MARKOV CHAINS AND THE ERGODIC THEOREM

CHAD CASAROTTO

Abstract. This paper will explore the basics of discrete-time Markov chains used to prove the Ergodic Theorem. Definitions and basic theorems will allow us to prove the Ergodic Theorem without any prior knowledge of Markov chains, although some knowledge about Markov chains will allow the reader better insight about the intuitions behind the provided theorems. Even for those familiar with Markov chains, the provided definitions will be important in providing the uses for the various notations used in this paper.

Contents

1. Basic Definitions and Properties of Markov Chains

Markov chains often describe the movements of a system between various states. In this paper, we will discuss *discrete-time* Markov chains, meaning that at each step our system can either stay in the state it is in or change to another state. We denote the random variable X_n as a sort of marker of what state our system is in at step n. X_n can take the value of any $i \in I$, where each i is a *state* in the *state-space*, I. States are usually just denoted as numbers and our state-space as a countable set.

We will call $\lambda = (\lambda_{i_1}, \lambda_{i_2}, \ldots) = (\lambda_i \,|\, i \in I)$ the *probability distribution* on X_n if: $\lambda_i = P(X_n = i)$ and $\sum_{i \in I} \lambda_i = 1$. Also, a matrix $P = \{p_{ij}\}$, where $i, j \in I$, is called *stochastic* if $\sum_{j \in I} \lambda_{ij} = 1$, $\forall i \in I$, i.e. every row of the matrix is a distribution. Now we can define a Markov chain explicitly.

Definition 1.1. $(X_0, X_1, \ldots) = (X_n)_{n \geq 0}$ is a *Markov chain with initial distribution* λ *and transition matrix* P, shortened to *Markov*(λ, P), if

- λ is the probability distribution on X_0;
- given that $X_n = i$, $(p_{ij} \,|\, i, j \in I)$ is the probability distribution on X_{n+1} and is independent of $X_k, 0 \leq k < n$, i.e. $P(X_{N+1} = j \,|\, X_n = i) = p_{ij}$.

Date: AUGUST 17, 2007.

Theorem 1.2. *$(X_n)_{0\le n\le N}$ is Markov(λ, P) if and only if*

$$P(X_0 = i_0, X_1 = i_1 \ldots, X_N = i_N) = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-1} i_N}. \tag{1.3}$$

Proof. First, suppose $(X_n)_{0\le n\le N}$ is Markov(λ, P), thus

$$\begin{aligned}
&P(X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N)\\
&\quad = P(X_0 = i_0)P(X_1 = i_1 \,|\, X_0 = i_0)\cdots P(X_N = i_N \,|\, X_0 = i_0, \ldots, X_{N-1} = i_{N-1})\\
&\quad = P(X_0 = i_0)P(X_1 = i_1 \,|\, X_0 = i_0)\cdots P(X_N = i_N \,|\, X_{N-1} = i_{N-1})\\
&\quad = \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-1} i_N}
\end{aligned}$$

Now assume that (1.3) holds for N, thus

$$\begin{aligned}
P(X_0 = i_0, X_1 = i_1 \ldots, X_N = i_N) &= \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-1} i_N}\\
\sum_{i_N \in I} P(X_0 = i_0, X_1 = i_1 \ldots, X_N = i_N) &= \sum_{i_N \in I} \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-1} i_N}\\
P(X_0 = i_0, X_1 = i_1 \ldots, X_{N-1} = i_{N-1}) &= \lambda_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{N-2} i_{N-1}}
\end{aligned}$$

And now by induction, (1.3) holds for all $0 \le n \le N$. From the formula for conditional probability, namely that $P(A \,|\, B) = P(A \cap B)/P(B)$, we can show that

$$\begin{aligned}
P(X_{N+1} = i_{N+1} \,|\, X_0 = i_0, \ldots, X_N = i_N) &= \frac{P(X_0 = i_0, \ldots, X_N = i_N, X_{N+1} = i_{N+1})}{P(X_0 = i_0, \ldots, X_N = i_N)}\\
&= \frac{\lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{N-1} i_N} p_{i_N i_{N+1}}}{\lambda_{i_0} p_{i_0 i_1} \cdots p_{i_{N-1} i_N}}\\
&= p_{i_N i_{N+1}}
\end{aligned}$$

Thus, by definition, $(X_n)_{0\le n\le N}$ is Markov(λ, P). □

The next theorem emphasizes the memorylessness of Markov chains. In the formulation of this theorem, we use the idea of the *unit mass at i*. It is denoted as $\delta_i = (\delta_{ij})$ where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Theorem 1.4. *Let $(X_n)_{n\ge 0}$ be Markov(λ, P). Then, given that $X_m = i$, $(X_l)_{l\ge m}$ is Markov(δ_i, P) and is independent of X_k, $0 \le k < m$.*

Proof. Let the event $A = \{X_m = i_m, \ldots, X_n = i_n\}$ and the event B be any event determined by $X_0, \ldots, X_m$. To prove the theorem, we must show that

$$P(A \cap B \,|\, X_m = i) = \delta_{i i_m} p_{i_m i_{m+1}} \cdots p_{i_{n-1} i_n} P(B \,|\, X_m = i)$$

thus the result follows from Theorem 1.2. First, let us consider any elementary event

$$B = B_k = \{X_0 = i_0, \ldots, X_m = i_m\}$$

Here we show that

$$P(A \cap B_k \text{ and } i = i_m \,|\, X_m = i) = \frac{\delta_{i i_m} p_{i_m i_{m+1}} \cdots p_{i_{n-1} i_n} P(B_k)}{P(X_m = i)}$$

which follows from Theorem 1.2 and the definition of conditional probability. Any event, B, determined by $X_0, \ldots, X_m$ can be written as a disjoint union of elementary events, $B = \bigcup_{k=1}^{\infty} B_k$. Thus, we can prove our above identity by summing up all of the different B_k for any given event. □

An additional idea that is going to be important later is the idea of conditioning on the initial state, X_0. We will let $P(A \,|\, X_0 = i) = P_i(A)$. Similarly, we will let $E(A \,|\, X_0 = i) = E_i(A)$.

2. Stopping Times and the Strong Markov Property

We start this section with the definition of a stopping time.

Definition 2.1. A random variable T is called a *stopping time* if the event $\{T = n\}$ depends only on $X_0, \ldots, X_n$ for $n = 0, 1, 2, \ldots$.

An example of a stopping time would be the *first passage time*

$$T_i = \inf\{n \geq 1 \,|\, X_n = i\}.$$

where we define $\inf \emptyset = \infty$. This is a stopping time since $\{T_i = n\} = \{X_k \neq i, X_n = i \,|\, 0 < k < n\}$. Now we will define an expansion of this idea that we will use later.

Definition 2.2. The *rth passage time* $T_i^{(r)}$ to state i is defined recursively using the first passage time.

$$T_i^{(0)} = 0, \qquad T_i^{(1)} = T_i$$

and, for $r = 1, 2, \ldots$,

$$T_i^{(r+1)} = \inf\{n \geq T_i^{(r)} + 1 \,|\, X_n = i\}.$$

This leads to the natural definition of the *length of the rth excursion* to i as

$$S_i^{(r)} = \begin{cases} T_i^{(r)} - T_i^{(r-1)} & \text{if } T_i^{(r-1)} < \infty \\ 0 & \text{otherwise.} \end{cases}$$

The following theorem shows how the Markov property holds at stopping times.

Theorem 2.3. *Let T be a stopping time of $(X_n)_{n\geq 0}$ which is Markov(λ, P). Then given $T < \infty$ and $X_T = i$, $(X_l)_{l\geq T}$ is Markov(δ_i, P) and independent of X_k, $0 \leq k < T$.*

Proof. First, we already have that $(X_l)_{l\geq T}$ is Markov(δ_i, P) by Theorem 1.4, so we just need to show the independence condition. Let the event $A = \{X_T = i_0, \ldots, X_{T+n} = i_n\}$ and the event B be any event determined by $X_0, \ldots, X_T$. It is important to notice that the event $B \cap \{T = m\}$ is determined by $X_0, \ldots, X_m$. We get that

$$P(A \cap B \cap \{T = m\} \cap \{X_T = i\}) = P_i(X_0 = i_0, \ldots, X_n = i_n) P(B \cap \{T = m\} \cap \{X_T = i\})$$

If we now sum over $m = 0, 1, 2, \ldots$ and divide each side by $P(T < \infty, X_T = i)$ using the definition of conditional probability, we obtain

$$P(A \cap B \,|\, T < \infty, X_T = i) = P_i(X_0 = i_0, \ldots, X_n = i_n) P(B \,|\, T < \infty, X_T = i)$$

which gives us the independence we desired. $\square$

3. Recurrence and Transience

Definition 3.1. Let $(X_n)_{n\geq 0}$ be Markov with transition matrix P. We say that a state i is *recurrent* if

$$P_i(X_n = i \text{ for infinitely many } n) = 1,$$

and we say that a state i is *transient* if

$$P_i(X_n = i \text{ for infinitely many } n) = 0.$$

The following results allow us to show that any state is necessarily either recurrent or transient.

Lemma 3.2. *For $r = 2, 3, \ldots$, given that $T_i^{(r-1)} < \infty$, $S_i^{(r)}$ is independent of X_k, $0 \leq k \leq T_i^{(r-1)}$ and*

$$P(S_i^{(r)} = n \,|\, T_i^{(r-1)} < \infty) = P_i(T_i = n).$$

Proof. We can directly apply Theorem 2.3 where $T_i^{(r-1)}$ is the stopping time T, since it is assured that $X_T = i$ when $T < \infty$. So, given that $T_i^{(r-1)} < \infty$, $(X_l)_{l\geq T}$ is Markov(δ_i, P) and independent of X_k, $0 \leq k < T$, the independence wanted. Yet, we know

$$S_i^{(r)} = \inf\{l - T \geq 1 \,|\, X_l = i\}$$

so $S_i^{(r)}$ is the first passage time of $(X_l)_{l\geq T}$ to state i, giving us our desired equality. □

Definition 3.3. The idea of the *number of visits to* i, V_i, is intuitive and can be easily defined using the indicator function

$$V_i = \sum_{n=0}^{\infty} 1_{\{X_n = i\}}$$

A nice property of V_i is that

$$E_i(V_i) = E_i\Big(\sum_{n=0}^{\infty} 1_{\{X_n=i\}}\Big) = \sum_{n=0}^{\infty} E_i(1_{\{X_n=i\}}) = \sum_{n=0}^{\infty} P_i(X_n = i).$$

Definition 3.4. Another intuitive and useful term is the *return probability to* i, defined as

$$f_i = P_i(T_i < \infty).$$

Lemma 3.5. $P_i(V_i > r) = (f_i)^r$ *for* $r = 0, 1, 2, \ldots$.

Proof. First, we know that our claim is necessarily true when $r = 0$. Thus, we can use induction and the fact that if $X_0 = i$ then $\{V_i > r\} = \{T_i^{(r)} < \infty\}$ to conclude that

$$\begin{aligned} P_i(V_i > r+1) &= P_i(T_i^{(r+1)} < \infty) \\ &= P_i(T_i^{(r)} < \infty \text{ and } S_i^{(r+1)} < \infty) \\ &= P_i(S_i^{(r+1)} < \infty \,|\, T_i^{(r)} < \infty) P_i(T_i^{(r)} < \infty) \\ &= f_i \cdot (f_i)^r = (f_i)^{r+1} \end{aligned}$$

using Lemma 3.2, so our claim is true for all r. □

Theorem 3.6. *The following two cases hold and show that any state is either recurrent or transient:*

1. *if $P_i(T_i < \infty) = 1$, then i is recurrent and $\sum_{n=0}^{\infty} P_i(X_n = i) = \infty$;*
2. *if $P_i(T_i < \infty) < 1$, then i is transient and $\sum_{n=0}^{\infty} P_i(X_n = i) < \infty$.*

Proof. If $P_i(T_i < \infty) = f_i = 1$ by Lemma 3.5, then

$$P_i(V_i = \infty) = \lim_{r\to\infty} P_i(V_i > r) = \lim_{r\to\infty} 1^r = 1$$

so i is recurrent and

$$\sum_{n=0}^{\infty} P_i(X_n = i) = E_i(V_i) = \infty.$$

In the other case, $f_i = P_i(T_i < \infty) < 1$ then using our fact about V_i

$$\sum_{n=0}^{\infty} P_i(X_n = i) = E_i(V_i) = \sum_{n=1}^{\infty} nP_i(V_i = n) = \sum_{n=1}^{\infty}\sum_{r=0}^{n-1} P_i(V_i = n)$$

$$= \sum_{r=0}^{\infty}\sum_{n=r+1}^{\infty} P_i(V_i = n) = \sum_{r=0}^{\infty} P_i(V_i > r) = \sum_{r=0}^{\infty} (f_i)^r = \frac{1}{1-f_i} < \infty$$

so $P_i(V_i = \infty) = 0$ and i is transient. □

4. Communication Classes and Recurrence

Definition 4.1. State i *can send to* state j, and we write $i \to j$ if

$$P_i(X_n = j \text{ for some } n \geq 0) > 0.$$

Also i *communicates with* j, and we write $i \leftrightarrow j$ if both $i \to j$ and $j \to i$.

Theorem 4.2. *For distinct states $i, j \in I$, $i \to j \iff p_{ii_1}p_{i_1i_2}\cdots p_{i_{n-1}j} > 0$ for some states $i_1, i_2, \ldots, i_{n-1}$. Also, $\leftrightarrow$ is an equivalence relation on I.*

Proof. ($\Rightarrow$)

$$0 < P_i(X_n = j \text{ for some } n \geq 0) \leq \sum_{n=0}^{\infty} P_i(X_n = j) = \sum_{n=0}^{\infty}\sum_{i_1,\ldots,i_{n-1}} p_{ii_1}p_{i_1i_2}\cdots p_{i_{n-1}j}$$

Thus, for some $p_{ii_1}p_{i_1i_2}\cdots p_{i_{n-1}j} > 0$ for some states $i_1, i_2, \ldots, i_{n-1}$.
($\Leftarrow$) Take some $i_1, i_2, \ldots, i_{n-1}$ such that

$$0 < p_{ii_1}p_{i_1i_2}\cdots p_{i_{n-1}j} \leq P_i(X_n = j) \leq P_i(X_n = j \text{ for some } n \geq 0).$$

Now it is clear from the proven inequality that $i \to j, j \to k \Rightarrow i \to k$. Also, it is true that $i \leftrightarrow i$ for any state i and that $i \leftrightarrow j \Rightarrow j \leftrightarrow i$. Thus, $\leftrightarrow$ is an equivalence relation on I. □

Definition 4.3. We say that $\leftrightarrow$ partitions I into *communication classes*. Also, a Markov chain or transition matrix P where I is a single communication class is called *irreducible*.

Theorem 4.4. *Let C be a communication class. Either all states in C are recurrent or all are transient.*

Proof. Take any distinct pair of states $i, j \in C$ and suppose that i is transient. Then there exist $n, m \geq 0$ such that $P_i(X_n = j) > 0$ and $P_j(X_m = i) > 0$, and for all $r \geq 0$

$$P_i(X_{n+r+m} = i) \geq P_i(X_n = j)P_j(X_r = j)P_j(X_m = i).$$

This implies that

$$\sum_{r=0}^{\infty} P_j(X_r = j) \leq \frac{1}{P_i(X_n = j)P_j(X_m = i)} \sum_{r=0}^{\infty} P_i(X_{n+r+m} = i) < \infty$$

by Theorem 3.6. So any arbitrary j is transient, again by Theorem 3.6, so the whole of C is transient. The only way for this not to be true is if all states in C are recurrent. □

This theorem shows us that recurrence and transience is a class property, and we will refer to it in the future as such.

Theorem 4.5. *Suppose P is irreducible and recurrent. Then for all $i \in I$ we have $P(T_i < \infty) = 1$.*

Proof. By Theorem 2.3 we have

$$P(T_i < \infty) = \sum_{j \in I} P_j(T_i < \infty)P(X_0 = j)$$

so we only need to show $P_j(T_i < \infty) = 1$ for all $j \in I$. By the irreducibility of P, we can pick an m such that $P_i(X_m = j) > 0$. From Theorem 3.6, we have

$$\begin{aligned}
1 &= P_i(X_n = i \text{ for infinitely many } n) \\
&= P_i(X_n = i \text{ for some } n \geq m+1) \\
&= \sum_{k \in I} P_i(X_n = i \text{ for some } n \geq m+1 \mid X_m = k)P_i(X_m = k) \\
&= \sum_{k \in I} P_k(T_i < \infty)P_i(X_m = k)
\end{aligned}$$

using Theorem 2.3 again. Since $\sum_{k \in I} P_i(X_m = k) = 1$ so we have that $P_j(T_i < \infty) = 1$. □

5. The Strong Law of Large Numbers and the Ergodic Theorem

The Strong Law of Large Numbers will be presented here without proof although an elementary proof needing much more background can be found in the journal *Wahrscheinlichketstheorie* by N. Etemadi (1981).

Theorem 5.1 (Strong Law of Large Numbers)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of independent, identically distributed, non-negative, random variables where $E(Y_k) = \mu$. Then*

$$P\left(\frac{Y_1 + \ldots + Y_n}{n} \to \mu \text{ as } n \to \infty\right) = 1$$

Definition 5.2. We will expand on the idea of V_i by defining

$$V_i(n) = \sum_{k=0}^{n-1} 1_{\{X_k = i\}},$$

representing the number of visits to i before step n.

The Ergodic theorem examines the long-run value of the ratio $V_i(n)/n$, which is the proportion of time spent in state i before step n.

Theorem 5.3 (Ergodic Theorem)**.** *Let P be irreducible and let λ be any distribution. If $(X_n)_{0\leq n\leq N}$ is Markov(λ, P) then*

$$P\left(\frac{V_i(n)}{n} \to \frac{1}{m_i} \text{ as } n \to \infty\right) = 1$$

where $m_i = E_i(T_i)$, i.e. the expected return time to state i.

Proof. First, we will consider the case where P is transient. We then know that with probability 1, $V_i < \infty$, so

$$\frac{V_i(n)}{n} \leq \frac{V_i}{n} \to 0 = \frac{1}{m_i}.$$

Now, we will consider the case that P is recurrent, and we will fix a state i. Let $T = T_i$, and we have $P_i(T < \infty) = 1$ by Theorem 3.6 and $(X_l)_{l\geq T}$ is Markov(δ_i, P) and independent of X_k, $0 \leq k < T$ by Theorem 2.3. Since the long-run proportion of time spent in state i is the same for $(X_l)_{l\geq T}$ as for $(X_n)_{n\geq 0}$, we need to only consider the case where $\lambda = \delta_i$.

By Lemma 3.2, we know that the non-negative, random variables $S_i^{(1)}, S_i^{(2)}, \ldots$ are independent and identically distributed with $E_i(S_i^{(r)}) = E_i(T_i) = m_i$. We know

$$T_i^{(V_i(n)-1)} = S_i^{(1)} + \ldots + S_i^{(V_i(n)-1)} \leq n - 1,$$

where the left-hand side of the inequality is the step of the last visit to i before step n. In addition,

$$T_i^{(V_i(n))} = S_i^{(1)} + \ldots + S_i^{(V_i(n))} \geq n,$$

where the left-hand side of the inequality is the step of the first visit to i after step $n - 1$. So we can squeeze the value of $n/V_i(n)$ using the following inequality:

$$\frac{S_i^{(1)} + \ldots + S_i^{(V_i(n)-1)}}{(V_i(n))} \leq \frac{n}{V_i(n)} \leq \frac{S_i^{(1)} + \ldots + S_i^{(V_i(n))}}{(V_i(n))} \tag{5.4}$$

Now we can use the Strong Law of Large Numbers to get

$$P\left(\frac{S_i^{(1)} + \ldots + S_i^{(n)}}{n} \to m_i \text{ as } n \to \infty\right) = 1,$$

and since P is recurrent,

$$P(V_i(n) \to \infty \text{ as } n \to \infty) = 1.$$

Thus, if we let $n \to \infty$ in (5.4), we get

$$P\left(\frac{n}{V_i(n)} \to m_i \text{ as } n \to \infty\right) = 1,$$

which implies that

$$P\left(\frac{V_i(n)}{n} \to \frac{1}{m_i} \text{ as } n \to \infty\right) = 1,$$

□

References

[1] Norris, James R. *Markov Chains.* University of Cambridge Press. 1998.