TECHNICAL RESEARCH REPORT

A Proof of the Markov Chain Tree Theorem

by

V. Anantharam and P. Tsoucas

SYSTEMS RESEARCH CENTER

UNIVERSITY OF MARYLAND

COLLEGE PARK, MARYLAND 20742

A Proof of the Markov Chain Tree Theorem

V. Anantharam * *and P. Tsoucas* **

Abstract

Let X be a finite set, P a stochastic matrix on X, and $\overline{P} = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k$. Let $G = (X, E)$ be the weighted directed graph on X associated to P, with weights p_{ij}. An *arborescence* is a subset $a \subseteq E$ which has at most one edge out of every node, contains no cycles, and has maximum possible cardinality. The weight of an arborescence is the product of its edge weights. Let $\mathcal{A}$ denote the set of all arborescences. Let $\mathcal{A}_{ij}$ denote the set of all arborescences which have j as a root and in which there is a directed path from i to j. Let $\|\mathcal{A}\|$, resp. $\|\mathcal{A}_{ij}\|$, be the sum of the weights of the arborescences in $\mathcal{A}$, resp. $\mathcal{A}_{ij}$. The Markov chain tree theorem states that $\overline{p}_{ij} = \frac{\|\mathcal{A}_{ij}\|}{\|\mathcal{A}\|}$. We give a proof of this theorem which is probabilistic in nature.

Keywords : arborescence, Markov chain, stationary distribution, tree.

---

* Field of Statistics, Center for Applied Mathematics, and School of Electrical Engineering, Cornell University, Ithaca, NY 14853.

** Systems Research Center and Department of Electrical Engineering, University of Maryland, College Park, MD 20742.

Address all correspondence to first author.

A Proof of the Markov Chain Tree Theorem [*]

1. Introduction

Let X be a finite set of cardinality n, and P a stochastic matrix on X. Let $\mathbf{x} = (X_n, n \geq 0)$ denote the canonical Markov chain on X with transition matrix P. Let $G = (X, E)$ be the weighted directed graph with vertex set X associated to P. This means that given $i, j \in X$ there is a directed edge from i to j iff $p_{ij} > 0$, and this edge has weight p_{ij}.

An *arborescence* is a subset $a \subseteq E$ which has at most one edge out of every node, contains no cycles, and has maximum possible cardinality. The nodes which have outdegree 0 in the arborescence are called its *roots*. It is easy to see that if there are α communicating classes in $\mathbf{x}$, then every arborescence has precisely one root in each communicating class and $n - \alpha$ edges. In particular, if P is irreducible then every arborescence has precisely one root and $n - 1$ edges. For basic facts about the decomposition of the state space of a Markov chain into communicating classes and transient states, see e.g. Freedman, (1983), Section 1.4.

The *weight* of an arborescence is the product of its edge weights. Let $\mathcal{A}$ denote the set of all arborescences and $\|\mathcal{A}\|$ the sum of the weights of the arborescences in $\mathcal{A}$. Let $\mathcal{A}_j$ denote the set of all arborescences which have j as a root, and $\|\mathcal{A}_j\|$ the sum of the weights of the arborescences in $\mathcal{A}_j$. Let $\mathcal{A}_{ij}$ denote the set of all arborescences in $\mathcal{A}_j$ in which there is a directed path from i to j, and $\|\mathcal{A}_{ij}\|$ the sum of the weights of the arborescences in $\mathcal{A}_{ij}$. We take $\mathcal{A}_{jj}$ to mean $\mathcal{A}_j$.

If the Markov chain $\mathbf{x}$ is started in the state $i \in X$, then it is well known that the long run average number of visits to any state j converges to a number $\bar{p}_{ij}$ given by the ij entry of

$$\overline{P} = \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k .$$

The existence of this limit is standard, see e.g. Freedman, (1983), Section 1.7. If $\mathbf{x}$ is irreducible all rows of the limit are identical and give the unique initial distribution from which $\mathbf{x}$ is stationary.

It turns out that there is a way to compute the entries of $\overline{P}$ in terms of the weights of arborescences in G. For irreducible P this fact appears to have been to originally discovered in the context of certain models for biological systems, see Kohler and Vollmerhaus, (1980), where it is called the diagram method, and attributed to Hill, (1966). It was also independently discovered by Shubert, (1975). This technique was extended to general Markov chains by Leighton and Rivest, (1983) and (1986), who call it the Markov chain tree theorem :

Theorem : Let the stochastic matrix P on the finite state space X determine the Markov

chain **x** with long run transition matrix $\overline{P}$. Then

$$\overline{p}_{ij} = \frac{\|\mathcal{A}_{ij}\|}{\|\mathcal{A}\|} .$$

(1)

If P is irreducible, then

$$\overline{p}_{ij} = \frac{\|\mathcal{A}_j\|}{\|\mathcal{A}\|}$$

(2)

for all i.

$\ominus$

Remark : To be precise we must assume that $\mathcal{A} \neq \emptyset$, i.e., at least one of the states of **x** is not isolated, where a state is called isolated if it cannot be accessed from any other state. One can avoid this assumption if (1) and (2) are interpreted suitably in this situation.

At first sight there does not appear to be an intuitive reason why the long run transition probabilities should be related to arborescences in the underlying directed graph. In fact, all proofs of the theorem that have appeared in the literature are algebraic or combinatorial in nature, and none of them provides a clear probabilistic reason for this unexpected connection. The purpose of this letter is to provide a simple proof of the theorem which is probabilistic in nature and makes the connection between long run transition probabilities and arborescences seem natural.

2. Proof

The probabilistic idea of our proof works for irreducible chains. From this we will get the general theorem by additional arguments at the end of this section.

Suppose P is irreducible, and let $\overline{\mathbf{x}} = (X_n, -\infty < n < \infty)$ be the canonical two sided chain with the stationary distribution. The basic probabilistic idea is to construct from this chain, in a canonical fashion, an $\mathcal{A}$ valued process $\overline{\mathbf{y}} = (Y_n, -\infty < \infty)$ that is a function of the past at any time. Define $f : \overline{\mathbf{x}} \to \mathcal{A}$ as follows : The root of $f(\overline{\mathbf{x}})$ is X_0. To find out where any other state $i \in X$ attaches we look for its most recent occurrence before time 0 and attach it to the succeeding state at that time. Formally, let $\tau(i) = \sup\{m < 0 : X_m = i\}$, and for $i \neq X_0$ attach i to $X_{\tau(i)+1}$. Clearly f is well defined almost surely. Then we define

$$Y_n = f(T^n(\overline{\mathbf{x}})) , \quad -\infty < n < \infty ,$$

where T denotes the left shift operator, $T(\overline{\mathbf{x}})_n = X_{n+1}$.

We check that $\overline{\mathbf{y}}$ is Markov. Indeed, given $\{Y_{n+m}, m \leq 0\}$, Y_{n+1} is got from Y_n by the following procedure : Let Y_n have root i. To Y_n attach the directed edge (i, j) with probability p_{ij} This creates a unique directed loop which contains i and j (possibly a self loop at i). Delete the unique directed edge out of j which breaks this loop. The resulting arborescence rooted at j is Y_{n+1}. The reader can easily write down the formal details. Let us call this procedure the *forward* procedure.

It is easy to see that an arborescence $b \in \mathcal{A}$ can be constructed from an arborescence $a \in \mathcal{A}$ by the forward procedure iff a can be constructed from b by the following procedure, called the *reverse* procedure : Let b have root j. To b attach the directed edge (j, k). This creates a unique directed loop containing j and k (possibly a self loop at j). To break this loop delete the unique edge directed *into* j which lies in this loop. The result is an arborescence (whose root need not be either j or k). If b can be constructed from a by the forward procedure there is a k such that a can be constructed from b by the reverse procedure on attaching (j, k).

Let $\hat{\mathbf{y}} = (\hat{Y}_n, -\infty < n < \infty)$ denote the time reversal of the Markov process $\overline{\mathbf{y}}$. For the definition of time reversal see e.g. Ross, (1983), Section 4.7. Let j denote the root of $\hat{Y}_n$. Then $\hat{Y}_{n+1}$ is got from $\hat{Y}_n$ by attaching (j, k) to $\hat{Y}_n$ with probability p_{jk} and then applying the reverse procedure. This is immediately obvious from applying the following well known lemma, which is sometimes called Kelly's lemma, see e.g. Walrand, (1988), Lemma 2.8.5 :

Lemma : Let $\overline{\mathbf{y}}$ be a finite state Markov chain with stochastic transition matrix Q and state space $\mathcal{A}$. Suppose one can find a distribution π on $\mathcal{A}$ and a stochastic matrix $\tilde{Q}$ such that

$$\pi_a Q_{ab} = \pi_b \tilde{Q}_{ba} \tag{3}$$

for all $a, b \in \mathcal{A}$. Then the time reversal of $\mathbf{y}$ has transition matrix $\tilde{Q}$ and $\mathbf{y}$ has stationary distribution proportional to π.

$\ominus$

In our situation, let a and b be arborescences such that b can be got from a by the forward procedure and a from b by the reverse procedure. If a has root i and b has root j, then $Q_{ab} = p_{ij}$. To get b from a we first attached (i, j) and then deleted the unique outgoing edge (j, k) from j. To get a from b we would attach (j, k) to b and delete (i, j). A moment's thought shows that (3) holds when we take $\pi(a)$ and $\pi(b)$ to be the weights of the arborescences a and b respectively and $\tilde{Q}_{ba} = p_{jk}$.

This verifies that the time reversal of $\overline{\mathbf{y}}$ is indeed as described. However it also verifies that the stationary distribution of $\overline{\mathbf{y}}$ is proportional to the weights of arborescences. But $\overline{\mathbf{x}}$ is a function of $\overline{\mathbf{y}}$, given by the function that takes arborescences to their roots. Under this map the preimage of j is the set of arborescences which have j as a root, namely $\mathcal{A}_j$. the theorem for irreducible P follows immediately.

If P is reducible, write X as a disjoint union, $X = \cup_{\theta=1}^{\Theta} C_\theta \cup T$, where $C_\theta, 1 \leq \theta \leq \Theta$ are communicating classes and T is the set of transient states. If $j \in T$, then $\mathcal{A}_{ij} = \emptyset$ and $\overline{p}_{ij} = 0$ for each i by standard theory, see e.g. Freedman, (1983), Section 1.4, so the theorem holds for such j. If $i \in C_\theta$ and $j \in C_{\theta'}$ with $\theta \neq \theta'$, then again $\mathcal{A}_{ij} = \emptyset$ and $\overline{p}_{ij} = 0$ by standard theory, so the theorem holds for such i and j. If $i, j \in C_\theta$, then $\mathcal{A}_{ij} = \mathcal{A}_j$ by standard theory, so the theorem holds for such i and j as a direct consequence of the theorem for irreducible P.

It remains to consider the case $i \in T$, $j \in C_\theta$. To deal with this we proceed by induction on the number of transient states. We have already proved that the theorem is true for finite Markov chains without transient states. Let T have cardinality m and

4

suppose the theorem is true for all finite Markov chains with $m-1$ or less transient states. From **x** we construct a modified Markov chain with state space $Z = X\backslash\{i\} \cup \{i^0, i^\omega\}$ and with transition matrix P^* having the following description : Think of Z as constructed from X by splitting the state i into two states i^0 and i^ω. Every transition out of i in P is a transition out of i^0 in P^*, while every transition into i in P is a transition into i^ω in P^*. To make P^* stochastic we also introduce a transition from i^ω to itself, with probability 1.

z has m transient states and one more communicating class than **x**, namely $\{i^\omega\}$. Let $\hat{\mathcal{A}}$ denote the set of arborescences in the weighted directed graph associated to P^*. Similarly define $\hat{\mathcal{A}}_k$, and $\hat{\mathcal{A}}_{kl}$. Let $\|\hat{\mathcal{A}}\|$, $\|\hat{\mathcal{A}}_k\|$, and $\|\hat{\mathcal{A}}_{kl}\|$ stand for the sums of the weights of arborescences in $\hat{\mathcal{A}}$, $\hat{\mathcal{A}}_k$, and $\hat{\mathcal{A}}_{kl}$ respectively.

We note that there is a one to one weight preserving correspondence between $\mathcal{A}$ and $\hat{\mathcal{A}}\backslash\hat{\mathcal{A}}_{i^0 i^\omega}$. Given $a \in \mathcal{A}$, split i into i^0 and i^ω, peel off the portion of a flowing into i as being rooted at i^ω and think of the unique directed edge out of i as being out of i^0. Conversely, given $\hat{a} \in \hat{\mathcal{A}}\backslash\hat{\mathcal{A}}_{i^0 i^\omega}$, coalesce i^0 and i^ω into i to get an element of $\mathcal{A}$. Note that under this correspondence $\hat{\mathcal{A}}_{i^0 j}$ corresponds to $\mathcal{A}_{ij}$.

Let $\overline{P^*} = \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{k=0}^{n-1} P^{*k}$. Then for any $k \in \cup_{\theta=1}^{\Theta} C_\theta$, we clearly have

$$\overline{p^*}_{i^0 k} = \overline{p}_{ik}[1 - \overline{p^*}_{i^0 i^\omega}] \ .$$

This means that if i is fixed, $\overline{p}_{ik}$ is proportional to $\overline{p^*}_{i^0 k}$ as k runs over $\cup_{\theta=1}^{\Theta} C_\theta$. Now consider the Markov chain got from **z** by erasing the row and column corresponding to i^0. (The resulting matrix is stochastic because the column corresponding to i^0 is zero). This has $m-1$ transient states. Let $\tilde{\mathcal{A}}$ denote the set of arborescences for this chain. By ignoring the edge out of i^0 we get a map from $\hat{\mathcal{A}}$ to $\tilde{\mathcal{A}}$. The preimages of any $\tilde{a} \in \tilde{\mathcal{A}}$ under this map differ by the different ways of going out of i^0. From this observation, and by induction hypothesis we see immediately that

$$\overline{p^*}_{i^0 k} = \sum_l p_{il}\overline{p^*}_{lk} = \sum_l p_{il}\frac{\|\hat{\mathcal{A}}_{lk}\|}{\|\hat{\mathcal{A}}\|} = \frac{\|\hat{\mathcal{A}}_{i^0 k}\|}{\|\hat{\mathcal{A}}\|}$$
$$= \frac{\|\mathcal{A}_{ik}\|}{\|\hat{\mathcal{A}}\|} \ .$$

This means that if i is fixed, $\overline{p}_{ik}$ is proportional to $\|\mathcal{A}_{ik}\|$ as k runs over $\cup_{\theta=1}^{\Theta} C_\theta$. This concludes the proof of the theorem.

References

Freedman, D. (1983), *Markov Chains* (Springer-Verlag, New York).

Hill, T. L. (1966), Studies in Irreversible Thermodynamics IV. Diagrammatic Representation of Steady State Fluxes for Unimolecular Systems, *Journal of Theoretical Biology*, **10**, pp. 442 -459.

Kohler, H. H., and Vollmerhaus, E. (1980), The Frequency of Cyclic Processes in Biological Multistate Systems, *Journal of Mathematical Biology*, **9**, pp. 275 -290.

Leighton F. T., and Rivest, R. L. (1983), The Markov Chain Tree Theorem, *M.I.T. Laboratory for Computer Science Technical Report*, **MIT/LCS/TM-249**, 13 p.

Leighton, F. T., and Rivest, R. L. (1986), Estimating a Probability using Finite Memory, *IEEE Transactions on Information Theory*, **IT-32,6**, pp. 733 -742.

Ross, S. (1983), *Stochastic Processes* (Wiley, New York).

Shubert, B. O. (1975), A Flow-Graph Formula for the Stationary Distribution of a Markov Chain, *IEEE Transactions on Systems, Man and Cybernetics*, pp. 565 -566.

Walrand, J. (1988), *An Introduction to Queueing Networks* (Prentice-Hall, New Jersey).