Estimating a Probability Using Finite Memory *

Extended Abstract

Frank Thomson Leighton and Ronald L. Rivest
Mathematics Department and Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, Mass. 02139

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 $i \geq 1$. We consider time-invariant finite-memory (i.e., finite-state) estimation procedures for the parameter p which take $X_1, \ldots$ as an input sequence. In particular, we describe an n-state deterministic estimation procedure that can estimate p with mean-square error $O(\frac{\log n}{n})$ and an n-state probabilistic estimation procedure that can estimate p with mean-square error $O(\frac{1}{n})$. We prove that the $O(\frac{1}{n})$ bound is optimal to within a constant factor. In addition, we show 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 the Markov Chain Tree Theorem.

1. 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

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

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

---

* This research was supported by the Bantrell Foundation and by NSF grant MCS-8006938.

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 $S = \{1, \ldots, n\}$, a start state $S_0 \in \{1, \ldots, n\}$, 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 Figure 1. This FSE has $n = \frac{(s+1)(s+2)}{2}$ 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 $\sigma^2(p) = \frac{pq}{s} = O(\frac{1}{\sqrt{n}})$.

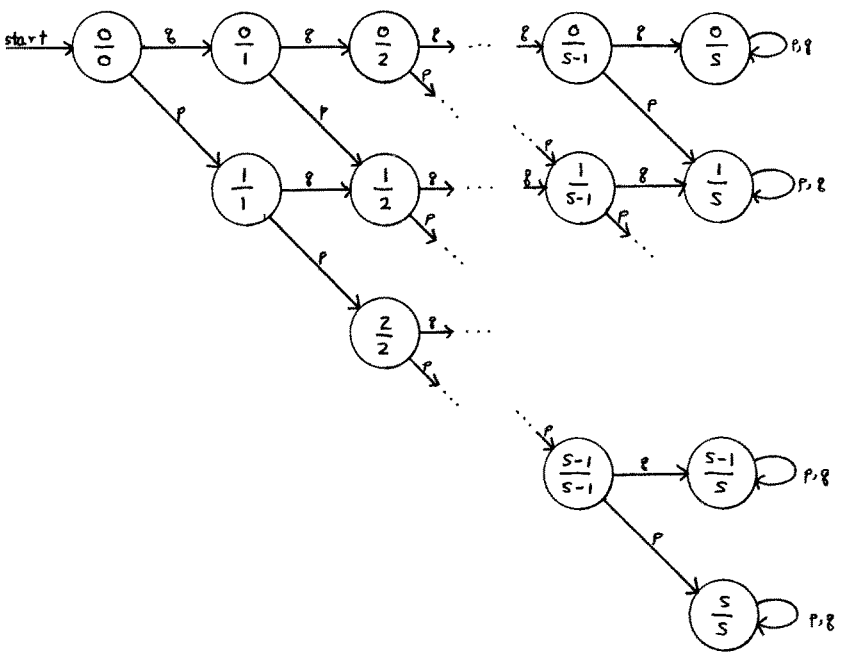

Figure 1: *An $\frac{(s+1)(s+2)}{2}$-state deterministic FSE with mean-square error $\sigma^2(p) = \frac{pq}{s}$. States are represented by circles. Arrows labeled with q denote transitions on input zero. Arrows labeled with p denote transitions on input one. Estimates are given as fractions and represent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

In [23], Samaniego considered probabilistic FSEs and constructed the probabilistic FSE shown in Figure 2. *Probabilistic* FSEs are similar to nonprobabilistic (or *deterministic*) FSEs 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, $\tau_{320} = \frac{2}{n-1}$ in Figure 2. So that τ is well-defined, we require that $\sum_{j=1}^{n} \tau_{ijk} = 1$ for all i and k.

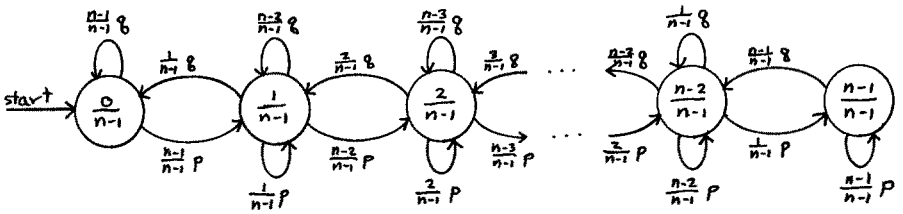

Figure 2: *A probabilistic n-state FSE with mean-square error* $\sigma^2(p) = \frac{pq}{n-1}$. *States are represented by circles in increasing order from left to right (e.g., state 1 is denoted by the leftmost circle and state n is denoted by the rightmost circle). State i estimates* $\frac{i-1}{n-1}$ *for* $1 \leq i \leq n$. *The estimates are shown as fractions within the 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, the probability of changing from state 2 to state 3 on input 1 is* $\frac{n-2}{n-1}$.

In this paper, we show that the mean-square error of the FSE shown in Figure 2 is $\sigma^2(p) = \frac{pq}{n-1} = O(\frac{1}{n})$, and that this 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), there is some value of p for which $\sigma^2(p) = \Omega(\frac{1}{n})$. Previously, the best lower bound known for $\sigma^2(p)$ was $\Omega(\frac{1}{n^2})$. The weaker bound is due to the "quantization problem", which provides a fundemental 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 $\frac{1}{2n}$. This means that the mean-square error for some p must be at least $\Omega(\frac{1}{n^2})$. Our result (which is based on the Markov Chain Tree Theorem [14]) 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 Figure 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 FSEs are the easiest FSEs 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 FSEs in the paper. Although we do not know how to achieve the $\Theta(\frac{1}{n})$ lower bound for deterministic FSEs, we can come close. In fact, we will construct an n-state deterministic FSE that has mean-square error $O(\frac{\log n}{n})$. The construction uses the input to deterministically simulate the probabilistic transitions of the FSE shown in Figure 2.

The remainder of the paper is divided into five sections. In Section 2, we present some background material on Markov chains (including the Markov Chain Tree Theorem) and prove that the FSE shown in Figure 2 has mean-square error $O(\frac{1}{n})$. In Section 3, we construct an n-state deterministic FSE with mean-square error $O(\frac{\log n}{n})$. The $\Omega(\frac{1}{n})$ lower bound for n-state FSEs is proved in Section 4. In Section 5, we demonstrate the universality of linear FSEs. We conclude in Section 6 with references and open questions.

2. Theory of Markov Chains

An n-state FSE act 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, $p_{33} = \frac{2}{n-1}p + \frac{n-3}{n-1}q$ for the FSE in Figure 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

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

This limit exists because P is stochastic (see Theorem 2 of [4]). The ijth entry of $\overline{P}$ is simply the long-run average probability $\overline{p}_{ij}$ 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 $\overline{P}$ 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 $\overline{P}$ may vary and we will use π to denote the S_0-th row of $\overline{P}$. 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, the mean-square error of an FSE can be expressed 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 [14] 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 $V = \{1, \ldots, n\}$ be the nodes of a directed graph G, with edge set $E = \{(i,j) \mid p_{ij} \neq 0\}$. 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 outdegree 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.

Leighton and Rivest proved the following in [14].

The Markov Chain Tree Theorem [14]: *Let the stochastic $n \times n$ matrix P define a finite Markov chain with long-run transition matrix $\overline{P}$. Then*

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

Corollary: *If the underlying graph is strongly connected, then*

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

As an example, consider once again the probabilistic FSE displayed in Figure 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

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

Summing over i, we find that

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

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 Figures 1 and 2 are equivalent (for $s = n-1$) in the long run! The FSE in Figure 2 has fewer states, however, and mean-square error $\sigma^2(p) = \frac{pq}{n-1} = O(\frac{1}{n})$.

The Markov Chain Tree Theorem will also be useful in Section 4 where we prove a lower bound on the worst-case mean-square error of an n-state FSE and in Section 5 where we establish the universality of linear FSEs.

3. An Improved Deterministic FSE

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

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 remains in state i with probability 1/2. (See Figure 3a.) Such a situation arises for states $i = \frac{n+1}{2}$ and $j = \frac{n+1}{2} + 1$ for odd n in the FSE of Figure 2. These transitions can be modelled by the deterministic transitions shown in Figure 3b.

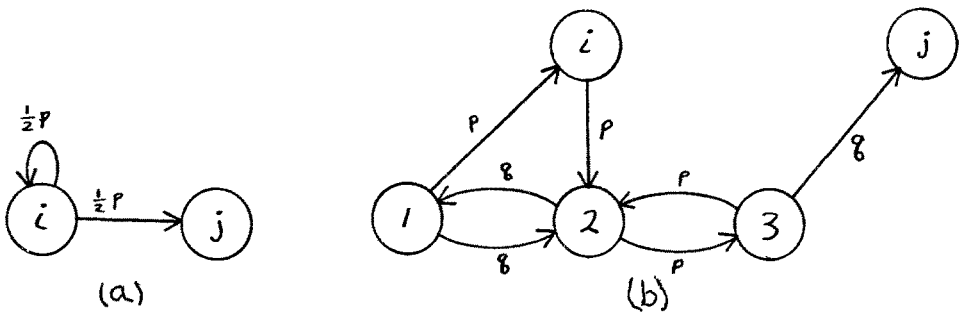

Figure 3: *Simulation of (a) probabilistic transitions by (b) deterministic transitions.*

The machine in Figure 3b 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 Figure 3b is equivalent to the probabilistic process in Figure 3a. (The trick of using a biased coin to simulate an unbiased coin has also been used by von Neumann in [18] and Hoeffding and Simons in [10].)

It is not difficult to generalize this technique to simulate transitions with other probabilities. For example, Figure 4b shows how to simulate a transition which has probability $\frac{3}{8}p$. 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" "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 0s and "10" pairs by 1s 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 5 or bigger is precisely $\frac{3}{8}$.

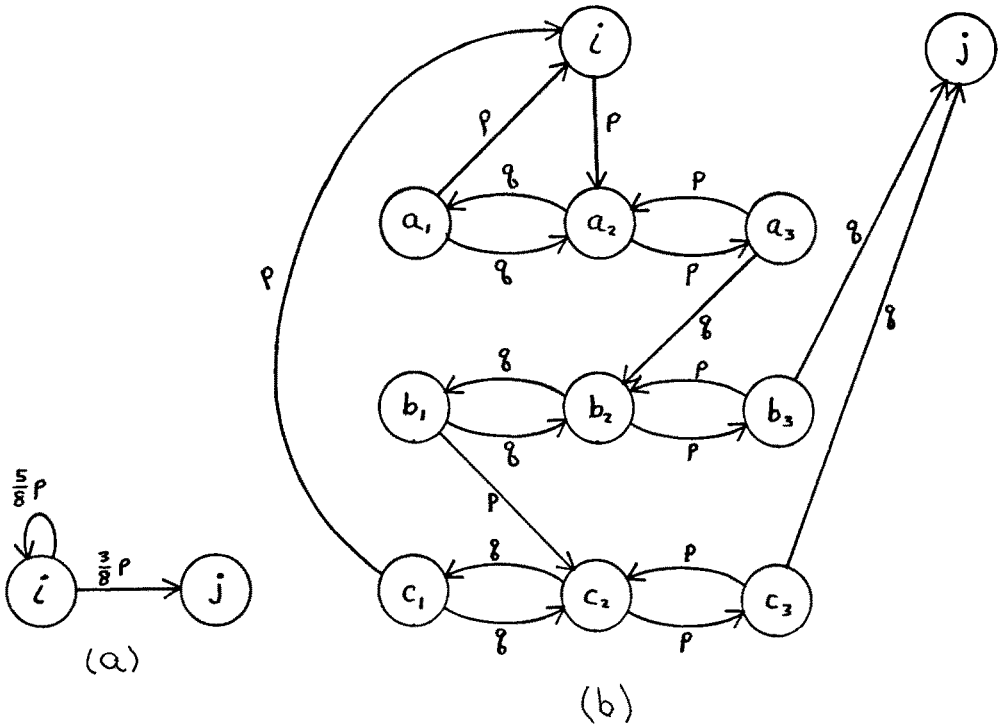

Figure 4: *Simulation of (a) probabilistic transitions by (b) deterministic transitions.*

In general, probabilistic transitions of the form shown in Figure 5 (where x is an integer) can be simulated with $3i$ extra deterministic states. Hence, when $n-1$ is a power of two, the n-state probabilistic FSE in Figure 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 $O(\frac{\log m}{m})$.

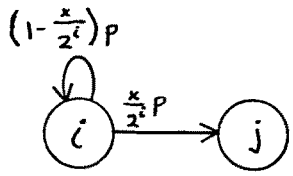

Figure 5: *General probabilistic transition.*

4. 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 $\Omega(\frac{1}{n})$. The proof is based on the Markov Chain Tree Theorem and the analysis of Section 2.

From the analysis of Section 2, we know that the mean-square error of an n-state FSE is

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

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

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

and $\|\mathcal{A}(V)\|$ 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}'s 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 $\|\mathcal{A}_{S_{0j}}(V)\|$ 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 $\|\mathcal{A}(V)\| = \sum_{j=1}^{n} \|\mathcal{A}_{S_{0j}}(V)\|$.

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

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

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

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

By dividing both sides by $p^n q^n$, this will prove the existence of a p for which

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

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

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

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

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

for all η.

The proof is now a straightforward computation.

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

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.

5. Universality of Linear FSEs

In Section 4, 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 $a_1, \ldots, a_n$ are nonnegative, then*

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

for all p and $\eta_1, \ldots, \eta_n$ where $a = \sum_{j=1}^{n} a_j$ and $\eta = \frac{1}{a}\sum_{j=1}^{n} a_j\eta_j$.

Proof: Since $a_1, \ldots, a_n$ are nonnegative, $a = 0$ if and only if $a_j = 0$ for $1 \leq j \leq n$. Thus

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

if and only if

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

which is true since

$$\begin{aligned} a\sum_{j=1}^{n} a_j(\eta_j - p)^2 - a^2(\eta - p)^2 &= a\sum_{j=1}^{n} a_j\eta_j^2 - a^2\eta^2 \\ &= \left(\sum_{i=1}^{n} a_i\right)\left(\sum_{j=1}^{n} a_j\eta_j^2\right) - \left(\sum_{i=1}^{n} a_i\eta_i\right)\left(\sum_{j=1}^{n} a_j\eta_j\right) \\ &= \sum^{1\leq i,j\leq n} a_ia_j(\eta_j^2 - \eta_i\eta_j) \\ &= \sum^{1\leq i<j\leq n} a_ia_j(\eta_j - \eta_i)^2 \end{aligned}$$

is nonnegative. ∎

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

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

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 Figure 6 can be written in this form by setting

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

for $1 \leq i \leq n$.

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. There are two possible difficulties. 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 $i, 1 < i < n$.

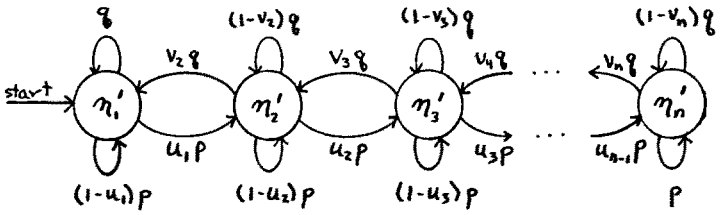

Figure 6: *Universal linear FSE.*

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

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

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

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

it can be easily verified that the mean-square error of the FSE shown in Figure 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 $1 \leq i \leq n$. 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 $j \leq i \leq k$, then the preceding scheme can be made to work by setting $u_1 = \cdots = u_{j-1} = 1$, $u_k = \cdots = u_{n-1} = 0$, $v_2 = \cdots = v_j = 0$, $v_{k+1} = \cdots = v_n = 1$,

$$u_i = \frac{a_{i+1}}{a_i} \quad \text{and} \quad v_{i+1} = 1 \quad \text{if} \quad a_i \geq a_{i+1} \quad \text{for} \quad j \leq i \leq k-1\,,$$

$$u_i = 1 \quad \text{and} \quad v_{i+1} = \frac{a_i}{a_{i+1}} \quad \text{if} \quad a_{i+1} \geq a_i \quad \text{for} \quad j \leq i \leq k-1\,,$$

$$\text{and} \qquad c = \frac{u_j \cdots u_{k-1}}{a_k}\,.$$

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 $j \leq i \leq k$. From the analysis in Sections 2 and 4, we know that $a_i \neq 0$ if and only if there is an arborescence 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 how, given any pair of arborescences A and A', to construct a sequence of arborescences $A_1, \ldots, A_m$ such that $A_1 = A$, $A_m = A'$, and A_i and A_{i+1} differ by at most one edge for $1 \leq i < m$. 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 $j \leq i \leq k$. 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 $0 \leq p \leq 1$.

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

Proof: The sequence of edge replacements proceeds in two phases. In the first phase, a node v in A_i is selected such that

1) v is neither a root of A_i nor a root of A',
2) the edge from v in A_i is different than the edge from v in A', and
3) the edges from all ancestors of v (if any) in A_i are edges in A'.

Then the edge from v in A_i is replaced by the edge from v in A' to form A_{i+1}.

The first phase continues until the supply of nodes that satisfy the three conditions is exhausted. At this point, every edge in A_i that is not on a path from a root of A' to a root of A_i is also in A'.

In the second phase, a root v of A_i that is not a root of A' is selected and the edge from v in A' is inserted to form A_{i+1}. In addition, the unique edge that enters v and that is descendent in A_i from a root of A' is removed in A_{i+1}. The ancestor of v in A_i that is a descendent of a root in A' then becomes a root of A_{i+1}. Note that the length of the path from the root of A' to the root of A_{i+1} is one less than the length of the path from the root of A' to v. Thus repetition of this process will eventually produce an arborescence A_m which has the same roots as A'. At this point, the procedure terminates. (For an example of this process, see Figure 7.)

Since every arborescence has exactly one root in each minimal closed subset of nodes, the preceding algorithm constructs a sequence of graphs $A_1, \ldots, A_m$ such that $A_1 = A$, $A_m = A'$, and A_i and A_{i+1} differ in at most one edge for $1 \leq i < m$. In order to complete the proof, we must show that each A_i is an arborescence. The proof is by induction, and shows that if A_i has no cycles or nodes with outdegree greater than one, then A_{i+1} has no cycles or nodes with outdegree greater than one. Since A_i and A_{i+1} have the same number of edges, we will have thus shown that A_{i+1} is an arborescence if A_i is an arborescence.

The outdegree constraint is straightforward to verify since, in the first phase, the outdegree of the nodes is not changed, and in the second phase, outgoing edges are added only to roots. It is also easy to verify that cycles are not introduced in the procedure. If a cycle were introduced in the first phase, it would have to consist of edges that are also in A' (by the third constraint on v), thus violating the acyclicity of A'. If a cycle were formed in the second phase, it could only contain edges which are not on the path from a root of A' to a root of A_i (since the last edge in this path was removed). Such edges are in A', however, again violating the acyclicity of A'. ∎

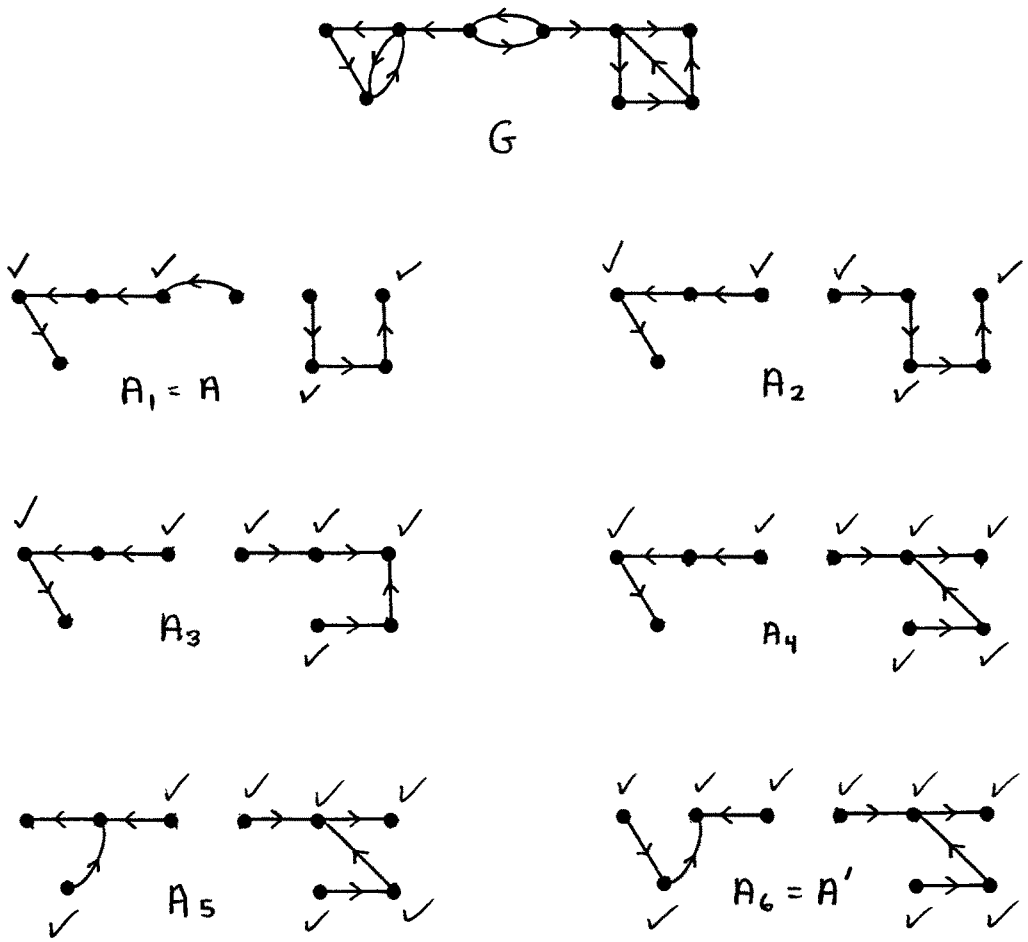

Figure 7: *Deforming A into A' by a sequence of edge replacements. Checkmarks denote nodes for which the outgoing edge is in A'. Arborescences A_2, A_3 and A_4 are formed during the first phase.*

6. Remarks

There is a large literature on problems related to estimation with finite memory. Most of the work thus far has concentrated on the *hypothesis testing problem* [1, 3, 9, 25, 26, 27]. 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 [2, 11, 12, 16, 21, 28]. Allowing transitions to be time-dependent greatly enhances the power of an automata. For example, a 4-state time-varying automata can estimate a probability with an arbitrarily small mean-square error.

As was mentioned previously, Samaniego [23] studied the problem of estimating the mean of a Bernoulli distribution using finite memory, and discovered the FSE shown in Figure 2. Hellman studied the problem for Gaussian distributions in [8], 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., [19, 20, 22]) and on approximate counting of large numbers [6, 15] 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 $o(\frac{\log n}{n})$ 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 Figure 2 has expected (over p) mean-square error $\Theta(\frac{1}{n})$. Prove that this is optimal (up to a constant factor) for n-state FSEs.
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.

Acknowledgements

We thank Tom Cover, Martin Hellman, Robert Gallager, and Peter Elias for helpful discussions.

References

[1] B. Chandrasekaran and C. Lam, "A Finite-Memory Deterministic Algorithm for the Symmetric Hypothesis Testing Problem," *IEEE Transactions on Information Theory*, Vol. IT-21, No. 1, January 1975, pp. 40-44.

[2] T. Cover, "Hypothesis Testing with Finite Statistics," *Annals of Mathematical Statistics*, Vol. 40, 1969, pp. 828-835.

[3] T. Cover and M. Hellman, "The Two-Armed Bandit Problem With Time-Invariant Finite Memory," *IEEE Transactions on Information Theory*, Vol. IT-16, No. 2, March 1970, pp. 185-195.

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

[5] W. Feller, *An Introduction to Probability Theory and its Applications*, Wiley, New York, 1957.

[6] P. Flajolet, "On Approximate Counting," INRIA Research Report No. 153, July 1982.

[7] R. Flower and M. Hellman, "Hypothesis Testing With Finite Memory in Finite Time," *IEEE Transactions on Information Theory*, May 1972, pp. 429-431.

[8] M. Hellman, "Finite-Memory Algorithms for Estimating the Mean of a Gaussian Distribution," *IEEE Transactions on Information Theory*, Vol. IT-20, May 1974, pp. 382-384.

[9] M. Hellman and T. Cover, "Learning with Finite Memory," *Annals of Mathematical Statistics*, Vol. 41, 1970, pp. 765-782.

[10] W. Hoeffding and G. Simons, "Unbiased Coin Tossing with a Biased Coin," *Annals of Mathematical Statistics*, Vol. 41, 1970, pp. 341-352.

[11] J. Koplowitz, "Necessary and Sufficient Memory Size for m-Hypothesis Testing," *IEEE Transactions on Information Theory*, Vol. IT-21, No. 1, January 1975, pp. 44-46.

[12] J. Koplowitz and R. Roberts, "Sequential Estimation With a Finite Statistic," *IEEE Transactions on Information Theory*, Vol. IT-19, No. 5, September 1973, pp. 631-635.

[13] S. Lakshmivarahan, *Learning Algorithms – Theory and Applications*, Springer-Verlag, New York, 1981.

[14] F. Leighton and R. Rivest, "The Markov Chain Tree Theorem," to appear.

[15] F. Morris, "Counting Large Numbers of Events in Small Registers," *Communications of the ACM*, Vol. 21, No. 10, October 1978, pp. 840-842.

[16] C. Mullis and R. Roberts, "Finite-Memory Problems and Algorithms," *IEEE Transactions on Information Theory*, Vol. IT-20, No. 4, July 1974, pp. 440-455.

[17] K. Narendra and M. Thathachar, "Learning Automata – A Survey," *IEEE Transactions on Systems*, Vol. SMC-4, No. 4, July 1974, pp. 323-334.

[18] J. von Neumann, "Various Techniques Used in Connection With Random Digits," *Monte Carlo Methods*, Applied Mathematics Series, No. 12, U. S. National Bureau of Standards, Washington D. C., 1951, pp. 36-38.

[19] A. Oppenheim and R. Schafe, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[20] L. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[21] R. Roberts and J. Tooley, "Estimation With Finite Memory," *IEEE Transactions on Information Theory*, Vol. IT-16, 1970, pp. 685-691.

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

[23] F. Samaniego, "Estimating a Binomial Parameter With Finite Memory," *IEEE Transactions on Information Theory*, Vol. IT-19, No. 5, September 1973, pp. 636-643.

[24] F. Samaniego, "On Tests With Finite Menory in Finite Time," *IEEE Transactions on Information Theory*, Vol. IT-20, May 1974, pp. 387-388.

[25] F. Samaniego, "On Testing Simple Hypothesis in Finite Time With Hellman-Cover Automata," *IEEE Transactions on Information Theory*, Vol. IT-21, No. 2, March 1975, pp. 157-162.

[26] B. Shubert, "Finite-Memory Classification of Bernoulli Sequences Using Reference Samples," *IEEE Transactions on Information Theory*, Vol. IT-20, May 1974, pp. 384-387.

[27] B. Shubert and C. Anderson, "Testing a Simple Symmetric Hypothesis by a Finite-Memory Deterministic Algorithm," *IEEE Transactions on Information Theory*, Vol. IT-19, No. 5, September 1973, pp. 644-647.

[28] T. Wagner, "Estimation of the Mean With Time-Varying Finite Memory," *IEEE Transactions on Information Theory*, Vol. IT-18, July 1972, pp. 523-525.