.

Thus for any $\omega \in \Omega$, if $c = X(\omega)$, then $\omega \in \{X = c\}$, so $\omega \in \{Y = c\}$ and $Y(\omega) = c$ also.

Thus $X = Y$.

Theorem 2 (The Ergodic Theorem) Let T be measure-preserving on $(\Omega, \mathfrak{F}, P)$. Then for any real-valued random variable X such that $E|X| < \infty$, with probability one

$$\lim \frac{1}{n} \sum_{k=0}^{n-1} X \circ T^k = E(X|\mathcal{S})$$

where $\mathcal{S}$ is the invariant sigma-algebra.

The theorem was proved in October 1931 by G.D. Birkhoff. A version of the ergodic theorem proving L^2-convergence instead of almost sure convergence was obtained slightly earlier by John von Neumann (Bergelson, 2004).

Lemma 2 (Maximal Ergodic Theorem) Let T be a measure preserving transformation and X a random variable with $E(|X|) < \infty$.

Define

$$S_k(\omega) = X(\omega) + \cdots + X(T^{k-1}(\omega))$$

$$M_n = \max(0, S_1, S_2, \ldots, S_n)$$

Then

$$\int_{\{M_n > 0\}} X(\omega) dP(\omega) \geq 0.$$

Proof of Lemma 2:

The proof will follow the method given in Breiman (1992). This proof of the Maximal Ergodic Lemma is due to Garsia (1965).

Let $S_k^1 = S_k \circ T$, so that

$S_k^1(\omega) = X(T(\omega)) + \cdots + X(T^k(\omega)) = S_k - X(\omega) + X(T^k(\omega))$.

Then

$$M_n \circ T = M_n^1 = \max(0, S_1^1, S_2^1, \ldots, S_n^1).$$

For $k = 1, \ldots, n, M_n^1 \geq S_k^1$, so

$$X + M_n^1 \geq X + S_k^1 = S_{k+1}.$$

For $k = 0$, trivially

$X + M_n^1 \geq X = S_1$.

Hence, for $k = 1, 2, \ldots, n + 1$,

$$X \geq S_k - M_n^1.$$

That shows

$$X \geq \max(S_1, \ldots S_n) - M_n^1.$$

Since $M_n^1 = M_n \circ T$,

$$\int_{\{M_n > 0\}} X(\omega)dP(\omega) \geq$$

$$\int_{\{M_n > 0\}} \Big(\max\big(S_1(\omega), \ldots S_n(\omega)\big) - M_n\big(T(\omega)\big)\Big) dP(\omega).$$

But we have $M_n(\omega) = \max \ (S_1, S_2, \ldots, S_n)$ when $M_n(\omega) > 0$.

Increasing the integration set does not change the integral of $M_n(\omega)$, while it can only make the integral of $M_n(T(\omega))$ larger.

Further, T is measure preserving. This implies that $M_n \circ T$ has the same distribution as M_n, and hence has the same integral.

Thus,

$$\int_{(M_n > 0)} X(\omega)dP(\omega) \geq$$

$$\int_{(M_n > 0)} \{M_n(\omega) - M_n(T(\omega))\}dP(\omega) \geq \int \{M_n(\omega) - M_n(T(\omega))\}dP(\omega) = 0$$

This proves the lemma.

The next lemma is given as Problem 9 in Section 12.2 of Fristedt and Gray (2013).

Lemma 3 Let $X_0, X_1, \ldots$ be a sequence of identically-distributed real-valued random variables such that $E|X_0| < \infty$. Then with probability one

$$\frac{X_n}{n} \to 0.$$

Proof of Lemma 3:

For any $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P(\frac{|X_n|}{n} > \varepsilon) = \sum_{n=1}^{\infty} P(|X_n| > n\varepsilon) = \sum_{n=1}^{\infty} P(|X_0| > n\varepsilon)$$

$$= \sum_{n=1}^{\infty} \frac{1}{\varepsilon} \int_{(n-1)\varepsilon}^{n\varepsilon} P(|X_0| > n\varepsilon) dt \leq \sum_{n=1}^{\infty} \frac{1}{\varepsilon} \int_{(n-1)\varepsilon}^{n\varepsilon} P(|X_0| > t) dt$$

$$= \frac{1}{\varepsilon} \int_0^{\infty} P(|X_0| > t) dt = \frac{1}{\varepsilon} E|X_0| < \infty.$$

Hence by the first Borel-Cantelli Lemma, with probability one, $\frac{|X_n|}{n} > \varepsilon$ occurs at most finitely many times. Since $\varepsilon > 0$ is arbitrary, this shows that with probability one, $\frac{X_n}{n} \to 0$, proving the lemma.

Proof of Theorem 2:

The proof follows the steps in Breiman (1992) with some additional details.

Let $Y = E(X|\mathcal{S})$. Then Y is measurable with respect to $\mathcal{S}$. By Lemma 1, $Y \circ T = Y$, and so it is obvious that

$$\lim \frac{1}{n} \sum_{k=0}^{n-1} Y \circ T^k = Y = E(Y|\mathcal{S}),$$

so the Ergodic Theorem holds for Y. Replacing X by $X - Y$, we assume from now on that $E(X|\mathcal{S}) = 0$. we must show that

$$\lim \frac{1}{n} \sum_{k=0}^{n-1} X \circ T^k = 0$$

holds with probability one.

We will first give the proof for a special case.

Special case:

Assume that

$$\frac{X \circ T^n}{n} \to 0$$

everywhere on Ω. This condition is obvious when X is bounded, but may not hold in general.

Let $\overline{X} = \overline{\lim} \frac{S_n}{n}$, and $D = \{\overline{X} > \varepsilon\}$, for any $\varepsilon > 0$.

Since $S_k \circ T = S_k + X \circ T^k - X$, it is easy to show that $\overline{X} = \overline{X} \circ T$. It follows that $\mathcal{X}_D \circ T = \mathcal{X}_D$.

Define $X^*(\omega) = (X(\omega) - \varepsilon)\mathcal{X}_D(\omega)$,

$S_k^*(\omega) = X^*(\omega) + \cdots + X^*(T^{k-1}\omega)$,

$M_n^*(\omega) = \max \ (0, S_1^*, S_2^*, \ldots, S_n^*)$.

Note that

$$S_k^* = (S_k - k\varepsilon)\mathcal{X}_D, \frac{S_k^*}{k} = \left(\frac{S_k}{k} - \varepsilon\right)\mathcal{X}_D, \sup_{k\geq 1}\frac{S_k^*}{k} = (\sup_{k\geq 1}\frac{S_k}{k} - \varepsilon)\mathcal{X}_D.$$

By the maximal ergodic theorem, $\int_{(M_n^*>0)} X^* dP \geq 0$.

Let $F_n = \{M_n^* > 0\} = \{\max_{1\leq k\leq n} S_k^* > 0\}$, and let

$$F = \{\sup_{k\geq 1} S_k^* > 0\} = \left\{\sup_{k\geq 1}\frac{S_k^*}{k} > 0\right\} = \left\{\sup_{k\geq 1}\frac{S_k}{k} > \varepsilon\right\} \cap D.$$

It is easy to see that F_n converges upward to F.

By definition F is a subset of D. On the other hand, clearly $\sup_{k\geq 1}\frac{S_k}{k} > \overline{X}$, so $\left\{\sup_{k\geq 1}\frac{S_k}{k} > \varepsilon\right\}$ contains D. Hence $F = D$.

Since $E|X^*| \leq E|X| + \varepsilon$, by the Dominated Convergence theorem,

$$\int_{F_n} X^* dP = \int \mathcal{X}_{F_n} X^* dP \to \int \mathcal{X}_F X^* dP = \int_F X^* dP$$

Thus,

$$0 \leq \int_D X^* dP = \int_D X dP - \varepsilon P(D) = \int_D E(X|\mathcal{S}) dP - \varepsilon P(D) = -\varepsilon P(D).$$

Hence $P(D) = 0$, and thus $\overline{X} \leq 0$ with probability one.

By a similar argument for the random variable $-X(\omega)$, we obtain $\overline{-X} \leq 0$.

Also $\overline{-X} = \text{limsup}_n \frac{-S_n}{n} = -\text{liminf}_n \frac{S_n}{n}$.

Since $\text{limsup}_n \frac{S_n}{n} \leq 0$ and $\text{liminf}_n \frac{S_n}{n} \geq 0$, with probability one

$\lim_n \dfrac{S_n}{n} = 0.$

this completes the proof under the special assumption that $\frac{X_n}{n} \to 0$ holds everywhere. To finish the proof in the general case, we must give an additional argument. We need to prove the Ergodic Theorem for a random variable X such that $E(X|\mathcal{S}) = 0$.

By Lemma 3,

$$\lim \frac{X_n}{n} = 0$$

holds with probability one.

Let $G = \left\{\lim \frac{X_n}{n} = 0\right\}$ and let $B = \Omega - G$. Then $P(B) = 0$.

Let $B_n = T^{-n}(B)$. Since T is measure-preserving, $P(B_n) = 0$ for all n.

Let H be the union of the sets B_n. Since P is countably additive, $P(H) = 0$.

Define the random variable X' by $X' = X$ on $\Omega - H$, $X' = 0$ on H.

Then $X' = X$ holds with probability one, so

$$\int_A X' dP = \int_A X dP = 0$$

for every $A \in \mathcal{S}$. Hence $E(X'|\mathcal{S}) = 0$.

Let $\omega \in \Omega - H$. Suppose that $T^k(\omega) \in H$ for some k. Then $T^k(\omega) \in B_n$ for some n, and so $T^n\left(T^k(\omega)\right) \in B$, i.e. $\omega \in B_{n+k}$, so $\omega \in H$, contradiction.

This shows that if $\omega \in \Omega - H$ then $T^k(\omega) \in \Omega - H$ for every $k = 0,1, \dots$.

Since $X' = X$ everywhere on $\Omega - H$, $X' \circ T^k = X \circ T^k$ everywhere on $\Omega - H$ for all $k = 0,1, \dots$.

Hence

$$\frac{1}{n}\sum_{k=0}^{n-1} X' \circ T^k = \frac{1}{n}\sum_{k=0}^{n-1} X \circ T^k$$

holds everywhere on $\Omega - H$.

Since $P(H) = 0$, if we can show that

$$\lim \frac{1}{n}\sum_{k=0}^{n-1} X' \circ T^k = 0$$

with probability one, then proof of the theorem will be complete.

Since $B \subset H$, $\Omega - H \subset G$, and so we know that $\frac{X' \circ T^n}{n} = \frac{X \circ T^n}{n} \to 0$ everywhere on $\Omega - H$.

On the other hand, suppose that $\omega \in H$. There are two possibilities. If $T^k(\omega) \in H$ for all n, then $\frac{X' \circ T^n}{n} = 0$ for all n. If instead $T^k(\omega) = \omega_1 \in \Omega - H$ for some k, then for all $n > k$ we have

$$\frac{X' \circ T^n(\omega)}{n} = \frac{X \circ T^{n-k}(\omega_1)}{n} = \frac{n-k}{n} \times \frac{X \circ T^{n-k}(\omega_1)}{n-k} \to 0.$$

Thus for all ω, $\frac{X' \circ T^n(\omega)}{n} \to 0$, i.e. $\frac{X' \circ T^n}{n} \to 0$ holds *everywhere*. Since the Ergodic Theorem has already been proved under this assumption, the proof of the theorem is complete.

Remarks:

Let $X = (X_0, X_1, \dots)$ be real-valued and stationary in the sense of Definition 1.

In Section 1 we showed that the coordinate representation *Z* for *X* is generated by a measure-preserving transformation, in the sense of Definition 1.

The probability measure for the sequence *Z* is $Q = P(X^{-1})$ which is the distribution of *X* on the sequence space.

Thus, by the ergodic theorem we have that

$$\frac{(Z_0 + \cdots + Z_{n-1})}{n}$$

converges with *Q*-probability one.

Thus,

$$\lim\sup \frac{(Z_0 + \cdots + Z_{n-1})}{n} = \lim\inf \frac{(Z_0 + \cdots + Z_{n-1})}{n}$$

holds with *Q*-probability one.

Since

$$X^{-1}\left(\left\{\lim\sup \frac{(Z_0 + \cdots + Z_{n-1})}{n} = \lim\inf \frac{(Z_0 + \cdots + Z_{n-1})}{n}\right\}\right.$$

$$= \left\{\lim\sup \frac{(X_0 + \cdots + X_{n-1})}{n} = \lim\inf \frac{(X_0 + \cdots + X_{n-1})}{n}\right\},$$

we have that

$$\lim\sup \frac{(X_0 + \cdots + X_{n-1})}{n} = \lim\inf \frac{(X_0 + \cdots + X_{n-1})}{n}$$

holds with P-probability one.

Therefore we can conclude that from the Ergodic Theorem for $Z_0, Z_1, \ldots$ that

$$\frac{(X_0 + \cdots + X_{n-1})}{n}$$

converges with probability one.

In Example a of Section 2, where X_n takes values in a general set G, and f is real-valued, assume that $E|f(X_0)| < \infty$. Then

$$\frac{f(X_0) + f(X_1) + \cdots + f(X_{n-1})}{n}$$

converges with probability one.

In Example b of Section 2, assume that $E|f(X_0, X_1)| < \infty$. Then

$$\frac{f(X_0, X_1) + f(X_1, X_2) + \cdots + f(X_{n-1}, X_n)}{n}$$

converges with probability one.

4 Ergodic Transformations

Let T be a measure-preserving transformation on a probability space $(\Omega, \mathfrak{F}, P)$.

T is said to be *ergodic* if for every invariant set A, $P(A) = 0$ or $P(A) = 1$.

For any stationary sequence $(X_0, X_1, \dots)$, the sequence will be said to be ergodic if the coordinate representation of the sequence on sequence space (described in Section 1) is ergodic with respect to the shift transformation on sequence space.

Equivalently, $(X_0, X_1, \dots)$ is ergodic if for any measurable subset C of sequence space which is invariant under the shift, either $P\big((X_0, X_1, \dots)\ in\ C\big) = 0$ or $P\big((X_0, X_1, \dots)\ in\ C\big) = 1$.

When X_n is real-valued with finite expectation, the Ergodic Theorem states that

$$\lim_{n\to\infty} \frac{S_n}{n} = E(X_0 | \mathcal{S}) \quad a.s.$$

Hence the limit is equal almost surely to a function which is measurable with respect to the invariant σ-algebra.

Thus for an ergodic transformation, the limit in the Ergodic Theorem is the constant function equal to $\int X_0 dP$.

Notice that for an ergodic stationary sequence $X_0, X_1, \dots$, the conclusion of the ergodic theorem is the same as the conclusion of the SLLN for an identical and independent distributed sequence $X_0, X_1, \dots$, but the assumption is more general.

5 Markov Chains and Stationary Processes

Let S be a countable set, and let $\mathfrak{S}$ be the collection of all subsets of S. A transition function with state space S is a function p from $S \times S$ to $[0,1]$ such that for all $x \in S$,

$$\sum_{y \in S} p(x,y) = 1.$$

A sequence of random variables $X_0, X_1, \ldots$ is said to be a **Markov Chain** with transition function *p* if for all $x_0, \ldots, x_{n-1}, x, y \in S$, whenever

$$P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) > 0$$

We have

$$P\{X_{n+1} = y | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x\} = p(x,y)$$

Equivalently, $X_0, X_1, \ldots$ is said to be a Markov Chain with transition function *p* if for all $x_0, \ldots, x_{n-1}, x, y \in S$,

$$P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x, X_{n+1} = y)$$

$$= P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x) p(x,y)$$

$$= P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) p(x_{n-1}, x) p(x,y). \qquad (\mathrm{M}-1)$$

By induction,

$$P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x_n)$$

$$= P(X_0 = x_0) p(x_0, x_1) \ldots p(x_{n-1}, x_n) \qquad (\mathrm{M}-2)$$

for any $x_0, \ldots, x_{n-1}, x_n \in S$. We call this the *path probability formula*. For a sequence $X_0, X_1, \ldots$, and a transition function p, it is easy to see that the path probability formula (M-2) holds if and only if the Markov property (M-1) holds.

A transition function p on a countable state space is said to be irreducible if for all points x, y in S, there exists a sequence of points $x_0, \ldots, x_k$ in S such that $x_0 = x, x_k = y$, and $p(x_{j-1}, x_j) > 0$ for $j = 1, \ldots, k$. Because of the path probability formula, to say that p is irreducible implies that a Markov chain with transition function p which starts at a point x can move with positive probability from x to y, for any points x and y. A Markov chain with irreducible transition function p is said to be an irreducible Markov chain.

For any positive integer k and any non-negative integer n, by the path probability formula we have

$$P\left(X_0 = x_{0,\ldots,}X_{n+k-1} = x_{n+k-1}, X_{n+k} = x_{n+k}\right)$$

$$= P(X_0 = x_0)p(x_0, x_1) \ldots p(x_{n+k-1}, x_{n+k}).$$

Since

$$P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x_n)$$

$$= P(X_0 = x_0)p(x_0, x_1) \ldots p(x_{n-1}, x_n),$$

this gives

$$P(X_0 = x_0, \ldots, X_{n+k-1} = x_{n+k-1}, X_{n+k} = x_{n+k})$$

$$= P(X_0 = x_0, \ldots, X_{n-1} = x_{n-1}, X_n = x_n)p(x_n, x_{n+1}) \ldots p(x_{n+k-1}, x_{n+k})$$

Summing over all possible values of $x_0, \ldots x_{n-1}$, then gives

$$P(X_n = x_n, \ldots, X_{n+k-1} = x_{n+k-1}, X_{n+k} = x_{n+k})$$

$$= P(X_n = x_n)p(x_n, x_{n+1}) \ldots p(x_{n+k-1}, x_{n+k}).$$

We will call this the path probability formula starting at time n.

Taking $k = 1$ in the path probability formula starting at time n, we have

$$P(X_n = x, X_{n+1} = y) = P(X_n = x)p(x, y).$$

If $P(X_n = x) > 0$, this is equivalent to

$$P(X_{n+1} = y | X_n = x) = p(x, y).$$

We will call this the *one-step* transition formula.

Summing over $x \in S$ in the equation $P(X_n = x, X_{n+1} = y) = P(X_n = x)p(x, y)$, we obtain

$$P(X_{n+1} = y) = \sum_{x \in S} P(X_n = x)p(x, y).$$

This equation also follows from the Law of Total Probability and the one-step transition formula. We will call this equation the *updating formula for distributions*.

A function $\mu: S \to [0,1]$ will be called a distribution on S if

$$\sum_{x \in S} \mu(x) = 1.$$

If $P(X_n = x) = \mu(x)$ for all x in S then we say that μ is the distribution of X_n.

A distribution π is said to be a stationary distribution if , for every $y \in S$,

$$\sum_x \pi(x)p(x, y) = \pi(y)$$

It can be shown that for an irreducible transition function p, if a stationary distribution exists then it is unique (Norris, 1998, Theorem 1.7.7).

It is an easy fact that, for a stationary distribution π, if $\pi(x) > 0$ and $p(x, y) > 0$ then $\pi(y) > 0$. Hence when the transition function is irreducible and the stationary distribution exists, it is positive everywhere.

If it happens that $P(X_n = x) = \pi(x)$ for all $x \in S$, so that π is the distribution of X_n, then the updating formula $P(X_{n+1} = y) = \sum_{x \in S} P(X_n = x)p(x, y)$, together with the definition of a stationary distribution, shows us that π is also the distribution of X_{n+1}. Thus if it happens that X_0 has distribution π, then X_n has distribution π for all n.

Theorem 3 Let $X_0, X_1, \dots$ be a discrete-time Markov Chain with stationary distribution π such that X_0 has distribution π. Then $X_0, X_1, \dots$ is stationary in the sense of Definition 1.

Proof of Theorem 3:

Let $x_0, \dots, x_k \in S$, and n be any non-negative integer.

It was already noted that for each n, X_n has distribution π. Thus,

$$P(X_n = x_0) = \pi(x_0) \text{ and } P(X_{n+1} = x_0) = \pi(x_0).$$

By the path probability formula starting at n,

$$P(X_n = x_0, \dots, X_{n+k-1} = x_{k-1}, X_{n+k} = x_k)$$

$$= P(X_n = x_0)p(x_0, x_1) \dots p(x_{k-1}, x_k)$$

$$= \pi(x_0)p(x_0, x_1) \dots p(x_{k-1}, x_k).$$

Similarly by the path probability formula starting at $n + 1$,

$$P(X_{n+1} = x_0, \dots, X_{n+k} = x_{k-1}, X_{n+k+1} = x_k)$$

$$= P(X_{n+1} = x_0)p(x_0, x_1) \dots p(x_{k-1}, x_k)$$

$$= \pi(x_0)p(x_0, x_1) \dots p(x_{k-1}, x_k).$$

Thus,

$$\begin{aligned} &P(X_n = x_0, \dots, X_{n+k-1} = x_{k-1}, X_{n+k} = x_k) \\ &\qquad = P(X_{n+1} = x_0, \dots, X_{n+k} = x_{k-1}, X_{n+k+1} = x_k). \end{aligned}$$

Let A be any subset of $\mathfrak{S}^{k+1}$.

Summing both sides of the preceding equation over all $x_0, \dots, x_k$ such that $(x_0, \dots, x_k) \in A$, we obtain,

$$P\big((X_n, \dots, X_{n+k}) \in A\big) = P\big((X_{n+1}, \dots, X_{n+k+1}) \in A\big).$$

Hence the distribution of $(X_n, \dots, X_{n+k})$ is equal to the distribution of $(X_{n+1}, \dots, X_{n+k+1})$. It follows that the distribution of $(X_n, \dots, X_{n+k})$ is the same for all n, and thus $X_0, X_1, \dots$ is stationary in the sense of Definition 1. This proves the Theorem.

6 Markov Chains on Sequence Space

Given a Markov transition function p on S and a distribution μ on S, the next lemma, which will not be proved, shows that there always exists a Markov chain with Markov transition function p and initial distribution μ. In fact, as the lemma shows, the chain can always be defined on the sequence space S^{∞} with a given transition function and a given initial distribution (Revuz, 1984, Theorem 2.8).

Lemma 4 Let p be a Markov transition function on a countable set S. Let $\mu: S \rightarrow [0,1]$ be a function such that $\sum_{y \in S} \mu(y) = 1$. Let $\mathfrak{G}$ be the collection of all subsets of S. Then exists a unique probability measure Q on $\mathfrak{G}^{\infty}$ such that $Z_0, Z_1, \ldots$ is a Markov chain with respect to Q having Markov transition function p, such that $Q(Z_0 = y) = \mu(y)$ for every $y \in S$.

The measure Q described in the lemma will often be denoted by Q_{μ}. If $x \in S$ and μ is the function such that $\mu(x) = 1$ and $\mu(y) = 0$ for all $y \neq x$, then Q_{μ} will also be denoted by Q_x.

The path probability formula shows easily that for any $x_0, \ldots, x_n \in S$,

$$Q_x(Z_0 = x_0, \ldots, Z_n = x_n) = p(x_0, x_1) \ldots p(x_{n-1}, x_n) \ if \ x_0 = x,$$

(Q-x-1)

and,

$$Q_x(Z_0 = x_0, \ldots, Z_n = x_n) = 0 \ if \ x_0 \neq x.$$

(Q-x-2)

Using this fact and the $\pi - \lambda$ theorem, it can be shown that if $\tilde{Q}$ is any probability measure on $\mathfrak{G}^\infty$ such that $Z_0, Z_1, \dots$ is a Markov chain with respect to $\tilde{Q}$ having Markov transition function p, and such that $\tilde{Q}(Z_0 = x) > 0$, then

$$Q_x(A) = \tilde{Q}(A | Z_0 = x)$$

for every $A \in \mathfrak{G}^\infty$.

Thus the chain with probability measure Q_x is obtained from the chain with probability measure $\tilde{Q}$ by restricting attention to paths starting at the point x.

We can use the probabilities Q_x to prove a more general form of the Markov property.

Lemma 5 (General Markov property) Let $x_0, \dots, x_n \in S$ and let $A \in \mathfrak{G}^\infty$. Let Q be any probability measure such that $Z_0, Z_1, \dots$ is a Markov chain with respect to Q and having transition function p. Then

$$Q(Z_0 = x_0, \dots, Z_n = x_n, (Z_n, Z_{n+1}, \dots) \in A)$$

$$= Q(Z_0 = x_0) p(x_0, x_1) \dots p(x_{n-1}, x_n) Q_{x_n}(A)$$

$$= Q(Z_0 = x_0, \dots, Z_n = x_n) Q_{x_n}(A).$$

Proof of Lemma 5:

First suppose that $A = \{(y_0, y_1, \dots, y_k)\} \times B$ for some $B \in \mathfrak{G}^\infty$.

If $y_0 = x_n$ then using the path probability formula for Q, the left side of the equation is

$$Q(Z_0 = x_0, \dots, Z_n = x_n, Z_{n+1} = y_1, \dots, Z_{n+k} = y_k)$$

$$= Q(Z_0 = x_0) p(x_0, x_1) \dots p(x_{n-1}, x_n) p(y_0, y_1) \dots p(y_{k-1}, y_k),$$

and this is also equal to the right side of the equation using the path probability formula for Q_{x_n}, by equation Q-x-1.

If $y_0 \neq x_n$, the left side is zero since the relevant set is empty, while the right side is zero since $Q_{x_n}(A)$ is zero by equation Q-x-2.

Now suppose that $A = H_0 \times H_1 \times \ldots \times H_n \times S^\infty$ for some subsets $S_0, S_1, \ldots, S_n$ of S. Summing both sides of the equation in the previous case over all

$y_0 \in H_0, y_1 \in H_1, \ldots, y_n \in H_n$, we see that the equation holds in this case also.

We have shown that the equation holds when A is any measurable rectangle in S^∞. The collection of all measurable rectangles is a π system. The collection of sets $A \in \mathfrak{G}^\infty$ such that the equation holds is easily seen to be a λ system. Hence, by the $\pi - \lambda$ theorem, the collection of sets $A \in \mathfrak{G}^\infty$ such that the equation holds contains the σ-algebra generated by the measurable rectangles, i.e. it contains all of $\mathfrak{G}^\infty$.

This proves the lemma.

7 Applying the Ergodic Theorem to Markov Chains

Lemma 6 Let p be an irreducible Markov transition function on S. Suppose that there exists a stationary distribution π for p. Let Q_π be the probability measure such that $Z_0, Z_1, \dots$ is a Markov chain with respect to Q_π having transition function p and initial distribution π. Then the sequence $Z_0, Z_1, \dots$ is ergodic.

Proof of Lemma 6:

Let C be a measurable subset of S^∞ such that $T^{-1}(C) = C$, where T is the shift. We must show that $Q_\pi(C) = 0$ or $Q_\pi(C) = 1$.

Let $z = (x_0, x_1, \dots) \in S^\infty$.

Since $T^{-1}(C) = C$,

$$z \in C \Leftrightarrow T(z) \in C.$$

That is, $(x_0, x_1, \dots) \in C \Leftrightarrow (x_1, x_2, \dots) \in C$.

Define the function φ on S by

$$\varphi(x) = Q_x(C) = Q_x\big((Z_0, Z_1, \dots) \in C\big).$$

By additivity,

$$\varphi(x) = \sum_{y \in S} Q_x(Z_1 = y, (Z_0, Z_1, \dots) \in C).$$

Since $Q_x(Z_0 = x) = 1$,

$$\varphi(x) = \sum_{y \in S} Q_x(Z_0 = x, Z_1 = y, (Z_0, Z_1, \dots) \in C).$$

Since C is invariant,

$$\varphi(x) = \sum_{y \in S} Q_x(Z_0 = x, Z_1 = y, (Z_1, Z_2, \dots) \in C).$$

By Lemma 5,

$$\varphi(x) = \sum_{y\in S} p(x,y)Q_y(C) = \sum_{y\in S} p(x,y)\varphi(y).$$

Let $x_0, \dots, x_n \in S$. Since C is invariant,

$$Q(\{Z_0 = x_0, \dots, Z_n = x_n\} \cap C)$$

$$= Q_\pi(Z_0 = x_0, \dots, Z_n = x_n, (Z_0, Z_1, \dots) \in C)$$

$$= Q_\pi(Z_0 = x_0, \dots, Z_n = x_n, (Z_n, Z_{n+1}, \dots) \in C).$$

By Lemma 5,

$$Q_\pi(Z_0 = x_0, \dots, Z_n = x_n, (Z_0, Z_1, \dots) \in C)$$

$$= Q_\pi(Z_0 = x_0)p(x_0, x_1) \dots p(x_{n-1}, x_n)Q_{x_n}(C)$$

$$= Q_\pi(Z_0 = x_0, \dots, Z_n = x_n)\ \varphi(x_n).$$

Thus if $Q_\pi(Z_0 = x_0, \dots, Z_n = x_n) > 0$,

$$Q_\pi(C|Z_0 = x_0, \dots, Z_n = x_n) = \varphi(x_n), \qquad (\mathrm{C}-1)$$

that is,

$$E_\pi(\mathcal{X}_C|Z_0 = x_0, \dots, Z_n = x_n) = \varphi(x_n).$$

Using the definition of conditional expectation as a random variable, we have proved that

$$E_\pi(\mathcal{X}_C|Z_0, \dots, Z_n) = \varphi(Z_n).$$

By the Martingale Theorem, $\varphi(Z_n)$ converges with probability one to some limit V. For any k, and any set $A \in \sigma(Z_0, \dots, Z_k)$,

$$\int_A \varphi(Z_n)dP = \int_A \mathcal{X}_C\, dP,$$

since $\varphi(Z_n) = E(\mathcal{X}_C|Z_0, \dots, Z_n)$.

By the Dominated Convergence Theorem, it follows that

$$\int_A VdP = \int_A \mathcal{X}_C dP.$$

Let H be the collection of all sets A such that for some non-negative integer k which may demand on A, $A \in \sigma(Z_0, \dots, Z_k)$. That is, let

$$H = \bigcup_{k=0}^{\infty} \sigma(Z_0, \dots, Z_k).$$

Then H is a π-system, and $\int_A VdP = \int_A \mathcal{X}_C dP$ for all $A \in H$. The collection L of all sets A such that $\int_A VdP = \int_A \mathcal{X}_C dP$ is easily seen to be a λ-system. By the $\pi - \lambda$ theorem, $\sigma(H) \subset L$. Thus $\sigma(Z_0, Z_1, \dots) \subset L$, so

$$\int_A VdP = \int_A \mathcal{X}_C dP$$

for every $A \in \sigma(Z_0, Z_1, \dots)$.

Since V and C are both measurable with respect to $\sigma(Z_0, Z_1, \dots)$, it follows that $V = \mathcal{X}_C$ holds almost surely. That is, $\varphi(Z_n) \to \mathcal{X}_C$ with probability one.

Since $Z_0, Z_1, \dots$ is stationary, $\varphi(Z_n)$ has the same distribution for all n. Hence the distribution of $\varphi(Z_n)$ is the same as the distribution of $\mathcal{X}_C$. Thus with probability one, $\varphi(Z_0) \in \{0,1\}$.

Let K be the set of $x \in S$ such that $\varphi(x) \notin \{0,1\}$. Then $P(Z_0 \in K) = 0$.

But $P(Z_0 \in x) = \pi(x) > 0$ for all x. Hence $K = \emptyset$.

Let $A_1 = \{x: \varphi(x) = 1\}$, $A_2 = \{x: \varphi(x) = 0\}$. Then $S = A_1 \cup A_2$.

Suppose that both A_1 and A_2 are nonempty.

Since p is irreducible, there exists points $x \in A_2$ and $y \in A_1$ such that $p(x, y) > 0$.

Hence

$$\varphi(x) = \sum_{z \in S} p(x,z)\varphi(z) \geq p(x,y)\varphi(y) > 0,$$

which is a contradiction.

Thus either $A_1 = \emptyset$ or $A_2 = \emptyset$.

Suppose $A_2 = \emptyset$. Then $\varphi(x) = 1$ for all x. By equation (C-1) with $n = 0$,

$$Q_\pi(C) = \sum_{x \in S} Q_\pi(Z_0 = x) Q_x(C) = \sum_{x \in S} Q_\pi(Z_0 = x)\varphi(x) = \sum_{x \in S} Q_\pi(Z_0 = x) = 1.$$

Similarly if $A_1 = \emptyset$ we have $Q_\pi(C) = 0$.

This proves the lemma.

8 Conclusions

Let the assumptions of Lemma 6 hold. Since the transition function p is assumed to be irreducible, we know that the chain $Z_0, Z_1, \dots$ is ergodic, and so by the Ergodic Theorem, for any $\mathfrak{G}^\infty$ measurable function f on S^∞ such that $\int |f| \, dQ_\pi < \infty$, we have

$$\begin{aligned}\frac{1}{n+1}&\left(f(Z_0, Z_1, \dots) + f(Z_1, Z_2, \dots) + \dots + f(Z_n, Z_{n+1}, \dots)\right) \\ &= \frac{1}{n+1}\left(f(Z_0, Z_1, \dots) + f(Z_0, Z_1, \dots) \circ T + \dots + f(Z_0, Z_1, \dots) \circ T^n\right) \\ &\to \int f \, dQ_\pi,\end{aligned}$$

Q_π-almost surely.

Let A be the set of points in S^∞ where

$$\frac{1}{n+1}\left(f(Z_0, Z_1, \dots) + f(Z_1, Z_2, \dots) + \dots + f(Z_n, Z_{n+1}, \dots)\right)$$

does not converge to $\int f \, dQ_\pi$.

Then $Q_\pi(A) = 0$, so $\sum_{x \in S} \pi(x) Q_x(A) = 0$.

Since $\pi(x) > 0$ for all $x \in S$, $Q_x(A) = 0$ for all $x \in S$.

Now let μ be any initial distribution. Then

$$Q_\mu(A) = \sum_{x \in S} \mu(x) Q_x(A) = 0.$$

This shows that

$$\frac{1}{n+1}\left(f(Z_0, Z_1, \dots) + f(Z_1, Z_2, \dots) + \dots + f(Z_n, Z_{n+1}, \dots)\right) \to \int f \, dQ_\pi,$$

Q_μ-almost surely.

If one is given a Markov chain $X_0, X_1, \dots$ on some general probability space having transition function p such that p is irreducible, we can use the arguments already made at the end of Section 3 about moving convergence proofs to the representation $Z_0, Z_1, \ldots$ of the chain defined on S^∞.

If the distribution of X_0 is μ, then the probability Q for the sequence $Z_0, Z_1, \ldots$ will be Q_μ.

Thus, assuming that p has stationary distribution π, it follows that

$$\frac{1}{n+1}\big(f(X_0, X_1, \ldots) + f(X_1, X_2, \ldots) + \cdots + f(X_n, X_{n+1}, \ldots)\big) \to \int f\, dQ_\pi,$$

P-almost surely.

If we let $f(x_0, x_1, \ldots) = r(x_0)$, this equation becomes

$$\frac{1}{n+1}\big(r(X_0) + \cdots + r(X_n)\big) \to \int r(X_0)\, dQ_\pi = \sum_{x \in S} \pi(x) r(x)$$

P-almost surely. This is the Strong Law for Markov Chains stated at the beginning of the paper.

References

Bergelson, Vitaly. Some historical remarks and modern questions around the ergodic theorem. *Dynamics of Complex Systems* (2004): 1-11.
Breiman, Leo. *Probability*, volume 7 of Classics in Applied Mathematics. *Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA*(1992).
Durrett, Richard. *Essentials of stochastic processes*. Springer Science & Business Media, 2012.
Fristedt, Bert E., and Lawrence F. Gray. *A modern approach to probability theory*. Birkhäuser, *Boston 1997.*
Garsia, Adriano M. A simple proof of E Hopf's maximal ergodic theorem. *Journal of Mathematics and Mechanics* 14, no. 3 (1965): 381.
Geyer, Charles J. Markov chain Monte Carlo lecture notes. *Course notes, Spring Quarter* (1998).
Norris, James R. *Markov chains*. Cambridge university press, 1998.
Revuz, Daniel. *Markov chains*. Elsevier, 1984.