Estimating a Probability Using Finite Memory

F. THOMSON LEIGHTON, MEMBER, IEEE, AND RONALD L. RIVEST

Abstract—Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent Bernoulli random variables with probability p that $X_i = 1$ and probability $q = 1 - p$ that $X_i = 0$ for all $i \geq 1$. Time-invariant finite-memory (i.e., finite-state) estimation procedures for the parameter p are considered which take $X_1, \cdots$ as an input sequence. In particular, an n-state deterministic estimation procedure is described which can estimate p with mean-square error $O(\log n/n)$ and an n-state probabilistic estimation procedure which can estimate p with mean-square error $O(1/n)$. It is proved that the $O(1/n)$ bound is optimal to within a constant factor. In addition, it is shown that linear estimation procedures are just as powerful (up to the measure of mean-square error) as arbitrary estimation procedures. The proofs are based on an analog of the well-known matrix tree theorem that is called the Markov chain tree theorem.

I. Introduction

LET $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent Bernoulli random variables with probability p that $X_i = 1$ and probability $q = 1 - p$ that $X_i = 0$ for all $i \geq 1$. Estimating the value of p is a classical problem in statistics. In general, an *estimation procedure* for p consists of a sequence of estimates $\{e_t\}_{t=1}^{\infty}$, where each e_t is a function of $\{X_i\}_{i=1}^{t}$. When the form of the estimation procedure is unrestricted, it is well-known that p is best estimated by

$$e_t = \frac{1}{t}\sum_{i=1}^{t} X_i.$$

As an example, consider the problem of estimating the probability p that a coin of unknown bias will come up heads. The optimal estimation procedure will, on the tth trial, flip the coin to determine X_t ($X_t = 1$ for heads and $X_t = 0$ for tails) and then estimate the proportion of heads observed in the first t trials.

The quality of an estimation procedure may be measured by its mean-square error $\sigma^2(p)$. The mean-square error of an estimation procedure is defined as

$$\sigma^2(p) = \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \sigma_i^2(p)$$

where

$$\sigma_t^2(p) = E\left((e_t - p)^2\right)$$

denotes the expected square error of the tth estimate. For example, it is well-known that $\sigma_t^2(p) = pq/t$ and $\sigma^2(p) = 0$ when $e_t = (1/t)\Sigma_{i=1}^{t} X_i$.

In this paper, we consider time-invariant estimation procedures which are restricted to use a finite amount of memory. A time-invariant finite-memory estimation procedure consists of a finite number of states $S = \{1, \cdots, n\}$, a start state $S_0 \in \{1, \cdots, n\}$, and a transition function τ which computes the state S_t at step t from the state S_{t-1} at step $t - 1$ and the input X_t according to

$$S_t = \tau(S_{t-1}, X_t).$$

In addition, each state i is associated with an estimate η_i of p. The estimate after the tth transition is then given by $e_t = \eta_{S_t}$. For simplicity, we will call a finite-state estimation procedure an "FSE."

As an example, consider the FSE shown in Fig. 1. This FSE has $n = (s + 1)(s + 2)/2$ states and simulates two counters: one for the number of inputs seen, and one for the number of inputs seen that are ones. Because of the finite-state restriction, the counters can count up to $s = \Theta(\sqrt{n})$ but not beyond. Hence all inputs after the sth input are ignored. On the tth step, the FSE estimates the proportion of ones seen in the first $\min(s, t)$ inputs. This is

$$e_t = \frac{1}{\min(s, t)}\sum_{i=1}^{\min(s, t)} X_i.$$

Hence the mean-square error of the FSE is $\sigma^2(p) = pq/s = O(1/\sqrt{n})$.

In [31], Samaniego considered probabilistic FSE's and constructed the probabilistic FSE shown in Fig. 2. Probabilistic FSE's are similar to nonprobabilistic (or deterministic) FSE's except that a probabilistic FSE allows probabilistic transitions between states. In particular, the transition function τ of a probabilistic FSE consists of probabilities τ_{ijk} that the FSE will make a transition from state i to state j on input k. For example, $\tau_{320} = 2/(n - 1)$

Manuscript received January 4, 1984; revised February 3, 1986. This work was supported in part by the Bantrell Foundation, in part by a National Science Foundation Presidential Young Investigator Award with matching funds from Xerox and IBM, and in part by the NSF under Grant MCS-8006938. A preliminary version of this paper was presented at the 1983 International Conference on Foundations of Computation Theory, Borgholm, Sweden.

F. T. Leighton is with the Mathematics Department and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

R. L. Rivest is with the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

IEEE Log Number 8609493.

0018-9448/86/1100-0733$01.00

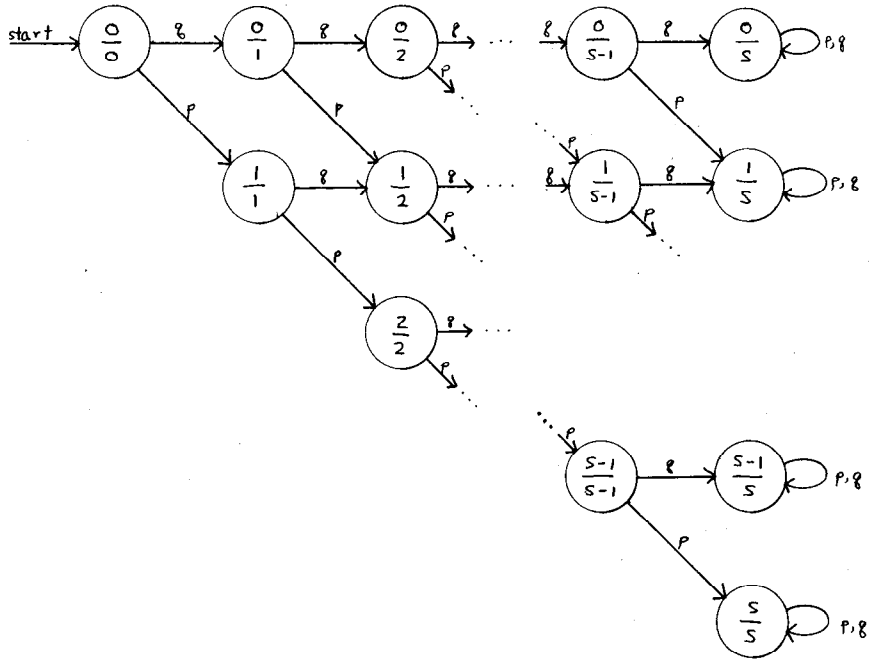

Fig. 1. $(s + 1)(s + 2)/2$-state deterministic FSE with mean-square error $\sigma^2(p) = pq/s$. States are represented by circles. Arrows labeled with q denote transitions on input zero. Arrows labeled with p denote transitions on input one. Estimates are given as fractions and represent proportion of inputs seen that are ones.

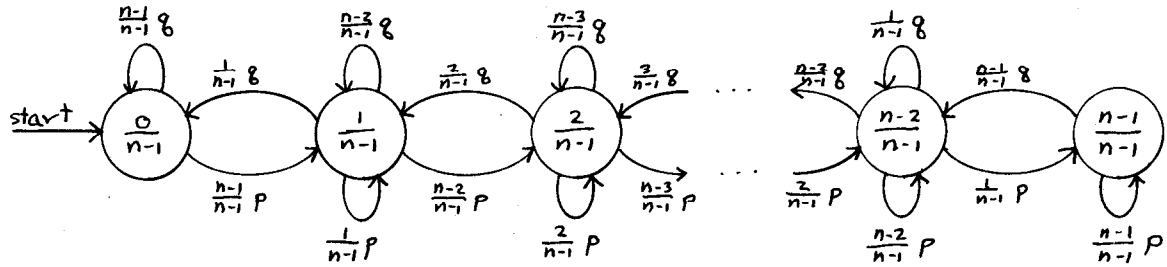

Fig. 2. Probabilistic n-state FSE with mean-square error $\sigma^2(p) = pq/(n - 1)$. States are represented by circles in increasing order from left to right (e.g., state 1 is denoted by leftmost circle and state n is denoted by rightmost circle). State i estimates $(i - 1)/(n - 1)$ for $1 \le i \le n$. The estimates are shown as fractions within circles. Arrows labeled with fractions of q denote probabilistic transitions on input zero. Arrows labeled with fractions of p denote probabilistic transitions on input one. For example, probability of changing from state 2 to state 3 on input 1 is $(n - 2)/(n - 1)$.

in Fig. 2. So that τ is well-defined, we require that $\sum_{j=1}^{n} \tau_{ijk} = 1$ for all i and k.

Samaniego [31] and others have shown that the mean-square error of the FSE shown in Fig. 2 is $\sigma^2(p) = pq/(n - 1) = O(1/n)$. In this paper, we prove that this method is the best possible (up to a constant factor) for an n-state FSE. In particular, we will show that for any n-state FSE (probabilistic or deterministic), some value of p exists for which $\sigma^2(p) = \Omega(1/n)$. Previously, the best lower bound known for $\sigma^2(p)$ was $\Omega(1/n^2)$. The weaker bound is due to the "quantization problem," which provides a fundamental limitation on the achievable performance of any FSE. Since the set of estimates of an n-state FSE has size n, there is always a value of p (in fact, there are many such values) for which the difference between p and the closest estimate is at least $1/2n$. This means that the mean-square error for some p must be at least $\Omega(1/n^2)$. Our result (which is based on an analog of the matrix tree theorem that we call the Markov chain tree theorem) proves that this bound is not achievable, thus showing that the quantization problem is not the most serious consequence of the finite-memory restriction.

It is encouraging that the nearly optimal FSE in Fig. 2 has such a simple structure. This is not a coincidence. In fact, we will show that for every probabilistic FSE with mean-square error $\sigma^2(p)$, there is a linear probabilistic FSE with the same number of states and with a mean-square error that is bounded above by $\sigma^2(p)$ for all p. (An FSE is said to be *linear* if the states of the FSE can be linearly ordered so that transitions are made only between consecutive states in the ordering. Linear FSE's are the easiest FSE's to implement in practice since the state information can be stored in a counter, and the transitions can be effected by a single increment or decrement of the counter.)

We also study deterministic FSE's in the paper. Although we do not know how to achieve the $\Theta(1/n)$ lower bound for deterministic FSE's, we can come close. In fact,

we will construct an n-state deterministic FSE that has mean-square error $O(\log n/n)$. The construction uses the input to deterministically simulate the probabilistic transitions of the FSE shown in Fig. 2.

The remainder of the paper is divided into sections as follows. In Section II, we present some background material on Markov chains and give a simple proof that the FSE shown in Fig. 2 has mean-square error $O(1/n)$. In Section III we construct an n-state deterministic FSE with mean-square error $O(\log n/n)$. The $\Omega(1/n)$ lower bound for n-state FSE's is proved in Section IV. In Section V, we demonstrate the universality of linear FSE's. In Section VI, we mention some related work and open questions. For completeness, we have included a proof of the Markov chain tree theorem in the Appendix.

II. Background Theory of Markov Chains

An n-state FSE acts like an n-state first-order stationary Markov chain. In particular, the transition matrix P defining the chain has entries

$$p_{ij} = \tau_{ij1}p + \tau_{ij0}q$$

where τ_{ijk} is the probability of changing from state i to state j on input k in the FSE. For example, $p_{33} = 2p/(n-1) + q(n-3)/(n-1)$ for the FSE in Fig. 2.

From the definition, we know that the mean-square error of an FSE depends on the limiting probability that the FSE is in state j given that it started in state i. (This probability is based on p and the transition probabilities τ_{ijk}.) The long-run transition matrix for the corresponding Markov chain is given by

$$\bar{P} = \lim_{t\to\infty} \frac{1}{t}(I + P + P^2 + \cdots + P^{t-1}).$$

This limit exists because P is stochastic (see [8, Theorem 2]). The ijth entry of $\bar{P}$ is simply the long-run average probability $\bar{p}_{ij}$ that the chain will be in state j given that it started in state i.

In the case that the Markov chain defined by P is ergodic, every row of $\bar{P}$ is equal to the same probability vector $\pi = (\pi_1 \cdots \pi_n)$ which is the stationary probability vector for the chain. In the general case, the rows of $\bar{P}$ may vary, and we will use π to denote the S_0th row of $\bar{P}$. Since S_0 is the start state of the FSE, π_i is the long-run average probability that the FSE will be in state i. Using the new notation, we can express the mean-square error of an FSE as

$$\sigma^2(p) = \sum_{i=1}^{n} \pi_i(\eta_i - p)^2.$$

Several methods are known for calculating long-run transition probabilities. For our purposes, the method developed by Leighton and Rivest in [21] is the most useful. This method is based on sums of weighted arborescences in the underlying graph of the chain. We review the method in what follows.

Let $V = \{1, \cdots, n\}$ be the nodes of a directed graph G, with edge set $E = \{(i, j) | p_{ij} \neq 0\}$. This is the usual directed graph associated with a Markov chain. (Note that G may contain self-loops.) Define the weight of edge (i, j) to be p_{ij}. An edge set $A \subseteq E$ is an *arborescence* if A contains at most one edge out of every node, has no cycles, and has maximum possible cardinality. The *weight* of an arborescence is the product of the weights of the edges it contains. A node which has out-degree zero in A is called a *root* of the arborescence.

Clearly, every arborescence contains the same number of edges. In fact, if G contains exactly k minimal closed subsets of nodes, then every arborescence has $|V| - k$ edges and contains one root in each minimal closed subset. (A subset of nodes is said to be closed if no edges are directed out of the subset.) In particular, if G is strongly connected (i.e., the Markov chain is irreducible), then every arborescence is a set of $|V| - 1$ edges that form a directed spanning tree with all edges flowing towards a single node (the root of the tree).

Let $\mathcal{A}(V)$ denote the set of arborescences of G, $\mathcal{A}_j(V)$ denote the set of arborescences having root j, and $\mathcal{A}_{ij}(V)$ denote the set of arborescences having root j and a directed path from i to j. (In the special case $i = j$, we define $\mathcal{A}_{jj}(V)$ to be $\mathcal{A}_j(V)$.) In addition, let $\|\mathcal{A}(V)\|$, $\|\mathcal{A}_j(V)\|$, and $\|\mathcal{A}_{ij}(V)\|$ denote the sums of the weights of the arborescences in $\mathcal{A}(V)$, $\mathcal{A}_j(V)$, and $\mathcal{A}_{ij}(V)$, respectively.

The relationship between steady-state transition probabilities and arborescences is stated in the following theorem. The result is based on the well-known matrix tree theorem and is proved in [21]. For the sake of completeness, we have provided a sketch of the proof in the Appendix.

The Markov Chain Tree Theorem: Let the stochastic $n \times n$ matrix P define a finite Markov chain with long-run transition matrix $\bar{P}$. Then

$$\bar{p}_{ij} = \frac{\|\mathcal{A}_{ij}(V)\|}{\|\mathcal{A}(V)\|}.$$

Corollary: If the underlying graph is strongly connected, then

$$\bar{p}_{ij} = \frac{\|\mathcal{A}_j(V)\|}{\|\mathcal{A}(V)\|}.$$

As a simple example, consider once again the probabilistic FSE displayed in Fig. 2. Since the underlying graph is strongly connected, the corollary means that

$$\pi_i = \frac{\|\mathcal{A}_i(V)\|}{\|\mathcal{A}(V)\|}.$$

In addition, each $\mathcal{A}_i(V)$ consists of a single tree with weight

$$\frac{n-1}{n-1}p \cdot \frac{n-2}{n-1}p \cdots \frac{n-(i-1)}{n-1}p \cdot \frac{i}{n-1}q \cdot \frac{i+1}{n-1}q \cdots \frac{n-1}{n-1}q,$$

and thus

$$\|\mathscr{A}_i(V)\| = \binom{n-1}{i-1}\frac{(n-1)!}{(n-1)^{n-1}}p^{i-1}q^{n-i}.$$

Summing over i, we find that

$$\begin{aligned}\|\mathscr{A}(V)\| &= \sum_{i=1}^{n}\binom{n-1}{i-1}\frac{(n-1)!}{(n-1)^{n-1}}p^{i-1}q^{n-i}\\ &= \frac{(n-1)!}{(n-1)^{n-1}}(p+q)^{n-1}\\ &= \frac{(n-1)!}{(n-1)^{n-1}}\end{aligned}$$

and thus that

$$\pi_i = \binom{n-1}{i-1}p^{i-1}q^{n-i}.$$

Interestingly, this is the same as the probability that $i-1$ of the first $n-1$ inputs are ones and thus the FSE in Figs. 1 and 2 are equivalent (for $s = n-1$) in the long run! The FSE in Fig. 2 has fewer states, however, and mean-square error $\sigma^2(p) = pq/(n-1) = O(1/n)$.

The Markov chain tree theorem will also be useful in Section IV, where we prove a lower bound on the worst-case mean-square error of an n-state, FSE and in Section V, where we establish the universality of linear FSE's.

III. An Improved Deterministic FSE

In what follows, we show how to simulate the n-state probabilistic FSE shown in Fig. 2 with an $O(n \log n)$-state deterministic FSE. The resulting m-state deterministic FSE will then have mean-square error $O(\log m/m)$. This is substantially better than the mean-square error of the FSE shown in Fig. 1, and we conjecture that the bound is optimal for deterministic FSE's.

The key idea in the simulation is to use the randomness of the inputs to simulate a fixed probabilistic choice at each state. For example, consider a state i which on input one changes to state j with probability 1/2, and which remains in state i with probability 1/2. (See Fig. 3(a).)

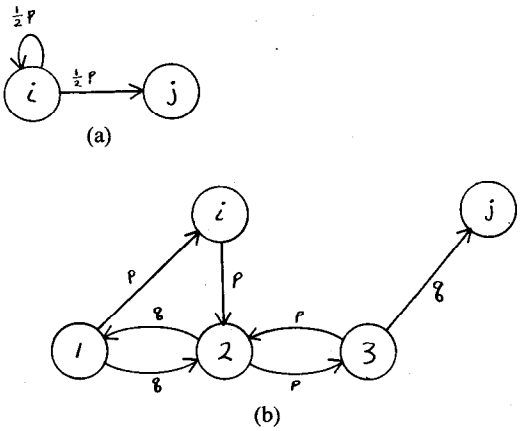

Fig. 3. Simulation of (a) probabilistic transitions by (b) deterministic transitions.

Such a situation arises for states $i = (n+1)/2$ and $j = (n+1)/2+1$ for odd n in the FSE of Fig. 2. These transitions can be modeled by the deterministic transitions shown in Fig. 3(b).

The machine in Fig. 3(b) starts in state i and first checks to see if the input is a one. If so, state 2 is entered. At this point, the machine examines the inputs in successive pairs. If 00 or 11 pairs are encountered, the machine remains in state 2. If a 01 pair is encountered, the machine returns to state i, and if a 10 pair is encountered, the machine enters state j. Provided that $p \neq 0, 1$ (an assumption that will be made throughout the remainder of the paper), a 01 or 10 pair will (with probability 1) eventually be seen, and the machine will eventually decide to stay in state i or move to state j. Note that, regardless of the value of p $(0 < p < 1)$, the probability of encountering a 01 pair before a 10 pair is identical to the probability of encountering a 10 pair before a 01 pair. Hence the deterministic process in Fig. 3(b) is equivalent to the probabilistic process in Fig. 3(a). (The trick of using a biased coin to simulate an unbiased coin has also been used by von Neumann in [26] and Hoeffding and Simons in [15].)

It is not difficult to generalize this technique to simulate transitions with other probabilities. For example, Fig. 4(b) shows how to simulate a transition which has probability $(3/8)p$. As before, the simulating machine first verifies that the input is a one. If so, state a_2 is entered and remaining inputs are divided into successive pairs. As before, 00 and 11 pairs are ignored. The final state of the machine depends on the first three 01 or 10 pairs that are seen. If the first three pairs are 10 10 10, 10 10 01, or 10 01

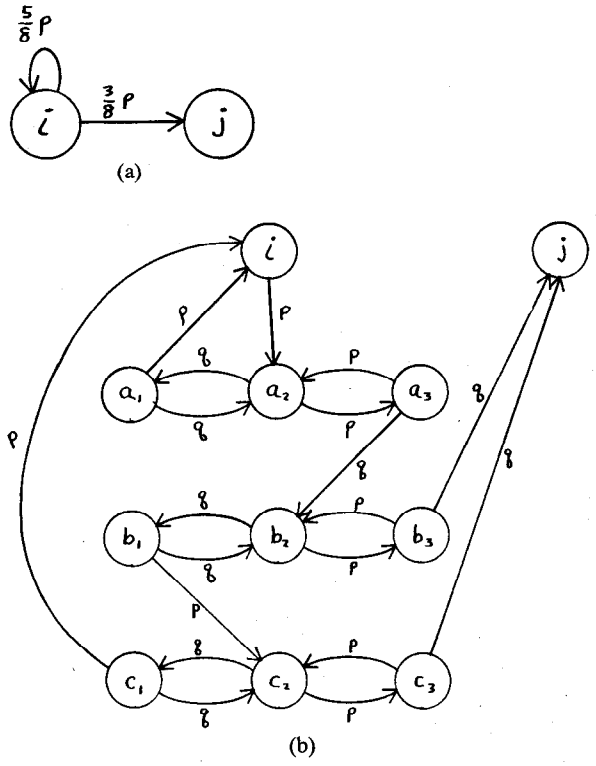

Fig. 4. Simulation of (a) probabilistic transitions by (b) deterministic transitions.

10 (in those orders), then the machine moves to state j. Otherwise, the machine returns to state i. Simply speaking, the machine interprets strings of 01's and 10's as binary numbers formed by replacing 01 pairs by zeros and 10 pairs by ones and decides if the resulting number is bigger than or equal to $101 = 5$. Since 01 and 10 pairs are encountered with equal probability in the input string for any p, the probability that the resulting number is five or bigger is precisely $3/8$.

In general, probabilistic transitions of the form shown in Fig. 5 (where x is an integer) can be simulated with $3k$ extra deterministic states, each with the same estimate. Hence when $n - 1$ is a power of two, the n-state probabilistic FSE in Fig. 2 can be simulated by a deterministic FSE with $6(n-1)\log(n-1) = O(n \log n)$ additional states. When n is not a power of two, the deterministic automata should simulate the next largest probabilistic automata that has 2^a states for some a. This causes at most a constant increase in the number of states needed for the simulation. Hence, for any m, there is an m-state deterministic automata with mean-square error $O(\log m/m)$.

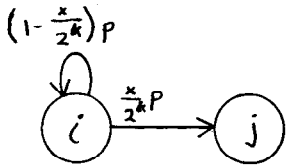

Fig. 5. General probabilistic transition.

IV. The Lower Bound

In this section, we show that, for every n-state probabilistic (or deterministic) FSE, there is a p such that the mean-square error of the FSE is $\Omega(1/n)$. The proof is based on the Markov chain tree theorem and the analysis of Section II.

From the analysis of Section II, we know that the mean-square error of an n-state FSE is

$$\sigma^2(p) = \sum_{j=1}^{n} \pi_j(\eta_j - p)^2 = \frac{\sum_{j=1}^{n} \|\mathscr{A}_{S_{0j}}(V)\|(\eta_j - p)^2}{\|\mathscr{A}(V)\|}$$

where $\|\mathscr{A}_{S_{0j}}(V)\|$ and $\|\mathscr{A}(V)\|$ are weighted sums of arborescences in the underlying graph of the FSE. In particular, each $\|\mathscr{A}_{S_{0j}}(V)\|$ is a polynomial of the form

$$f_j(p,q) = \sum_{i=1}^{n} a_{ij} p^{i-1} q^{n-i},$$

and $\|\mathscr{A}(V)\|$ is a polynomial of the form

$$g(p,q) = \sum_{i=1}^{n} a_i p^{i-1} q^{n-i}$$

where $a_i = \sum_{j=1}^{n} a_{ij}$ and $a_{ij} \geq 0$ for all $1 \leq i, j \leq n$. The nonnegativity of the a_{ij} follows from the fact that every edge of the graph underlying the FSE has weight $p_{ij} = \tau_{ij1}p + \tau_{ij0}q$, where τ_{ij1} and τ_{ij0} are nonnegative. Since every arborescence in the graph has $m \leq n-1$ edges, every term in the polynomial for $\|\mathscr{A}_{S_{0j}}(V)\|$ has the form $ap^r q^s$, where $r + s = m$. Multiplying by $(p+q)^{n-1-m} = 1$ then puts $f_j(p,q)$ in the desired form. The identity for $g(p,q)$ follows from the fact that $\|\mathscr{A}(V)\| = \sum_{j=1}^{n} \|\mathscr{A}_{S_{0j}}(V)\|$.

From the preceding analysis, we know that

$$\sigma^2(p) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij} p^{i-1} q^{n-i}(\eta_j - p)^2}{\sum_{i=1}^{n} a_i p^{i-1} q^{n-i}}$$

where $a_i = \sum_{j=1}^{n} a_{ij}$ and $a_{ij} \geq 0$ for $1 \leq i, j \leq n$. In what follows, we will show that

$$\int_{p=0}^{1} \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij} p^{n+i-1} q^{2n-i}(\eta_j - p)^2\, dp \geq \Omega\left(\frac{1}{n}\right) \int_{p=0}^{1} \sum_{i=1}^{n} a_i p^{n+i-1} q^{2n-i}\, dp$$

for all $a_{ij} \geq 0$ and η_j. Since the integrands are always nonnegative, we will have thus proved the existence of a p $(0 < p < 1)$ for which

$$\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij} p^{n+i-1} q^{2n-i}(\eta_j - p)^2 \geq \Omega\left(\frac{1}{n}\right) \sum_{i=1}^{n} a_i p^{n+i-1} q^{2n-i}.$$

Dividing both sides by $p^n q^n$ proves the existence of a p for which

$$\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij} p^{i-1} q^{n-i}(\eta_j - p)^2 \geq \Omega\left(\frac{1}{n}\right) \sum_{i=1}^{n} a_i p^{i-1} q^{n-i}$$

and thus for which $\sigma^2(p) \geq \Omega(1/n)$.

The proof relies heavily on the following well-known identities:

$$\int_0^1 p^i (1-p)^j\, dp = \frac{i!\,j!}{(i+j+1)!} \qquad (*)$$

and

$$\int_0^1 p^i (1-p)^j (p-\eta)^2\, dp \geq \frac{(i+1)!\,(j+1)!}{(i+j+3)!\,(i+j+2)} \qquad (**)$$

for all η.

The proof is now a straightforward computation:

$$\begin{aligned}
&\int_{p=0}^{1} \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} p^{n+i-1} q^{2n-i} (\eta_j - p)^2 \, dp \\
&\quad = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} \int_0^1 p^{n+i-1} (1-p)^{2n-i} (p - \eta_j)^2 \, dp \\
&\quad \geq \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{a_{ij}(n+i)!(2n-i+1)!}{(3n+2)!(3n+1)} \quad \text{by } (**) \\
&\quad = \sum_{i=1}^{n} \frac{a_i (n+i)!(2n-i+1)!}{(3n+2)!(3n+1)} \\
&\quad = \sum_{i=1}^{n} \frac{(n+i)(2n-i+1)}{(3n+2)(3n+1)^2} \frac{a_i (n+i-1)!(2n-i)!}{(3n)!} \\
&\quad \geq \frac{2n(n+1)}{(3n+2)(3n+1)^2} \sum_{i=1}^{n} \frac{a_i (n+i-1)!(2n-i)!}{(3n)!} \\
&\quad = \Omega\left(\frac{1}{n}\right) \sum_{i=1}^{n} a_i \int_0^1 p^{n+i-1} (1-p)^{2n-i} \, dp \text{ by } (*) \\
&\quad = \Omega\left(\frac{1}{n}\right) \int_{p=0}^{1} \sum_{i=1}^{n} a_i p^{n+i-1} q^{2n-i} \, dp.
\end{aligned}$$

It is worth remarking that the key fact in the preceding proof is that the long-run average transition probabilities of an n-state FSE can be expressed as ratios of $(n-1)$-degree polynomials with nonnegative coefficients. This fact comes from the Markov chain tree theorem. (Although it is easily shown that the long-run probabilities can be expressed as ratios of $(n-1)$-degree polynomials, and as infinite polynomials with nonnegative coefficients, the stronger result seems to require the full use of the Markov chain tree theorem.) The remainder of the proof essentially shows that functions of this restricted form cannot accurately predict p. Thus the limitations imposed by restricting the class of transition functions dominate the limitations imposed by quantization of the estimates.

V. Universality of Linear FSE's

In Section IV, we showed that the mean-square error of any n-state FSE can be expressed as

$$\sigma^2(p) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} p^{i-1} q^{n-i} (\eta_j - p)^2}{\sum_{i=1}^{n} a_i p^{i-1} q^{n-i}}$$

where $a_i = \sum_{j=1}^{n} a_{ij}$ and $a_{ij} \geq 0$ for $1 \leq i, j \leq n$. In this section, we will use this fact to construct an n-state *linear* FSE with mean-square error at most $\sigma^2(p)$ for all p. We first prove the following simple identity.

Lemma 1: If $a_1, \cdots, a_n$ are nonnegative, then

$$\sum_{j=1}^{n} a_j (\eta_j - p)^2 \geq a(\eta - p)^2$$

for all p and $\eta_1, \cdots, \eta_n$ where $a = \sum_{j=1}^{n} a_j$ and $\eta = (1/a) \sum_{j=1}^{n} a_j \eta_j$.

Proof: This is just a special case of the general theorem [12, Theorem 16] that an sth power mean is greater than an rth power mean if $s > r$. The lemma also follows from Cauchy's inequality [12, Theorem 6], or it can be proved using the observation that $f(x) = (x-p)^2$ is a convex function.

Let $\eta_i' = (1/a_i) \sum_{j=1}^{n} a_{ij} \eta_j$ for $1 \leq i \leq n$. From Lemma 1, we can conclude that

$$\sigma^2(p) \geq \frac{\sum_{i=1}^{n} a_i p^{i-1} q^{n-i} (p - \eta_i')^2}{\sum_{i=1}^{n} a_i p^{i-1} q^{n-i}}$$

for $0 \leq p \leq 1$. This ratio of sums is similar to the mean-square error of a linear FSE which never moves left on input one and never moves right on input zero. For example, the mean-square error of the linear FSE in Fig. 6 can be written in this form by setting

$$a_i = u_1 \cdots u_{i-1} v_{i+1} \cdots v_n$$

for $1 \leq i \leq n$.

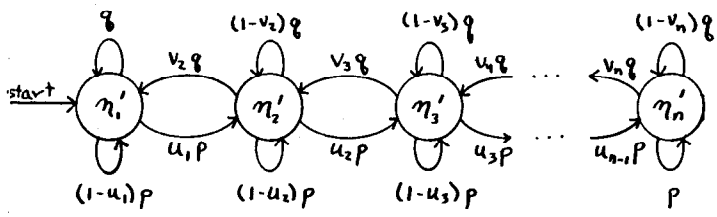

Fig. 6. Universal linear FSE.

Given a nonnegative set $\{a_i\}_{i=1}^{n}$, it is not always possible to find sets $\{u_i\}_{i=1}^{n-1}$ and $\{v_i\}_{i=2}^{n}$ such that $0 \leq u_i, v_i \leq 1$ and $a_i = u_1 \cdots u_{i-1} v_{i+1} \cdots v_n$ for all i. Two possible difficulties may arise. The first problem is that a_i might be larger than one for some i. This would mean that some u_j or v_j must be greater than one, which is not allowed. The second problem involves values of a_i which are zero. For example, if $a_1 \neq 0$ and $a_n \neq 0$, then each u_i and v_i must be nonzero. This would not be possible if $a_i = 0$ for some $i, 1 < i < n$.

Fortunately, both difficulties can be overcome. The first problem is solved by observing that the mean-square error corresponding to the set $\{ca_i\}_{i=1}^{n}$ is the same as the mean-square error corresponding to $\{a_i\}_{i=1}^{n}$ for all $c > 0$. By setting

$$u_i = \frac{a_{i+1}}{a_i}, \quad v_{i+1} = 1 \quad \text{if } a_i \geq a_{i+1},$$

$$u_i = 1, \quad v_{i+1} = \frac{a_i}{a_{i+1}} \quad \text{if } a_{i+1} \geq a_i,$$

and

$$c = \frac{u_1 \cdots u_{n-1}}{a_n},$$

we can easily verify that the mean-square error of the FSE

shown in Fig. 6 is

$$\frac{\sum_{i=1}^{n} ca_i p^{i-1}q^{n-i}(p-\eta_i')^2}{\sum_{i=1}^{n} ca_i p^{i-1}q^{n-i}} = \frac{\sum_{i=1}^{n} a_i p^{i-1}q^{n-i}(p-\eta_i')^2}{\sum_{i=1}^{n} a_i p^{i-1}q^{n-i}}$$

provided that $a_i > 0$ for $1 \le i \le n$. This is because

$$\begin{aligned} u_1 \cdots u_{i-1}v_{i+1} \cdots v_n &= \frac{ca_n}{u_i \cdots u_{n-1}} v_{i+1} \cdots v_n \\ &= ca_n\left(\frac{v_{i+1}}{u_i}\right) \cdots \left(\frac{v_n}{u_{n-1}}\right) \\ &= ca_n\left(\frac{a_i}{a_{i+1}}\right) \cdots \left(\frac{a_{n-1}}{a_n}\right) \\ &= ca_i. \end{aligned}$$

If $a_1 = \cdots = a_{j-1} = 0$ and $a_{k+1} = \cdots = a_n = 0$ but $a_i \neq 0$ for $j \le i \le k$, then the preceding scheme can be made to work by setting $u_1 = \cdots = u_{j-1} = 1$, $u_k = \cdots = u_{n-1} = 0$, $v_2 = \cdots = v_j = 0$, $v_{k+1} = \cdots = v_n = 1$,

$$u_i = \frac{a_{i+1}}{a_i}, \qquad v_{i+1} = 1 \text{ if } a_i \ge a_{i+1} \text{ for } j \le i \le k-1,$$

$$u_i = 1, \qquad v_{i+1} = \frac{a_i}{a_{i+1}} \text{ if } a_{i+1} \ge a_i \text{ for } j \le i \le k-1,$$

and

$$c = \frac{u_j \cdots u_{k-1}}{a_k}.$$

To overcome the second problem then, it is sufficient to show that if $a_j \neq 0$ and $a_k \neq 0$ for some FSE, then $a_i \neq 0$ for every i in the range $j \le i \le k$. From the analysis in Sections II and IV, we know that $a_i \neq 0$ if and only if an arborescence exists in the graph underlying the FSE which has $i-1$ edges weighted with a fraction of p and $n-i$ edges weighted with a fraction of q. In Lemma 2, we will show that, given any pair of arborescences A and A', one can construct a sequence of arborescences $A_1, \cdots, A_m$ such that $A_1 = A$, $A_m = A'$, and A_i and A_{i+1} differ by at most one edge for $1 \le i < m$. Since every edge of the graph underlying an FSE is weighted with a fraction of p or q or both, this result will imply that a graph containing an arborescence with $j-1$ edges weighted with a fraction of p and $n-j$ edges weighted with a fraction of q, and an arborescence with $k-1$ edges weighted with a fraction of p and $n-k$ edges weighted with a fraction of q, must also contain an arborescence with $i-1$ edges weighted with a fraction of p and $n-i$ edges weighted with a fraction of q for every i in the range $j \le i \le k$. This will conclude the proof that for every n-state FSE with mean-square error $\sigma^2(p)$, there is an n-state linear FSE with mean-square error at most $\sigma^2(p)$ for $0 \le p \le 1$.

Lemma 2: Given a graph with arborescences A and A', a sequence of arborescences $A_1, \cdots, A_m$ exists such that $A_1 = A$, $A_m = A'$, and A_{i+1} can be formed from A_i for $1 \le i < m$ by replacing a single edge of A_i with an edge of A'.

Proof: Given A_i, we construct A_{i+1} as follows. First we identify an edge $e = (u, v)$ from the set $A' - A_i$. Next, we consider the graph $A_i' = A_i + e$, which must contain either two edges directed out of u, or a directed cycle, or both. We claim that it is possible to have chosen e so that at most one of these cases arise by choosing e to be directed out of a root of A_i if possible (so we get only a cycle), or else by choosing the edge $e = (u, v)$ from $A' - A_i$ with u as near (in A') to a root of A' as possible. In the latter case, v and all its successors have as out-edges their edges from A', and the root of A_i that v leads to is a root of A', so that no cycles can arise by adding the edge e. We assume such an appropriate choice of e has been made. If u has out-degree two in A_i', we create A_{i+1} by deleting from A_i' the other edge out of u (which of necessity cannot belong to A', since A' is an arborescence). If A_i' contains a cycle, we create A_{i+1} by deleting from A_i' an edge in the directed cycle which does not belong to A'. (There must be such an edge, since A' contains no cycles.) This process terminates because the number of edges in common between A_i and A' increases by one at each step.

VI. Remarks

The literature on problems related to estimation with finite memory is extensive. Most of the work thus far has concentrated on the hypothesis testing problem [3], [6], [14], [33], [34], [36]. Generally speaking, the hypothesis testing problem is more tractable than the estimation problem. For example, several constructions are known for n-state automata which can test a hypothesis with long-run error at most $O(\alpha^n)$, where α is a constant in the interval $0 < \alpha < 1$ that depends only on the hypothesis. In addition, several researchers have studied the time-varying hypothesis testing problem [5], [18], [19], [24], [29], [37]. Allowing transitions to be time-dependent greatly enhances the power of an automata. For example, a four-state time-varying automata can test a hypothesis with an arbitrarily small long-run error.

As was mentioned previously, Samaniego [31] studied the problem of estimating the mean of a Bernoulli distribution using finite memory, and discovered the FSE shown in Fig. 2. Hellman studied the problem for Gaussian distributions in [13] and discovered an FSE which achieves the lower bound implied by the quantization problem. (Recall that this is not possible for Bernoulli distributions.) Hellman's construction uses the fact that events at the tails of the distribution contain a large amount of information about the mean of the distribution. The work on digital filters (e.g., [27], [28], [30]) and on approximate counting of large numbers [10], [23] is also related to the problem of finite-memory estimation.

We conclude with some questions of interest and some topics for further research:

1) Construct an n-state deterministic FSE with mean-square error $o(\log n/n)$ or show that no such construction is possible.
2) Construct a truly optimal (in terms of worst-case mean-square error) n-state FSE for all n.

3) Consider estimation problems where a prior distribution on p is known. For example, if the prior distribution on p is known to be uniform, then the n-state FSE in Fig. 2 has expected (over p) mean-square error $\Theta(1/n)$. Prove that this is optimal (up to a constant factor) for n-state FSE's.
4) Consider models of computation that allow more than constant storage. (Of course, the storage should also be less than logarithmic in the number of trials to make the problem interesting.)
5) Can the amount of storage used for some interesting models be related to the complexity of representing p? For example, if $p = a/b$, then $\log a + \log b$ bits might be used to represent p. Suppose that the FSE may use an extra amount of storage proportional to the amount it uses to represent its current prediction.

Acknowledgment

We thank Seth Chaiken, Tom Cover, Peter Elias, Robert Gallager, Martin Hellman, Dan Kleitman, Gary Miller, Larry Shepp, and Lorie Snell for helpful remarks and references. We also thank the referees for their helpful comments.

Appendix

Proof of the Markov Chain Tree Theorem

The Markov chain tree theorem was originally proved in [21] but was never published, so for completeness, we will sketch the proof in this Appendix. The proof is based on the matrix tree theorem (e.g., see [2]) and thus is similar to a number of derivative results in the literature. In fact, Corollary 1 is also proved in [17] and [35], although the result is not as well-known as one might expect. We commence with some elementary definitions and lemmas.

It is well known that the states of any Markov chain can be decomposed into a set T of transient states and sets $B_1, B_2, \cdots, B_m$ of minimal closed subsets of states. For any subset of states $W \subseteq V$, define $c(W)$ to be the number of minimal closed subsets of states contained in W. For example, every arborescence has $|V| - c(V)$ edges. The following lemma states a simple but important fact about $c(W)$.

Lemma A1: If U and W are disjoint subsets of V and if there are no edges from W to U in E, then $c(U \cup W) = c(U) + c(W)$.

Proof: Every minimal closed subset in U or W is a minimal closed subset in $U \cup W$. Thus $c(U \cup W) \geq c(U) + c(W)$. If a closed subset of $U \cup W$ contains nodes in both U and W, then the portion of the subset in W is also closed (since there are no edges from W to U). Thus the original subset is not minimal, implying that $c(U \cup W) \leq c(U) + c(W)$. Thus $c(U \cup W) = c(U) + c(W)$, as claimed.

Given any subset of nodes $W \subseteq V$, define an arborescence from W to be an acyclic subgraph of $G = (V, E)$ for which the out-degree of nodes in W is at most one and for which the out-degree of nodes in $V - W$ is zero. Let $\mathscr{A}^r(W)$ denote the set of arborescences from W with r edges, $\mathscr{A}_j^r(W)$ denote the set of arborescences from W with root j and r edges, and $\mathscr{A}_{ij}^r(W)$ denote the set of arborescences from W with root j, a path from i to j and r edges. (If $i = j$, then $\mathscr{A}_{ij}^r(W)$ is defined to be $\mathscr{A}_j^r(W)$.) As we are particularly interested in arborescences with $|W| - c(W)$ edges, we use $\mathscr{A}(W)$, $\mathscr{A}_j(W)$, and $\mathscr{A}_{ij}(W)$ to denote the sets $\mathscr{A}^{|W|-c(W)}(W)$, $\mathscr{A}_j^{|W|-c(W)}(W)$, and $\mathscr{A}_{ij}^{|W|-c(W)}(W)$, respectively. For example, $\mathscr{A}_{ij}(W)$ denotes the set of arborescences from W with root j, a path from i to j, and $|W| - c(W)$ edges.

Notice that the definitions for $\mathscr{A}(V)$, $\mathscr{A}_j(V)$, and $\mathscr{A}_{ij}(V)$ provided here are equivalent to those given in Section II. This is because every maximum arborescence has $|V| - c(V)$ edges. Also notice that $\mathscr{A}_j(W)$ and $\mathscr{A}_{ij}(W)$ may be empty for some W. This happens when node j is not contained in a minimal closed subset of W and/or when there is no path from i to j in G. When W is nonempty, $\mathscr{A}(W)$ is nonempty. In general, $\mathscr{A}^r(W)$ will be empty precisely when $r > |W| - c(W)$.

The weight of an arborescence from W and the $\|\mathscr{A}\|$ notation are defined as in Section II. Using Lemma A1, we easily establish the following identities.

Lemma A2: Let U and W be disjoint subsets of V such that there are no edges from W to U. Also let $i, i' \in U$ and $j, j' \in W$ be arbitrary vertices. Then

$$\|\mathscr{A}(U \cup W)\| = \|\mathscr{A}(U)\| \cdot \|\mathscr{A}(W)\|$$

$$\|\mathscr{A}_i(U \cup W)\| = \|\mathscr{A}_i(U)\| \cdot \|\mathscr{A}(W)\|$$

$$\|\mathscr{A}_j(U \cup W)\| = \|\mathscr{A}(U)\| \cdot \|\mathscr{A}_j(W)\|$$

$$\|\mathscr{A}_{ii'}(U \cup W)\| = \|\mathscr{A}_{ii'}(U)\| \cdot \|\mathscr{A}(W)\|$$

$$\|\mathscr{A}_{jj'}(U \cup W)\| = \|\mathscr{A}(U)\| \cdot \|\mathscr{A}_{jj'}(W)\|$$

$$\|\mathscr{A}_{ij}(U \cup W)\| = \sum^{j' \in W} \|\mathscr{A}_{ij'}(U)\| \cdot \|\mathscr{A}_{j'j}(W)\|.$$

Proof: The union of an arborescence from U with $|U| - c(U)$ edges and an arborescence from W with $|W| - c(W)$ edges is an arborescence from $U \cup W$ with $|U| - c(U) + |W| - c(W) = |U \cup W| - c(U \cup W)$ edges. (No cycles can be formed in the union since there are no edges from W to U.) Conversely, an arborescence from $U \cup W$ with $|U \cup W| - c(U \cup W)$ edges can have at most $|U| - c(U)$ edges from nodes in U and at most $|W| - c(W)$ edges from W. Hence the arborescence can be uniquely expressed as the union of an arborescence from U with $|U| - c(U)$ edges and an arborescence from W with $|W| - c(W)$ edges. Thus $\|\mathscr{A}(U \cup W)\| = \|\mathscr{A}(U)\| \cdot \|\mathscr{A}(W)\|$. The remaining identities can be similarly proved.

At first glance, it is not at all clear why sums of weighted arborescences should be related to long-run transition probabilities. Nor will the connection be made clear from our proof, which relies on the matrix tree theorem. In fact, both quantities are related to sums of weighted paths in the chain. We refer the reader to [21] for a longer but more enlightening proof.

Let X be an arbitrary real-valued $n \times n$ matrix. We let $C_k(X)$ denote the $n \times n$ matrix obtained from X by replacing its kth column by a length n vector of ones. We let $D_{ij}(X)$ denote the $(n-1) \times (n-1)$ matrix obtained from X by deleting its ith row and jth column. If A and B are sets, we also let $D_{AB}(X)$ denote the matrix obtained from X by deleting all rows in A and all columns in B. The following lemma contains some simple identities for the determinants of these matrices. (The determinant of a matrix X is denoted by $|X|$.)

Lemma A3: Let X be an $n \times n$ stochastic matrix. Then

$$|C_i(X)| = |C_j(X)| \qquad \text{for } 1 \le i, j \le n$$

$$|D_{ij}(X)| = (-1)^{i+j}|D_{ii}(X)| \qquad \text{for } 1 \le i, j \le n$$

$$|C_k(X)| = \sum_{i=1}^{n} |D_{ii}(X)| \qquad \text{for } 1 \le k \le n.$$

Proof: The proof is straightforward.

A general version of the matrix tree theorem [1] can be stated as follows.

Matrix Tree Theorem: Let the $n \times n$ matrix X have entries x_{ij}, where

$$x_{ij} = -y_{ij} \qquad \text{for } i \ne j,$$

and

$$x_{ii} = -y_{ii} + \sum_{k=1}^{n} y_{ik}.$$

Define an associated graph G with $V = \{1, \cdots, n\}$ and $E = \{(i, j) | y_{ij} \ne 0\}$ having weight y_{ij} on edge (i, j). Let $B \subseteq V$, i, $j \in V - B$ and $r = n - |B|$. Then

$$|D_{B,B}(X)| = \|\mathscr{A}^r(V - B)\|$$

and

$$(-1)^{i+j}|D_{B+j,B+i}(X)| = \|\mathscr{A}_{ij}^{r-1}(V - B - \{j\})\|.$$

Proof: See [1].

We now proceed with the proof of the Markov chain tree theorem, starting first with the case that the Markov chain M is irreducible (Corollary 1). In this case each row of $\bar{P}$ is equal to the vector π which is defined as the unique solution to

$$\pi P = \pi \qquad \sum_{k=1}^{n} \pi_k = 1.$$

The vector π is the stationary probability vector for M if M is aperiodic.

Since P is stochastic, the defining conditions on π can be combined to read

$$\pi C_k(I - P) = \epsilon_k$$

where I denotes the identity matrix and ϵ_k denotes the vector having a one in column k and zeros elsewhere. This equation uniquely defines π, for any $k, 1 \le k \le n$.

We now use Cramer's rule to solve for π:

$$\pi_k = \frac{|D_{kk}(I - P)|}{|C_k(I - P)|}.$$

Note that Lemma A3 implies that $|C_k(I - P)| = |C_l(I - P)|$ even if $k \ne l$, so the denominators of the equations for the π_k are all the same. A simple application of the Matrix tree theorem to the evaluation of $|D_{kk}(I - P)|$ then completes the proof for irreducible Markov chains.

We now generalize our result to include all Markov chains. As before, partition the states of M into a set T of transient states, and sets $B_1, \cdots, B_m$ of minimal closed subsets of states.

We let P_k denote the $|B_k| \times |B_k|$ submatrix of P giving the transition probabilities within B_k, Q denote the $|T| \times |T|$ matrix of transition probabilities within T, and R_k denote the $|B_k| \times |T|$ matrix of transition probabilities from B_k to T. By appropriate reordering the rows and columns of P we have

$$P = \begin{pmatrix} Q & R_1 & R_2 & \cdots & R_m \\ 0 & P_1 & 0 & \cdots & 0 \\ 0 & 0 & P_2 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & P_m \end{pmatrix}.$$

It is well-known that $\bar{P}$ then has the following form:

$$\bar{P} = \begin{pmatrix} 0 & U_1 & U_2 & \cdots & U_m \\ 0 & \bar{P}_1 & 0 & \cdots & 0 \\ 0 & 0 & \bar{P}_2 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \bar{P}_m \end{pmatrix}$$

where $\bar{P}_k$ is the long-run transition matrix for P_k,

$$U_k = NR_k\bar{P}_k$$

and

$$N = (I + Q + Q^2 + \cdots) = (I - Q)^{-1}.$$

Here n_{ij} is the average number of times M will visit state j, when M starts in state i. The matrix N always exists [16, Lemma III.4.1]. In fact, we will show in what follows that

$$n_{ij} = \frac{\|\mathscr{A}_{ij}(T - \{j\})\|}{\|\mathscr{A}(T)\|}.$$

By definition,

$$\begin{aligned} n_{ij} &= \left((I - Q)^{-1}\right)_{ij} \\ &= \frac{(-1)^{i+j}|D_{ji}(I - Q)|}{|I - Q|} \\ &= \frac{(-1)^{i+j}|D_{V-T+\{j\},V-T+\{i\}}(I - P)|}{|D_{V-T,V-T}(I - P)|} \\ &= \frac{\|\mathscr{A}_{ij}(T - \{j\})\|}{\|\mathscr{A}(T)\|} \end{aligned}$$

by the matrix tree theorem.

Clearly, both $\bar{p}_{ij}$ and $\|\mathscr{A}_{ij}(V)\|$ are zero unless $i, j \in B_k$ (one of the closed subsets), or $i \in T$ (the set of transient states) and $j \in B_k$. In the former case, $\bar{p}_{ij} = (\bar{P}_k)_{ij}$. From the analysis of irreducible chains, this means that $\bar{p}_{ij} = \|\mathscr{A}_{ij}(B_k)\|/\|\mathscr{A}(B_k)\|$ and thus that $\bar{p}_{ij} = \|\mathscr{A}_{ij}(V)\|/\|\mathscr{A}(V)\|$.

If $i \in T$ and $j \in B_k$, then (using Lemma A2)

$$\begin{aligned} \bar{p}_{ij} &= (NR_k\bar{P}_k)_{ij} \\ &= \sum_{l \in B_k} \sum_{l' \in T} \frac{\|\mathscr{A}_{il'}(T - \{l'\})\|}{\|\mathscr{A}(T)\|} \cdot \|\mathscr{A}_{l'l}(\{l'\})\| \cdot \frac{\|\mathscr{A}_{lj}(B_k)\|}{\|\mathscr{A}(B_k)\|} \\ &= \sum_{l \in B_k} \frac{\|\mathscr{A}_{il}(T)\|}{\|\mathscr{A}(T)\|} \cdot \frac{\|\mathscr{A}_{lj}(B_k)\|}{\|\mathscr{A}(B_k)\|} \\ &= \frac{\|\mathscr{A}_{ij}(T \cup B_k)\|}{\|\mathscr{A}(T \cup B_k)\|} \\ &= \frac{\|\mathscr{A}_{ij}(V)\|}{\|\mathscr{A}(V)\|}. \end{aligned}$$

This concludes the proof of the Markov chain tree theorem.

REFERENCES

[1] S. Chaiken, "A combinatorial proof of the all minors matrix tree theorem," *SIAM J. Algebraic Discrete Methods*, vol. 13, pp. 319–329, Sept. 1982.

[2] S. Chaiken and D. Kleitman, "Matrix tree theorems," *J. Comb. Theory*, Series A, vol. 24, pp. 377–381, May 1978.

[3] B. Chandrasekaran and C. Lam, "A finite-memory deterministic algorithm for the symmetric hypothesis testing problem," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 40–44, Jan. 1975.

[4] C. Coates, "Flow-graph solutions of linear algebraic equations," *IRE Trans. Circuit Theory*, vol. CT-6, pp. 170–187, 1959.

[5] T. Cover, "Hypothesis testing with finite statistics," *Ann. Math. Statist.*, vol. 40, pp. 828–835, 1969.

[6] T. Cover and M. Hellman, "The two-armed bandit problem with time-invariant finite memory," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 185–195, Mar. 1970.

[7] D. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs, Theory and Applications*. New York: Academic, 1979.

[8] J. Doob, *Stochastic Processes*. New York: Wiley, 1953.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1957.

[10] P. Flajolet, "On approximate counting," INRIA Research Rep. 153, July 1982.

[11] R. Flower and M. Hellman, "Hypothesis testing with finite memory in finite time," *IEEE Trans. Inform. Theory*, pp. 429–431, May 1972.

[12] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*. London: Cambridge Univ. Press, 1952.

[13] M. Hellman, "Finite-memory algorithms for estimating the mean of a Gaussian distribution," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 382–384, May 1974.

[14] M. Hellman and T. Cover, "Learning with finite memory," *Ann. Math. Statist.*, vol. 41, pp. 765–782, 1970.

[15] W. Hoeffding and G. Simons, "Unbiased coin tossing with a biased coin," *Ann. Math. Statist.*, vol. 41, pp. 341–352, 1970.

[16] D. Isaacson and R. Madsen, *Markov Chains—Theory and Applications*. New York: Wiley, 1976.

[17] H. Kohler and E. Vollmerhaus, "The frequency of cyclic processes in biological multistate systems," *J. Math. Biology*, no. 9, pp. 275–290, 1980.

[18] J. Koplowitz, "Necessary and sufficient memory size for *m*-hypothesis testing," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 44–46, Jan. 1975.

[19] J. Koplowitz and R. Roberts, "Sequential estimation with a finite statistic," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 631–635, Sept. 1973.

[20] S. Lakshmivarahan, *Learning Algorithms—Theory and Applications*. New York: Springer-Verlag, 1981.

[21] F. Leighton and R. Rivest, "The Markov chain tree theorem," Mass. Inst. Technol., Cambridge, MIT Tech. Memo. 249, Nov. 1983.

[22] Q. Minping and Q. Min, "Circulation for recurrent Markov chains," *Z. Varshinenaka*, vol. 59, no. 2, pp. 203–210, 1982.

[23] F. Morris, "Counting large numbers of events in small registers," *Commun. ACM*, vol. 21, pp. 840–842, Oct. 1978.

[24] C. Mullis and R. Roberts, "Finite-memory problems and algorithms," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 440–455, July 1974.

[25] K. Narendra and M. Thathachar, "Learning automata—A survey," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-4, pp. 323–334, July 1974.

[26] J. von Neumann, "Various techniques used in connection with random digits," *Monte Carlo Methods*, Applied Mathematics Series, no. 12. Washington, DC: Nat. Bureau of Standards, 1951, pp. 36–38.

[27] A. Oppenheim and R. Schafe, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[28] L. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[29] R. Roberts and J. Tooley, "Estimation with finite memory," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 685–691, 1970.

[30] A. Sage and J. Melsa, *Estimation Theory With Applications to Communications and Control*. New York: McGraw-Hill, 1971.

[31] F. Samaniego, "Estimating a binomial parameter with finite memory," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 636–643, Sept. 1973.

[32] ——, "On tests with finite memory in finite time," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 387–388, May 1974.

[33] ——, "On testing simple hypothesis in finite time with Hellman-Cover automata," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 157–162, Mar. 1975.

[34] B. Shubert, "Finite-memory classification of Bernoulli sequences using reference samples," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 384–387, May 1974.

[35] B. Shubert, "A flow-graph formula for the stationary distribution of a Markov chain," *IEEE Trans. Syst., Man, Cybern.*, pp. 555–556, Sept. 1975.

[36] B. Shubert and C. Anderson, "Testing a simple symmetric hypothesis by a finite-memory deterministic algorithm," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 644–647, Sept. 1973.

[37] T. Wagner, "Estimation of the mean with time-varying finite memory," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 523–525, July 1972.

[38] J. Koplowitz, "Estimation of the mean with the minimum size finite memory," in *Proc. IEEE Computer Society Conf. on Pattern Recognition and Image Processing*, 1977, pp. 318–320.