FINITE MARKOV CHAINS

BEN NORDSTROM

ABSTRACT. This paper provides basic information and theorems about finite Markov chains. The inspiration for this paper came from Professor Laci Babai's Discrete Mathematics lecture notes from the REU program of 2003. After reading through his chapter on Markov chains, I decided to proceed by answering as many exercises from the notes as possible. Below is what I have finished.

CONTENTS

1. FORMAL DEFINITION AND BASIC PROPERTIES

First we will start with the formal definition of a Markov chain and then we will go through some basic properties.

A *Markov chain* is a memoryless, homogeneous, stochastic process with a finite number of states. A *process* is a system that changes after each time step t, and a *stochastic process* is a process in which the changes are random. The states are labelled by the elements of the set $[n] = \{1, \ldots, n\}$, with X_t denoting the state at time t. If the process undergoes the transition $i \to j$ at time t, this is indicated by $X_t = i$ and $X_{t+1} = j$.

The transition data of a Markov chain is encoded by an $n \times n$ transition matrix $T = (p_{ij})$, where $p_{ij} = P(X_{t+1} = j | X_t = i)$ is the probability of transitioning from state i to state j. The initial distribution is given by an $1 \times n$-vector $q(0) = (q_{01}, \ldots, q_{0n})$, where $q_{0i} = P(X_0 = i)$ is the probability that i is the initial state. Note that $\sum_{i=1}^{n} q_{0i} = 1$; that is, we start at some vertex with probability 1.

A process is *memoryless* if the probability of an $i \to j$ transition does not depend on the history of the process. A process is *homogeneous* if it does not depend on the time t.

Definition 1.1. An $n \times n$ matrix $T = (a_{ij})$ is *stochastic* if its entries are nonnegative real numbers and the sum of each row is constant. That is, $\sum_{j=1}^{n} a_{ij} = c$ where $c \in \mathbb{R}$.

More simply, T being a stochastic matrix means that we will traverse to another vertex with probability one. Recall that a directed graph G is a pair (V, E) where V

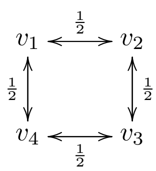
Date: AUGUST 22, 2008.

is the set of vertices and $E \subseteq V \times V$ is a set of ordered pairs of vertices representing edges. We have $E \subseteq V \times V$ which we present in an $n \times n$ matrix called the adjacency matrix. The data of a directed graph is presented by an $n \times n$ adjacency matrix $A = (a_{ij})$ with $a_{ij} = 0$ or 1 depending on whether there is an edge $i \to j$ or not. However, a Markov chain transition matrix does more than just tell you if there is an edge between two vertices; it also gives the probability of traversing that edge. So, instead of assigning a 1 to represent an edge, we assign some *weight* $p \in [0, 1]$.

A Markov chain may be represented by a directed graph in which the vertices correspond to the states of the Markov chain. For example, a directed cycle of length n corresponds to a Markov chain with n states. In this case the weight on each edge is 1 because there is only one direction you can travel at every state. Let's look at an example:

Example 1.2. Consider a person on a square where the corners are the vertices and the lines are the edges connecting the vertices as in the following diagram:

(1.3)



Suppose this person starts at v_1, i.e., $q(0) = (1, 0, 0, 0)$, and flips a coin to decide between going one way or the other way. So the probability of going to v_2 is $\frac{1}{2}$ and for v_4 is $\frac{1}{2}$. After n steps, regardless of where our person has been, the probability of the going one of the two possible directions is still $\frac{1}{2}$ and this *only* depends on the current state. So $P(X_n = v_j) = P(X_n = v_j | X_{n-1} = v_i)$ where $i \to j$. Note that the transition matrix for the diagram above is represented as $\begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$.

2. Results

Remark 2.1. The matrix representing a Markov chain is stochastic, with every row summing to 1.

Before proceeding with the next result I provide a generalized version of the theorem.

Proposition 2.2. *The product of two $n \times n$ stochastic matrices is a stochastic matrix.*

Proof. Let $A = (a_{ij})$ and $B = (b_{ij})$ be $n \times n$ stochastic matrices where $\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} b_{ij} = 1$. We know that $(A \cdot B)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = (a_{i1} b_{1j} + \ldots + a_{in} b_{nj})$.

Take:

$$\begin{aligned}\sum_{j=1}^{n}(A\cdot B)_{ij} &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{ik}b_{kj}\\ &= a_{i1}b_{11}+\ldots+a_{in}b_{n1}+\ldots+a_{in}b_{nn}\\ &= a_{i1}(b_{11}+\ldots+b_{1n})+\ldots+a_{in}(b_{n1}+\ldots+b_{nn})\\ &= a_{i1}+\ldots+a_{in}=1.\end{aligned}$$

Which proves the claim.

□

Theorem 2.3. *If T is a stochastic matrix then T^k is a stochastic matrix for all k.*

Proof. Again, we will proceed by induction. Our first case is when $k = 1$ which is trivial. Assume T^{k-1} is a stochastic matrix. We know that T and T^{k-1} are stochastic, so $T^k = T^{k-1} \cdot T$ is also stochastic. □

Let's look at the initial distribution.

Theorem 2.4. *If $q(0) = (q_{01}, \ldots, q_{0n})$ is an initial distribution for a stochastic process with transition matrix T, then the distribution at time t is $q(t) = q(0) \cdot T^t$.*

Proof. We proceed by induction. The first case when $t = 0$ is trivial. Assume that $q(k-1) = q(0) \cdot T^{k-1}$ and look at $q(k-1) \cdot T$. From our hypothesis we have that $q(k-1) \cdot T = q(0)T^{k-1} \cdot T = q(0) \cdot T^k$. We must show that $q(k) = q(k-1) \cdot T$. We have:

$$\begin{aligned}(q(k-1)\cdot T)_i &= \sum_{j=1}^{n} q(k-1)_j \cdot p_{ji}\\ &= \sum_{j=1}^{n} P(X_{k-1}=j)\cdot P(X_k=i|X_{k-1}=j)\\ &= \sum_{j=1}^{n} P(X_k=i|X_{k-1}=j)\\ &= q(k)_i,\end{aligned}$$

which proves the claim. □

This means that the distribution of the states after k steps can be determined by taking the initial distribution and multiplying it by the transition matrix raised to the k^{th} power. Thus $(T^k)_{ij}$ represents the probability of being in state j after k steps given that the starting state is i.

3. Eigenvectors and eigenvalues

Definition 3.1. A left *eigenvector* of an $n \times n$ matrix T is a $1 \times n$ vector $x \neq \mathbf{0}$ such that $x \cdot T = \lambda \cdot x$ for some complex number λ called the *eigenvalue* corresponding to the eigenvector x. Similarly, a right eigenvector of T is an $n \times 1$ vector $y \neq \mathbf{0}$ such that $T \cdot y = \mu \cdot y$ for some complex number μ which is again called an eigenvalue.

Remark 3.2. $\mathbf{0} = (0, \ldots, 0)$ is never an eigenvector.

Remark 3.3. The left and right eigenvalues of an $n \times n$ matrix are the same. The left and right eigenvectors, however, are not.

Theorem 3.4. *If T is an $n \times n$ stochastic matrix (in which the rows sum to 1) then $\lambda = 1$ is a right eigenvalue.*

Proof. We are solving for x in the equation $T \cdot x = x$, where x is an eigenvector to the eigenvalue 1. Let x be a vector such that $x_i = 1$ for all $i = 1, \ldots, n$. Then $(T \cdot x)_i = \sum_{j=1}^n a_{ij} x_j = \sum_{j=1}^n a_{ij} 1 = 1 = x_i$. □

Consequently $\lambda = 1$ is also a left eigenvalue.

Theorem 3.5. *If T is an $n \times n$ matrix and λ and μ are left and right eigenvalues to eigenvectors x and y (respectively) with $\lambda \neq \mu$, then $x \cdot y = 0$; i.e., x and y are orthogonal.*

Proof. Consider the product

$$x \cdot T \cdot y.$$

Using associativity and the fact that x and y are left/right eigenvectors, we know that:

$$\begin{aligned} (x \cdot T) \cdot y &= x \cdot (T \cdot y) \\ \lambda x \cdot y &= \mu x \cdot y \\ (\lambda - \mu) x \cdot y &= 0. \end{aligned}$$

Since $\lambda \neq \mu$ it follows that $x \cdot y = 0$. □

Theorem 3.6. *If $\lambda \in \mathbb{C}$ is an eigenvalue of a stochastic matrix $T = (p_{ij})$ then $|\lambda| \leq 1$.*

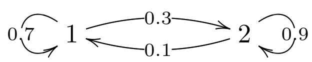
Proof. Let x be a right eigenvector corresponding to eigenvalue λ and let $x_k := \max_{i \in [n]} x_i$. Since $T \cdot x = \lambda \cdot x$ it follows that,

$$\begin{aligned} (\lambda \cdot x)_k &= (T \cdot x)_k \\ &= (p_{k1} x_1 + \ldots + p_{kn} x_n) \\ \text{It follows that } |\lambda| \cdot |x_k| &= |p_{k1} x_1 + \ldots + p_{kn} x_n| \\ &\leq |p_{k1} x_1| + \ldots + |p_{kn} x_n| \\ &= p_{k1} |x_1| + \ldots + p_{kn} |x_n| \\ &\leq |x_k| \cdot (p_{k1} + \ldots + p_{kn}) \\ &= |x_k|, \end{aligned}$$

which proves the claim. □

Let's look at an example:

Example 3.7. Consider the following 2-vertex graph.



The transition matrix for the graph above is $T = \begin{pmatrix} 0.7 & 0.3 \\ 0.1 & 0.9 \end{pmatrix}$. Observe that large powers of T appear to converge to $\begin{pmatrix} 0.25 & 0.75 \\ 0.25 & 0.75 \end{pmatrix}$. To find eigenvalues for T we will use the characteristic polynomial $f_A(x) = \det(xI - A) = \begin{vmatrix} x-0.7 & -0.3 \\ -0.1 & x-0.9 \end{vmatrix} =$ $x^2 - 1.6x + 0.6 = (x-1)(x-0.6)$. So the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = \frac{3}{5}$. Let's try and solve for the right eigenvectors $v = (v_1, v_2)$ and $v' = (v'_1, v'_2)$. For v we have that $T \cdot v = 1 \cdot v$ which gives rise to two equations:

$$\begin{aligned} 0.7v_1 + 0.3v_2 &= v_1 \\ 0.1v_1 + 0.9v_2 &= v_2 \\ 0.1v_1 &= 0.1v_2 \\ v_1 &= v_2. \end{aligned}$$

So we have $v = c(1,1)$ where $c \in \mathbb{R}$. To solve for $T \cdot v' = \frac{3}{5} \cdot v'$:

$$\begin{aligned} 0.7v'_1 + 0.3v'_2 &= 0.6v'_1 \\ 0.1v'_1 + 0.9v'_2 &= 0.6v'_2 \\ 0.1v'_1 &= -0.3v'_2 \\ v'_1 &= -3v'_2. \end{aligned}$$

Similarly to above, we have $v' = c(-3,1)$, where $c \in \mathbb{R}$.

For the the left eigenvectors we will do the same thing, but also normalize the eigenvector to eigenvalue 1; i.e., $|x_1| + |x_2| = 1$ where $x = (x_1, x_2)$ is the left eigenvector to eigenvalue 1. So our first set of equations for x are:

$$\begin{aligned} 0.7x_1 + 0.1x_2 &= x_1 \\ 0.3x_1 + 0.9x_2 &= x_2 \\ |x_1| + |x_2| &= 1 \\ 0.1x_2 &= 0.3x_1 \\ x_2 &= 3x_1 \\ \Rightarrow |x_1| + |3x_1| &= 1 \\ 4|x_1| &= 1 \\ |x_1| &= \frac{1}{4} \\ \Rightarrow |x_2| &= \frac{3}{4}. \end{aligned}$$

Observe that the vector $(|x_1|, |x_2|) = (\frac{1}{4}, \frac{3}{4})$ represents the rows of the matrix that T^k converges to. This will come up later.

Lemma 3.8. *Let T be a 2×2 stochastic matrix whose rows sum to 1. Observe that T is of the form $T = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$. For all $n \geq 1$ we have:*

$$T^n = \frac{1}{a+b} \begin{pmatrix} b & a \\ b & a \end{pmatrix} + \frac{(1-a-b)^n}{a+b} \begin{pmatrix} a & -a \\ -b & b \end{pmatrix}.$$

Proof. We must show that the formula for T holds for $n = 1$. I will only prove this for the $1, 1$ entry.

$$\begin{aligned}(T^1)_{1,1} &= \frac{b}{a+b} + \frac{a(1-a-b)^1}{a+b} \\ &= \frac{b+a-a^2-ab}{a+b} = \frac{a(1-a)+b(1-a)}{a+b} \\ &= \frac{(1-a)(a+b)}{a+b} = (1-a).\end{aligned}$$

Suppose it holds for $n = k - 1$. Now we must show that it holds for k.

$$\begin{aligned}\text{We know } (T^k)_{1,1} &= (T^{k-1} \cdot T)_{1,1} \\ &= \frac{(1-a)(b+a(1-a-b)^{k-1}) + ab - ab(1-a-b)^{k-1}}{a+b} \\ &= \frac{b + ((1-a-b)^{k-1})(a-a^2-ab)}{a+b} \\ &= \frac{b+a(1-a-b)^k}{a+b},\end{aligned}$$

which proves the claim. □

Proposition 3.9. *Under the hypotheses of Lemma 3.8, powers of T converge to the matrix* $\frac{1}{a+b}\begin{pmatrix} b & a \\ b & a \end{pmatrix}$.

Proof. Using Lemma 3.8 we can write

$$\lim_{n\to\infty} T^n = \frac{1}{a+b}\begin{pmatrix} b & a \\ b & a \end{pmatrix} + \lim_{n\to\infty} \frac{(1-a-b)^n}{a+b}\begin{pmatrix} a & -a \\ -b & b \end{pmatrix}.$$

Since $|(1-a-b)| < 1$ it follows that $(1-a-b)^n \to 0$ as $n \to \infty$ and thus $\lim_{n\to\infty} T^n = \frac{1}{a+b}\begin{pmatrix} b & a \\ b & a \end{pmatrix}$. □

Remark 3.10. Not all transition matrices converge. If we look at $W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, notice that the powers of W alternate between $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$; so W never converges.

Definition 3.11. A stationary distribution for a Markov chain is a vector $q = (q_1, \ldots, q_n)$ such that $q_i \geq 0$ and $\sum_{i=1}^n q_i = 1$ which is also a left eigenvector to eigenvalue 1; i.e., $q \cdot T = q$.

Remark 3.12. Observe that the left eigenvector to eigenvalue 1 in Example 3.7 is the stationary distribution.

Remark 3.13. If at time t, the distribution $q(t)$ is stationary then $q(t) = q(t + x)$ for all $x \in \mathbb{N}$.

Proof. Suppose we are at time t in our transition step. From Theorem 2.4 we know that $q(t) = q(0) \cdot T^t$. Because $q(t)$ is a stationary distribution it follows that:

$$\begin{aligned} q(t) &= q(t) \cdot T \\ &= q(0) \cdot T^t \cdot T \\ &= q(0) \cdot T^{t+1} \\ &= q(t+1). \end{aligned}$$

Using the same argument it follows that $q(t+1) = q(t+2) = \ldots = q(t+x)$. □

Proposition 3.14. *Suppose T is a stochastic matrix. If $T^\infty = \lim_{t\to\infty} T^t$ exists, then every row of T^∞ is a stationary distribution.*

Proof. We must establish that $T \cdot T^\infty = T^\infty$. $T^\infty = \lim_{n\to\infty} T^n$, so $T \cdot T^\infty = T \cdot \lim_{n\to\infty} T^n = \lim_{n\to\infty} T^{n+1} = T^\infty$. The claim immediately follows from this argument. □

Notice that this holds for T from Example 3.7.

References

[1] László Babai. Discrete Mathematics Lecture Notes. 2003.
[2] Olle Häggström. Finite Markov Chains and Algorithmic Applications. Cambridge University Press. 2002.