The Annals of Statistics
1974 Vol. 2, No. 2, 323-336

MULTIPLE HYPOTHESIS TESTING BY FINITE MEMORY ALGORITHMS

BY S. YAKOWITZ

University of Arizona and Naval Postgraduate School

In recent years, a theory has been emerging concerning the statistical power of small computers. In the present paper it is proven that in the sense peculiar to this literature, small computers (mathematically equivalent to finite automata) can in general be designed to solve multiple simple hypothesis testing problems. In many cases, only one state for each hypothesis is needed. In a more conventional sense, we reveal the construction of finite automata which implement sequential decision procedures having the capacity to distinguish between any given finite set of probabilities with any desired accuracy. Finally, some results on the ability of finite automata to track time—changing hypotheses are outlined.

1. Introduction and preliminaries. In the tradition of studies by H. Robbins [8], Isbell [7], Samuels [10], and Hellman and Cover [6], this work is devoted to the investigation of some of the statistical powers of finite automata (FA's). Specifically, let $(\mathscr{X}, \mathscr{A})$ be a measurable space on which is defined a finite family $\mathscr{P} = \{P_j\colon 1 \leqq j \leqq M\}$ of probability functions and suppose $(S, \mathscr{X}, f)$ to be a finite automaton (FA) with (finite) state set S, the input set being the sample space $\mathscr{X}$ of the measurable space $(\mathscr{X}, \mathscr{A})$, and f an $\mathscr{A}$-measurable function mapping $S \times \mathscr{X}$ into S. An initial state $s(0) \in S$ and an $\mathscr{X}$-valued sequence $\{X(i)\}$ having been specified, a FA determines a sequence of states $\{s(n)\}$ (called a *trajectory*) by the following recursive rule: $s(n+1) = f(s(n), X(n))$. Our definition of FA differs from that in Arbib [1], for example, only in the inessential aspects that we do not define an "output" and do not restrict the input set to be finite. It is to be understood that an algorithm has finite memory (by definition) if it can be implemented by some FA.

In the aforementioned studies as well as this, a FA is employed for hypothesis testing as follows: Let $\{X_i\}$ be an independent random sequence of $\mathscr{X}$-values, each member of which has the common probability distribution $P \in \mathscr{P}$. The random sequence $\{X(n)\}$ serves as input to the FA whose initial state is some arbitrarily chosen state in S. The reader will recognize that the random trajectory $\{\mathbf{s}(i)\}$ so induced is a stationary Markov chain—a fact which will be used without comment in our analysis. The state set S having been partitioned into subsets $S_1, S_2, \cdots, S_M$ indexed by the same set as $\mathscr{P}$, the quality of performance of a FA is measured by the asymptotic relative frequency that the $s(n) \in S_k$, k being the index of the common distribution $P_k \in \mathscr{P}$ of the X_i's. Let us precisely

Received May 1972; revised January 1973.

AMS 1970 *subject classifications*. Primary 62F05; Secondary 93E10.

Key words and phrases. Multiple hypothesis testing, finite automata, finite memory distinguishability.

define a quality number of a FA for the hypothesis testing task just described.

DEFINITION. Let $(\mathscr{X}, \mathscr{A})$, $\mathscr{P}$, $(S, \mathscr{X}, f)$, $S_1, S_2, \cdots, S_M$ be as above and assume that each $P_j \in \mathscr{P}$ induces a unique invariant probability measure $v_j(\cdot)$ on $S_1, S_2, \cdots, S_M$. The *probability of error* of the FA is

$$\max\{1 - v_j(S_j) : 1 \leqq j \leqq M\} .$$

The probability of error is taken to be the "worst case" relative frequency that the FA is not in the set $S_k \subset S$ having the same index as P_k, the probability in $\mathscr{P}$ governing the distribution of the X_i's. Let us say that the FA hypothesis testing problem is *solvable* if for every positive number e there is some FA having probability of error less than e.

In the course of our analysis, we will see, for example, that if randomization is used, for any $\mathscr{P}$, the hypothesis testing problem is solvable. Further, if for every $e > 0$ and any two members P_i and $P_j \in \mathscr{P}$, there is an event A_e such that $\min\{P_i[A_e]/P_j[A_e],\ P_j[A_e]/P_i[A_e]\} < e$, then the hypothesis testing problem is solvable within the family of M state machines, M being the cardinal number of $\mathscr{P}$.

In Section 4, statistical hypothesis testing by FA's is put into a more conventional framework, namely that of Wald's [12] statistical decision functions. Specifically, $(\mathscr{X}, \mathscr{A})$, $\mathscr{P}$, and a positive number e having been given we show the construction of a FA with a state q_j for each P_j such that with input sequence $\{X_i\}$ of i.i.d. variables, eventually the FA enters some state q_v never to leave (i.e., the q_j's are absorbing). This signifies the estimate that $\{X_i\}$ has the distribution P_v. We have that for any $e > 0$, $\mathscr{P}$, and the FA we construct,

$$P_j[\text{absorbing state is } q_j] > 1 - e , \qquad 1 \leqq j \leqq M .$$

The paper closes with some preliminary studies of the capability of FA's to keep track of the underlying law P_j for the sequence $\{X_i\}$ for which, from time to time, the underlying law changes to other distributions in $\mathscr{P}$.

For historical perspective, let us review results and techniques of some of the fundamental publications on statistical FA's. H. Robbins proposed [8] a two-armed bandit problem of selecting, on the basis of the past r tosses (r a fixed number), which of two coins (whose Bernoulli parameters are initially unknown) should be tossed at time n ($n = 1, 2, \cdots$). The object is to achieve the maximum proportion of heads, or equivalently to choose the coin having the higher probability of heads the greatest proportion of time. He proposed the strategy, "Start tossing with coin 1. Stop if the first toss is tails, otherwise continue tossing until the first run of r successive tails occurs and then stop. This defines the first block of tosses with coin 1. Now start tossing with coin 2 and apply the same rule, obtaining the first block of tosses with coin 2. Then start again with coin 1, and so on indefinitely, thus generating an infinite sequence of tosses consisting of alternative blocks of tosses with coins 1 and 2." In order to put this strategy into our framework, it is necessary to make some artifices; let us assume that actually both coins are tossed at each epoch, but for each $s \in S$ $f(s, X)$ depends only on either the first or second coordinate of X. Then $\mathcal{X}$ is taken to be the set of 2-tuples whose coordinates are either H or T—the first coordinate corresponding to an outcome on coin 1 and the second to the outcome of coin 2. The FA illustrated in Figure 1 implements Robbins' strategy. Robbins

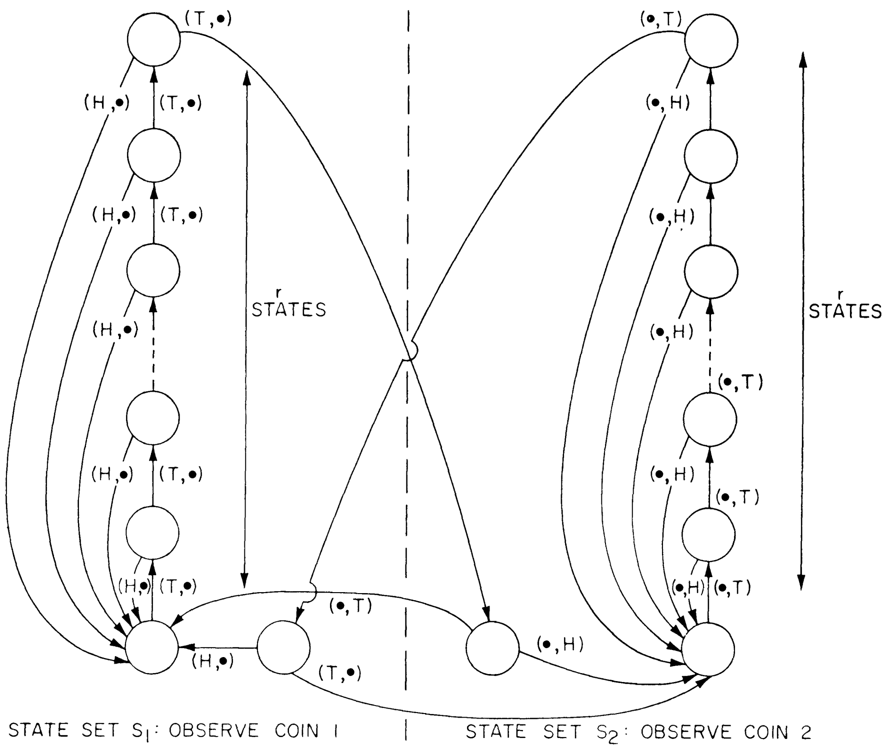

FIG. 1. FA implementation of Robbins' finite memory TAB strategy.

proved that the probability of error converges to 0 as r becomes arbitrarily large.

Isbell [7], Smith and Pyke [10], Samuels [11], Cover [3] and Cover and Hellman [4], improved on Robbins' strategy in the sense that their strategies give lower error for fixed memory length, where "memory length" is the upper bound to the number of successive past outcomes determined by the current state. Cover and Hellman provided a lower bound to the achievable probability of error by FA's with a given number of states, and gave the construction of a class of FA's having members whose performance comes arbitrarily close to the unachievable lower bound.

Cover [3] and Hellman and Cover [6] extended the finite memory hypothesis testing problem to a general class of two hypothesis problems. In particular, the latter paper is definitive in that it develops a lower bound to the probability of error of an n-state FA (n fixed) for a given problem and gives the construction of n state FA's which perform within any given tolerance of this (unachievable) bound.

Sagalowicz extended the FA theory to multi-hypothesis problems. In his Ph. D. thesis [9] under the direction of Professor Cover, he conjectured our Theorem 1, proving it in certain cases (such as the three-hypothesis problem) and outlining its wide domain of applicability.

Our appraisal is that an automata theory viewpoint on statistics is important because ultimately statistical schemes must be implemented by computers and automata theory is the theory of what can be accomplished through digital computation. There may be some delicacy required in going from an arbitrary statistical rule to a computable procedure. For example, Cover [3] shows that in a certain problem if the decimal expansions of the observations are truncated the usual test is no longer consistent, but a new one may be devised which is. FA's are the accepted mathematical model in computer science for small, special purpose digital computers. On the other hand, Turing machines, or equivalently recursive functions, are thought to represent large digital computers.

2. FA hypothesis testing solvability.

LEMMA. *Let an irreducible, aperiodic Markov chain with states* $1, 2, \cdots, M-1$ *be augmented to allow passage between state* $M-1$ *and a new state, state* M. *That is, if* $A = [a_{ij}]$ *and* $B = [b_{ij}]$ *are respectively the original and augmented Markov transition matrices,* $a_{ij} = b_{ij}$ *if* $i, j < M$ *and* i *or* $j < M-1$. $b_{M-1,M-1} = a_{M-1,M-1} - b_{M-1,M}$. $b_{Mk} = 0$ *for* $k < M-1$. *Under these circumstances, if* $v(\cdot)$ *is the unique invariant distribution for* A, B *has a unique invariant distribution* $u(\cdot)$ *determined by the conditions,*

$$u(j) = (1 - u(M))v(j)\,, \qquad j < M\,,$$
$$u(M)/u(M-1) = b_{M-1,M}/b_{M,M-1}\,.$$

The proof is computational and follows directly by substitution of the asserted solution for $u(\cdot)$ into B.

ASSUMPTION 1. For every positive number e, $1 \leqq i, j \leqq M$, there are sets $A_{ij}(e) \in \mathscr{A}$ which have the properties

(i) $P(A_{ij}(e)) > 0$, $P \in \mathscr{P}$;
(ii) for $j \neq k$ and all $e' > 0$, $A_{ij}(e) \cap A_{ik}(e')$ is empty, and
(iii) $P_i(A_{ij}(e))/P_i(A_{ji}(e)) < e$.

THEOREM 1. *Assumption* 1 *is a sufficient condition that the* FA *hypothesis testing problem be solvable by a class of* FA'*s having one state for each hypothesis.*

PROOF. The proof proceeds by induction on M, the number of hypotheses. It would be sufficient to start with the trivial case ($M = 1$), but it is mildly instructive to consider the case $M = 2$. Assume e is a positive number, $i, j \in S = \{1, 2\}$, and $f(i, x) = j$ if and only if $x \in A_{ij}(e)$ when $i \neq j$. The Markov chain induced by the FA has the transition matrix (as in Feller's [5] definition).

$$\begin{bmatrix} 1 - P_j(A_{12}(e)) & P_j(A_{21}(e)) \\ P_j(A_{12}(e)) & 1 - P_j(A_{21}(e)) \end{bmatrix}^T$$

where superscript T denotes "transpose". As all the coordinates are positive, it is known that the matrix is aperiodic and has a unique invariant probability, $u(\cdot)$, on S. Slight algebraic manipulation yields

$$u(1) = [P_j(A_{21}(e))/P_j(A_{12}(e))]u(2)\,, \qquad j = 1, 2\,.$$

Assume P_1 is the active hypothesis; recalling Assumption 1, we write

$$u(1) > (1/e)u(2)\,.$$

The side condition that $u(1) + u(2) = 1$ gives us

$$u(1) > (1/e)(1 - u(1))$$

or

$$u(1) > 1/(1 + e)\,.$$

Consequently, $u(1) > 1 - e$. The situation is symmetric so that under P_2,

$$u(2) > 1 - e \quad \text{also.}$$

Now let us discuss the inductive step. Suppose that for any $M - 1$ hypothesis problem satisfying Assumption 1 and for any positive number d, there is some $M - 1$ state FA having probability of error less than d. Let us suppose further, as part of the inductive hypothesis (note it holds for the $M = 2$ case) that if transition from state i to state k, $i \neq k$, has positive probability, then

$$\{x : f(i, x) = k\}$$

is of the form $A_{ik}(h)$, for some $h > 0$, as described in Assumption 1. These conditions are sufficient to ensure that under each $P \in \mathscr{S}$ the resulting Markov chain has a unique invariant probability, $v(\cdot)$. Now let us suppose that an extra probability function, P_M, is augmented to the set $\{P_1, \cdots, P_{M-1}\}$ and that Assumption 1 still prevails. Let t be some hypothesis number such that for any $d > 0$, there is some FA (of the above type) having probability of error less than d and having invariant probability under P_M which satisfies

$$v(t) \geqq v(j)\,, \qquad 1 \leqq j \leqq M - 1\,.$$

There must be some such t inasmuch as there are only finitely many hypotheses. For the rest of this proof, we will restrict our attention to the class of $M - 1$ state machines satisfying the above inequality. To construct a FA for the M hypothesis problem we are going to augment a state, call it M, to the chain with states $\{1, 2, \cdots, M - 1\}$, as in the lemma; that is, we allow passage to state M only through state t. Such transition takes place whenever the system is in state t and event $A_{tM}(c)$ occurs (the constant c will be determined presently) or when the system is in state M and event $A_{Mt}(c)$ occurs. Denoting the invariant probability of the augmented chain by $u(\cdot)$, under P_M

$$u(M) = [P_M(A_{tM}(c))/P_M(A_{Mt}(c))]u(t) .$$

Recalling Assumption 1 and the lemma, we write

$$u(M) > (1/c)v(t)(1 - u(M)) .$$

We defined t so that under P_M, $v(t) \geqq (M-1)^{-1}$; so the above inequality implies

$$u(M) > (c(M-1))^{-1}(1 - u(M))$$

and solving explicitly for $u(M)$ we have $u(M) > 1/(1 + (M-1)c)$ whence

(1) $$u(M) > 1 - (M-1)c , \qquad \text{under} \quad P_M .$$

Under P_t (recalling the inductive hypothesis that the original FA has probability of error $< d$), $u(t) = (1 - u(M))v(t) > (1 - u(M))(1 - d)$. But from the lemma, $u(M) = P_t(A_{tM}(c))/P_t(A_{Mt}(c))\, u(t) < cu(t)$, which allows us to substitute for $u(M)$ as follows: $u(t) > (1 - cu(t))(1 - d)$, which, when solved explicitly for $u(t)$ can be reduced to

(2) $$u(t) > 1 - d - c , \qquad \text{under} \quad P_t .$$

Finally, under P_k, $k \neq t$ or M, define $r_k = P_k(A_{tM}(c))/P_k(A_{Mt}(c))$. Then, $u(M) = r_k u(t) = r_k(1 - u(M))v(t)$, and solving for $u(M)$, we find $u(M) = r_k v(t)/(1 + r_k v(t)) < r_k v(t)$, and $u(k) = (1 - u(M))v_k > (1 - r_k v(t))v_k$, which, by the inductive hypothesis that the $M-1$ state machine has error less than d, assures us that

(3) $$u(k) > (1 - r_k d)(1 - d) , \qquad \text{under} \quad P_k .$$

The proof is completed by showing that given $e > 0$, positive numbers c and d can be selected so that from (1), (2), and (3), $u(j) > 1 - e$ under P_j, $j = 1, 2, \cdots, M$. First select c so that $(M-1)c < e$ (and thus $u(M) > 1 - e$ under P_M) and also so that $c < e/2$. Note that if $d < e/2$, $u(t) > 1 - e$, under P_t. c having been selected, the r_k's are determined and, letting $r = \max\{r_k : k = 1, \cdots, M-1, k \neq t\}$ choose d' so that $(1 - rd')(1 - d') \geqq 1 - d'(r+1) > 1 - e$ and observe that if $d < \min\{e/2, d'\}$ then by (2) and (3) under P_j, $u(j) > 1 - e$, for $j < M$.

PROPOSITION 2. *If, for each triple of hypothesis numbers i, j, k, $i \neq j$, there are sets A_{ij} (disjoint in j) such that*

(4) $$P_i[A_{ij}]/P_i[A_{ji}] < 1 \qquad \textit{and} \qquad P_k[A_{ij}] > 0 ,$$

then the FA *hypothesis testing problem is solvable.*

OUTLINE OF PROOF. We only sketch the proof inasmuch as it follows the ideas of the preceding proof. Again, the argument depends on finite induction. For $M = 2$, consider the machine in Figure 2. Observe that if T_{ij} is the expected time for passage from $i1$ to $j1$, then T_{ij} is the expected waiting time for N

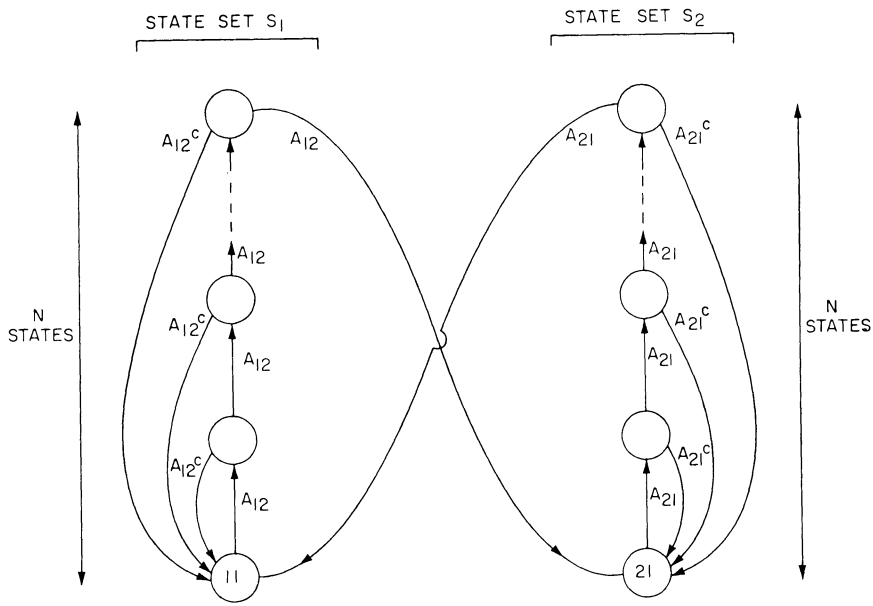

FIG. 2. FA for two hypothesis problem.

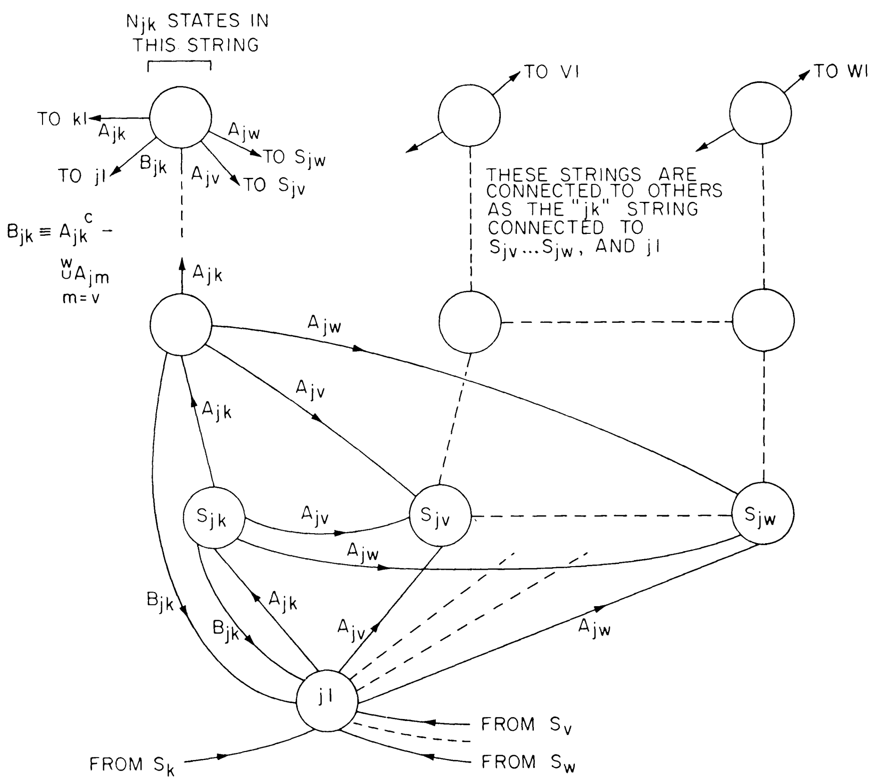

FIG. 3. Construction of state set S_j for multihypothesis tester.

successive occurrences of A_{ij}, whose formula is found in Feller ([5], page 321). If $u(\cdot)$ is the invariant probability, one readily confirms that

$$u(S_1)/u(S_2) = T_{12}/T_{21}\,, \tag{5}$$

and as N increases, this ratio converges to infinity or zero according to whether the underlying probability is P_1 or P_2, and consequently we have solvability.

Now assume that for every positive number d, and for any $M - 1$ hypothesis problem satisfying the conditions of the proposition, there is a FA (whose generic set S_j is illustrated in Figure 3) which has probability of error less than d. We take $N_{jk} = N_{kj}$ and observe that the FA of Figure 2 is a special case of this machine. As in the proof of Theorem 1, given a hypothesis set $\mathscr{P} = \{P_1, \cdots, P_M\}$, we begin by restricting attention to a subclass of the above FA's which solve the testing problem for $\{P_1, \cdots, P_{M-1}\}$ and additionally, for some fixed hypothesis number t and under P_M, have invariant probabilities v which satisfy

$$v(S_t) \geqq v(S_j)\,, \qquad 1 \leqq j \leqq M-1\,. \tag{6}$$

To modify this class of FA's to solve the M hypothesis problem, first append a string of N states to $t1$ in the manner illustrated in Figure 4. For now, leave out state set S_M and branches between S_M and S_t. Verify that this modification has been done so that the transition probabilities between those states Q of the FA not shown in Figure 4 are unchanged by the modification. Further, if $q \in Q$ and $j \in Q^c$, the probability of transition from j to q in the partially

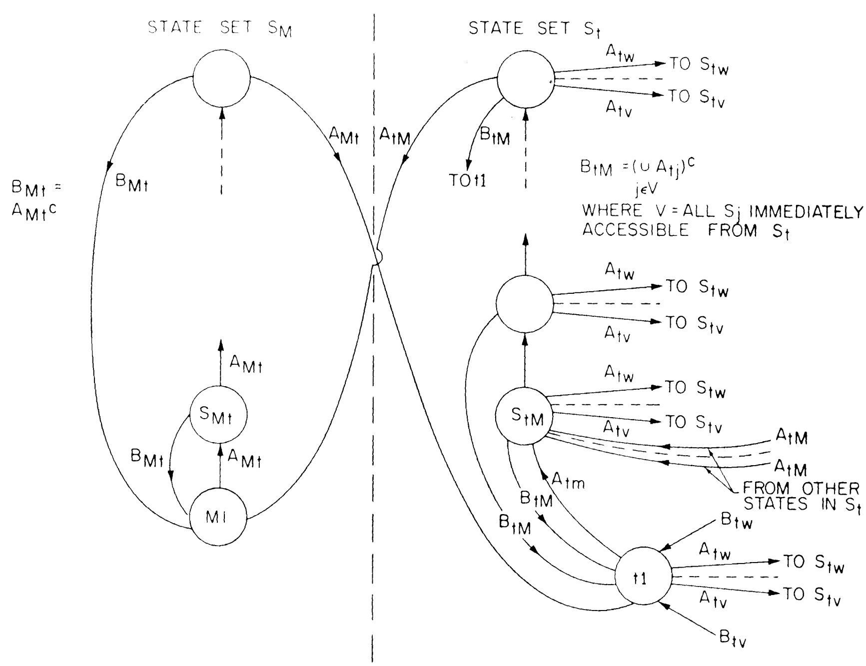

FIG. 4. Augmenting FA for M hypothesis problem.

augmented FA is the same as that from $t1$ to j in the original. From these and related developments, one may conclude that under any member of $\mathscr{P}$, the invariant probabilities $v(S_j)$, $1 \leqq j \leqq M-1$, are unchanged by the partial modification.

Now to complete the modification attach set S_M as shown in Figure 4 and denote invariant probabilities of this final machine by $u(\cdot)$. By studying the expected waiting time in S_t (not counting excursions to S_j, $j \neq M$) until passage from $t1$ to $M1$ occurs, one may verify that under $P_k \in \mathscr{P}$,

$$(7) \qquad Q_k(N) \equiv u(S_M)/u(S_t) = (P_k[A_{tM}]/P_k[A_{Mt}])^N \frac{1-(P_k[A_{Mt}])^N}{1-(P_k[A_{tM}])^N}\,.$$

Also, from generalization of the lemma, $u(S_j) = (1 - u(S_M))v(S_j)$, $1 \leqq j \leqq M$. Recalling that $v(S_t) \geqq 1/M - 1$ under P_M, verify that (under P_M),

$$(8) \qquad u(S_M) \geqq ([(M-1)/Q_M(N)] + 1)^{-1}\,.$$

Let d denote the probability of error of the original FA for the $M-1$ hypothesis problem. Then under P_t, it is quickly seen that

$$(9) \qquad u(S_t) > (1 - Q_t(N))(1-d)\,.$$

Similarly,

$$(10) \qquad u(S_k) \geqq (1 - Q_k(N)d)(1-d)\,.$$

The argument is completed by showing that given $e > 0$, N and then d may be selected so that the resulting FA has probability of error less than e. Notice that as N grows, $Q_M(N) \to \infty$ and $Q_t(N) \to 0$. Therefore, N may be selected so that by (8), under P_M, $u(S_M) > 1 - e$ and also $Q_t(N) < e/2$. For N so chosen, $d < e/2$ (and so by (9), under P_t, $u(S_t) > 1 - e$) in such a manner that the right-hand side of (10) is greater than $1 - e$ for each $k \neq t$, M. From this we conclude that the probability of error of the FA is less than e, i.e., under P_j, $1 \leqq j \leqq M$,

$$(11) \qquad u(S_j) > 1 - e\,.$$

3. Randomization and the domain of FA solvability. While arbitrary finite subsets of some important probability families satisfy Assumption 1 (for instance a multivariate normal family with fixed covariance matrix), subsets of others (notably a multivariate normal family with fixed mean vector) do not. We can, nevertheless, satisfy Assumption 1 for this and many other families by adhering to the tradition of Hellman and Cover [6] and Samuels [10] in allowing randomization. To achieve this, we assume that the events in the product field $\mathscr{A}' = \mathscr{A} \times \mathscr{B}$, where $\mathscr{A}$ is the σ-field of the hypothesis testing problem and $\mathscr{B}$ is the Borel set for the unit interval. For $P_j \in \mathscr{P}$, the associated probability $P_j' \in \mathscr{P}'$ on $\mathscr{A}'$ is the product probability of P_j and P_U where P_U is the uniform probability distribution. If $\mathscr{P}$ is some given set of probabilities on $\mathscr{A}$, $\mathscr{P}'$ denotes the associated set on $\mathscr{A}'$ obtained from $\mathscr{P}$ as just described.

Proposition 3. *If $\mathscr{P}$ is a finite set of probabilities such that for each distinct pair of indices,*

$$\inf_A P_i[A]/P_j[A] \quad \text{or} \quad \inf_A P_j[A]/P_i[A] = 0\,, \tag{12}$$

then $\mathscr{P}'$ satisfies Assumption 1.

Proof. Let P_k, P_j be any two probabilities in $\mathscr{P}$. Suppose $\inf_A P_k[A]/P_j[A] = 0$, and e is any positive number. Let A be some event such that

$$0 < P_i[A]/P_j[A] < e^2/2.$$

We leave it to the reader to see how randomization can be used so that the following method goes through even if $P_i[A] = 0$. Let B be any event in $\mathscr{B}$ such that $P_U[B] = (e/2)P_j[A]$. Then

$$A'_{ij} \equiv A \times [0, 1] \qquad \text{and} \qquad A'_{ji} \equiv \mathscr{X} \times B\,.$$

So

$$P_j'[A'_{ji}]/P_i'[A'_{ji}] = P_U[B]/P_j[A] = \frac{(e/2)P_j[A]}{P_j[A]} < e\,,$$

and

$$P_i'[A'_{ij}]/P_i'[A'_{ji}] = P_i[A]/P_U[B] < \frac{(e^2/2)P_j[A]}{e/2P_j[A]} = e\,.$$

Thus we have found A'_{ij} such that for each i, j

$$P_i'[A'_{ij}]/P_i'[A'_{ji}] < e$$

and such that A'_{ij} is either of the form $A \times [0, 1]$, $(A \in \mathscr{A})$, or $\mathscr{X} \times B$, $(B \in \mathscr{B})$. It remains to establish sets A''_{ij} retaining the inequality properties of A'_{ij} and such that A''_{ij} is disjoint in j for each i. Toward this end, let $B_1, \cdots, B_M$ be M disjoint events in $\mathscr{B}$ such that

$$P_U[B_j] = 1/M\,, \qquad 1 \leqq j \leqq M\,.$$

If A'_{ij} is of the form $A \times [0, 1]$, we set $A^*_{ij} = A \times B_j$ and otherwise $A^*_{ij} = \mathscr{X} \times B'$, where $B' \subset B_j$ is some event such that $P_U[B'] = M^{-1}P_U[B]$. It is evident that $P_k'[A^*_{ij}] = P_k[A'_{ij}]/M$ and so the inequalities of Assumption 1 still hold, and the A^*_{ij} are disjoint in j.

At this point we have seen that for every positive number e, and distinct indices i, j, we can choose $A_{ij}(e) \equiv A^*_{ij}$ from $\mathscr{A}'$ in such a way that

$$P_i'[A_{ij}(e)]/P_i'[A_{ji}(e)] < e$$

and the $A_{ij}(e)$'s are disjoint in j. In the development above, let us specifically denote the dependence of A on e by writing $A(e)$. It is possible to select the $A(e)$'s so that if $e' > e$, $A(e) \subset A(e')$. By further making obvious constraints on the selection of $B' = B(e)$, things can be arranged so that for all positive e, e'

$$A_{ij}(e)'' \qquad \text{and} \qquad A_{ik}(e')''$$

are disjoint if $k \neq i$ or j. This completes the realization of the properties of Assumption 1.

Let us call condition (12) Assumption 1′, by virtue of its equivalence (via randomization) to Assumption 1. While we have not attempted a comprehensive cataloguing of different probability families with respect to Assumption 1′, we have done some study in this direction. For example, it is possible to quickly demonstrate that many families of the Koopman exponential form do obey Assumption 1′. A family $\mathscr{F} = \{f(x, \theta)\}$ of probability densities with respect to either counting measure or Lebesgue measure and indexed by a real vector θ is of the Koopman type if there exist real-valued functions $h(x)$, $B(\theta)$, and real vector-valued functions, $R(x)$ and $Q(\theta)$ such that

$$f(x, \theta) = B(\theta)h(x) \exp (Q(\theta)^T R(x)) .$$

PROPOSITION 4. *Let $\mathscr{P}$ be a finite subset of a family of the Koopman type, and take $R = \{R(x): h(x) > 0\}$ and $Q = \{Q(\theta) - Q(\theta'): \text{all } \theta, \theta'\}$. Then $\mathscr{P}$ satisfies Assumption* 1′ *if and only if the projection of R on q is unbounded for each $q \in Q$.*

We omit the easy proof, but mention that the proposition implies that the following families satisfy Assumption 1′:

(i) Multivariate normal	(ii) Gamma	(iii) Geometric
(iv) Poisson	(v) Rayleigh	(vi) Lognormal

Sagalowicz [9] shows that the Cauchy family fails to satisfy Assumption 1′.

Finally let us note that for any two distinct probability distributions P_i and P_j, for some A_{ij}, A_{ji}

$$P_j(A_{ji})/P_j(A_{ij}) , \qquad P_i(A_{ij})/P_i(A_{ji}) < 1 .$$

Using randomization, we can readily construct events A'_{ij} which also satisfy the above equation and are furthermore disjoint in the second subscript and consequently (using randomization) the conditions of Proposition 2 hold for *any* finite set $\mathscr{P}$ of probability functions.

4. Sequential analysis by FA's. The object of this section is to show how FA's can be designed to implement sequential decision functions (in the tradition of A. Wald [12]). Specifically, for each hypothesis in $\mathscr{P}$, we associate one absorbing state in a FA. Now the FA operates by eventually entering one of these absorbing states, and when this happens, the "terminal decision" is taken to be the hypothesis labeled by the subscript of the particular absorbing state entered.

In this section, therefore, a hypothesis testing problem and a FA having been specified, we define the *probability of error* (for the sequential analysis problem) to be the maximum (over $P_j \in \mathscr{P}$), probability, under P_j, that the terminal state entered is not labeled j.

THEOREM 2. *Given any sequential analysis testing problem and any positive number e, there exists a randomized* FA *having probability of error less than e; if Assumption* 1′ *is satisfied, the maximum number of states required is twice the number of probabilities in $\mathscr{P}$.*

PROOF. Let A_k denote the transition matrix under P_k of a FA having probability of error (in the sense of Section 3) less than $e/4$ for the testing problem. If u_k is the vector of invariant probabilities for A_k, then u_k is the square matrix whose rows are the vectors u_k. We will use the matrix norm $||A - B|| = \sum_{i,j} |a_{ij} - b_{ij}|$. From the foregoing definitions and elementary properties of Markov chains, there is some N such that $||U_k - A_k{}^n|| < e/4$ for every $n > N$ and $1 \leqq k \leqq M$, where as usual, M is the number of hypotheses. Let h be the parameter of some geometric variable T' such that $P[T' < N] < e/2$. The FA above is now augmented so that each state in S_j has a branch leading to a state labeled q_j, $1 \leqq j \leqq M$, which is followed whenever some event B_h in the randomization experiment, occurs, where $P_{l}[B_h] = h$. We leave it to the reader to develop the details of how the B_h's are embedded in the product field $\mathscr{P}'$, and how they are chosen to preserve the disjointness of the events A_{ij}.

With this construction, we have, letting T' be the variable (with the geometric distribution) of the time immediately preceding entry into $\{q_k: 1 \leqq k \leqq M\}$,

$$\begin{aligned} P_j[\text{terminal state} \neq q_j] &\leqq P[T' \leqq N] + P[T' \geqq N \text{ and } \mathbf{s}(T') \notin S_j] \\ &= P[T' \leqq N] + P[T' \geqq N] P[\mathbf{s}(T') \notin S_j \,|\, T' \geqq N] . \end{aligned}$$

From the definition of N, for every initial state θ, every $t > N$, and set Q of states,

$$P_j[\mathbf{s}(t) \in Q] - u_j(Q) \leqq ||U_j - A_j{}^t|| < e/4 .$$

In particular, $u_j(S_j{}^c) < e/4$ and thus $P[\mathbf{s}(t) \notin S_j] < e/2$.

In summary, recalling that $P[T' \leqq N] < e/2$,

$$\begin{aligned} P[\text{terminal state} \neq q_j] &< e/2 + P[T' > N] P[\mathbf{s}(T') \notin S_j / T' > N] \\ &\leqq e/2 + e/2 = e . \end{aligned}$$

5. FA adaptive control. Let $\{T_i\}$ be a strictly increasing (perhaps random) sequence of integers with $T_1 = 1$, and take $\{P(i)\}$ to denote a sequence with range in the set $\mathscr{P}$ of probabilities. In this section we will suppose the random sequence $\{X_i\}$ to be generated so that for $T_j \leqq n \leqq T_{j+1}$, X_n is i.i.d., with probability law $P(j) \in \mathscr{P}$. Using methods of Section 4 for every positive e we will be able to give conditions on T_j such that one can design a FA of the type in Figure 3 with average probability of error less than e. That is, the proportion of time the FA is in the set S_j having the same index as $P(n) \in \mathscr{P}$ converges asymptotically to a number greater than $1 - e$, even for time changing hypotheses.

An engineering application of a FA with such capabilities might be in an assembly line setting where one portion of the process has a number of different possible modes of operation, each of which is acceptable provided the operation at a later section of the line takes into account which of these modes is indeed in effect (and uses the procedure appropriate for the mode at the earlier section). Specifically, for example, waste treatment plants have several characteristic types of influx, each of which induces a particular type of decay flora and each of which requires, therefore, addition of different sorts of reagents. One might seek as cheap a computer as possible which would examine signals from sensors to determine which mode of influx is currently active, and send the result of this decision to stations "downstream" in the process. If the sensors are not totally effective for distinguishing between the processes (i.e., the probabilities of the various modes are not orthogonal), a statistical point of view would be appropriate. If further, the mode described above is time-changing, the theory of this section seems indicated.

One problem (and the one apparently suggested by the assembly line description) of the genre we have been discussing arises when we assume $T_{j+1} - T_j$ is i.i.d., as some fixed integer-valued random variable T_0. Formally,

PROPOSITION 5. *Let $(\mathscr{Z}, \mathscr{A})$ and $\mathscr{P}$ be the parameters of an* FA *hypothesis testing problem whose hypotheses are time-changing. Suppose the times of change, T_j, have the property that $T_{j+1} - T_j$ are* i.i.d., *as a random variable T_0. In the notation of Section* 4, *let A_j be the transition matrix, under P_j, of some* FA *having probability of error less than $e/2$ for the fixed hypothesis problem, and let N be a number such that for all $t > N$, $\|U_j - A_j^t\| < e/2$, $1 \leqq j \leqq M$. Then for this* FA,

(13) *the asymptotic relative frequency of error* $\leqq NE[T_0^{-1}](1 - e) + e$.

PROOF. The trajectory $\mathbf{s}(k)$ of the FA above is conveniently partitioned into blocks, the kth such block consisting of the segment $\{\mathbf{s}(i): T_k \leqq i < T_{k+1}\}$. We proceed by proving that no matter what $P(j)$ is, the expected average error of the jth block is bounded by (13).

Let k, $1 \leqq k \leqq M$, be the index of $P(j)$. Define $\theta(i) = 1$ if $\mathbf{s}(i) \notin S_k$ and 0 otherwise. Then

$$\sum \theta(i): T_j \leqq i \leqq n \leqq N + \sum \theta(i): T_j + N < i \leqq n .$$

But, if $u_k(\cdot)$ is the invariant probability of the transition matrix A_k, for $i \geqq T_j + N$

$$E[\theta(i)] \leqq u(S_k^c) + (e/2) < e ,$$

and consequently for n any positive number

$$E[\sum \theta(i): T_j \leqq i \leqq n] \leqq N + (n - N)e ,$$

from whence the expected average error in the block is bounded by

$$\textstyle\sum_{n=1}^{\infty}[(N + e(n - N))/n]P_{T_0}(n) = E(1/T_0)N(1 - e) + e ,$$

which, by the law of large numbers, must therefore also be bound to the asymptotic relative frequency of error for the entire $\{\mathbf{s}(i)\}$ sequence.

There are many ways in which these developments may be extended. Under slightly further restrictions on T_0, one may find bounds as above for the probability of error at some fixed sample time n. Another avenue for extension concerns the cases in which actions are chosen at each time n on the basis of a history $\{X_i\}_{i \leq n}$ of observations where $\{X_i\}$ is as earlier in this section. At each time n and selection a, the loss $L(a, j)$ is realized eventually (but not in time to use it for the statistical problem), j being the index of the "true" underlying distribution at time n. Then, letting $L^*(j)$ denote the $\min_a L(a, j)$ and π_j be the relative frequency of P_j in $\{P(n)\}$, the best possible performance would result in the average loss $\sum_{j=1}^{M} \pi_j L^*(j) = L^*$. With respect to the time-changing model, it is possible to compute bounds on the average error of a FA controller for such a control problem. The FA controller is supposed to choose the best action a_j^* for j when $s(n)$ is in S_j.

6. Acknowledgments. The author completed the major portion of this study while under the sponsorship of a National Academy of Science Postdoctoral Research Fellowship at the Naval Postgraduate School. Also, the work profits from the author's conversations with D. Sagalowicz and T. Cover.

REFERENCES

[1] Arbib, M. (1969). *Theories of Abstract Automata.* Prentice-Hall, Englewood Cliffs.

[2] Cover, T. M. (1968). A note on the two-armed bandit problem with finite memory. *Information and Control* **12** 371–377.

[3] Cover, T. M. (1969). Hypothesis testing with finite memory. *Ann. Math. Statist.* **40** 828–835.

[4] Cover, T. M. and Hellman, M. E. (1970). The two-armed bandit problem with time-invariant finite memory. *IEEE Trans. Information Theory*, **IT-6** 185–197.

[5] Feller, W. (1969). *An Introduction to Probability Theory and its Applications*, (3rd. ed.). Wiley, New York.

[6] Hellman, M. E. and Cover, T. M. (1970). Learning with finite memory. *Ann. Math. Statist.* **41** 765–782.

[7] Isbell, J. S. (1959). On a problem of Robbins. *Ann. Math. Statist.* **30** 606–610.

[8] Robbins, H. (1956). A sequential decision problem with finite memory. *Proc. Natl. Acad. Science* **42** 920–923.

[9] Sagalowicz, D. (1970). Hypothesis testing with finite memory. Ph. D. thesis, Electrical Engr. Dept., Stanford Univ.

[10] Samuels, S. M. (1969). Randomized rules for the two-armed bandit with finite memory. *Ann. Math. Statist.* **39** 2103–2107.

[11] Smith, C. V. and Pyke, R. (1965). The Robbins–Isbell two armed bandit problem with finite memory. *Ann. Math. Statist.* **36** 1375–1386.

[12] Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Systems and Industrial Engineering
University of Arizona
Tucson, Arizona 85721