LEARNING WITH FINITE MEMORY

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING

AND THE COMMITTEE ON THE GRADUATE DIVISION

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Martin Edward Hellman

March 1969

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Thomas M Cover

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Joseph W. Goodman

Approved for the University Committee on the Graduate Division:

Virgil K. Whitaker

Dean of the Graduate Division

Abstract

This paper lays a foundation for the theory of hypothesis testing with finite memory by solving the following problem under a finite memory constraint. Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables drawn according to a probability measure $\mathcal{P}$. The problem is to decide between the two simple hypotheses $\mathcal{P} = \mathcal{P}_0$ and $\mathcal{P} = \mathcal{P}_1$. The X_i's are observed sequentially and a new decision must be formulated after each observation. It may be shown that any rule that minimizes the probability of error requires infinite memory (in the nondegenerate case), even if sufficient statistics are utilized. Motivated by a desire to keep memory finite, let the data be summarized after each new observation by an m-valued statistic T which is updated according to the rule $T_n = f(T_{n-1}, X_n)$, where f may be a randomized function. Let the decision rule take action $d(T_n)$ at time n. The objective is to find the pair (f,d) which minimizes the asymptotic probability of error $P(e)$. This algorithm may be thought of as a finite-state automaton, in which the inputs are the observations, the outputs are the decisions, and the states constitute the memory.

Letting $\bar{\ell}$ and $\underline{\ell}$ be the (a.e.) maximum and minimum likelihood ratios, define $\gamma = \bar{\ell}/\underline{\ell}$. Furthermore let π_0 and π_1 be the a priori probabilities of the two hypotheses and m be the number of states in memory. Then it is shown that P^* is the greatest lower bound for $P(e)$, where

$$P^* = \min \left\{ \frac{2\sqrt{\pi_0 \pi_1 \gamma^{m-1}} - 1}{\gamma^{m-1} - 1} \ , \ \pi_0 \ , \ \pi_1 \right\} \ .$$

Thus it is seen that γ is a measure of the separation between the two hypotheses, and that P^* decreases almost exponentially in m. Moreover, a class of ϵ-optimal automata is demonstrated (i.e., for any $\epsilon > 0$ there exists an automaton in this class with $P(e) \leq P^* + \epsilon$). It is further shown that, except for certain degenerate cases, no machine can actually achieve P^* and that the ϵ-optimal class is essentially unique. The solution is ϵ-optimal for both the Bayesian and Neyman-Pearson formulations of the problem.

CONTENTS

	Page
Abstract	iii
Introduction	1
A Lower Bound for P(e)	7
A Class of ϵ-Optimal Automata	16
Uniqueness of the ϵ-Optimal Class	29
Conclusions	40
Appendix I: Definitions and Facts from the Theory of Markov Chains	42
Appendix II: Extensions of Theorem 1 to the Nonergodic Case	45
Appendix III: Rate of Convergence	51
Appendix IV: The case of $\gamma = \infty$	56
References	58

ILLUSTRATIONS

Figure

1.	Saturable counter	17
2.	Distribution on the automaton's states	18
3.	The canonical form of the ϵ-optimal machine	26

ACKNOWLEDGEMENT

Sincerest thanks are given to Professor Thomas M. Cover, without whose guidance this work would not have been possible. Thanks are also due to Professor J. Goodman and Professor M. Arbib for their patient reading of the manuscript.

Introduction.

Let $X_1, X_2,$ be a sequence of independent, identically distributed random observations drawn according to a probability measure $\mathcal{P}$ defined on an arbitrary probability space. Consider the simple hypothesis testing problem

$$H: \mathcal{P} = \mathcal{P}_0 \quad \text{vs} \quad H_1: \mathcal{P} = \mathcal{P}_1$$

Let the prior probabilities of the null hypothesis H_0 and the alternative hypothesis H_1 be denoted by π_0 and π_1 respectively. The usual goal is to find a sequence of decision rules $d_1(X_1)$, $d_2(X_1,X_2),\ldots$ which minimizes the asymptotic probability of error $P(e)$. (Although a Bayesian formulation will be used for the development in this paper, the results of the theory apply directly to the Neyman-Pearson formulation in which the probability of error under H_0 is fixed and the probability of error under H_1 is to be minimized.)

Since d_n may depend on $(X_1,X_2,\ldots,X_n)$, as n increases, the amount of data to be stored increases without bound. Some means of data reduction may be desirable. Sufficient statistics can sometimes be used to reduce the required size of memory. When used, such statistics lose no information. However, as will be shown in the following example, apparent data reduction is sometimes misleading.

Consider the problem in which X is univariate normally distributed with variance equal to one. Let $\pi_0 = \pi_1 = 1/2$. Under H_0, let the distribution have mean $\mu = +1$; and under H_1 let the mean $\mu = -1$. A statistic, sufficient for this problem, is $T_n = \sum_{i=1}^{n} X_i$, where T_n is the value of the statistic after n observations. A simple optimal

decision scheme is for d_n to decide H_0 if $T_n \geq 0$ and to decide H_1 if $T_n < 0$. Furthermore, a simple updating scheme is given by

$$T_{n+1} = T_n + X_{n+1} \quad . \tag{1}$$

Thus T_n contains all the desired information about $(X_1, X_2, \ldots, X_n)$, and only T_n need be remembered. Thus at time n, instead of storing n real numbers, it is necessary to store only one. This is an apparent n-fold reduction in the required data.

However, T is real-valued. Thus infinite storage is needed for it alone. Furthermore there exist [1] uniformly continuous one-to-one mappings of $\mathbb{R}^n$ onto $\mathbb{R}$, so that if memory can store one real number it can store any number of real numbers. One might think that in spite of this theoretical lack of reduction of the data, there might be a real saving in the sense that a truncated version of T would yield acceptable probability of error. However, here too, T fails. No matter how T is truncated, the results of this paper show how to construct a better rule.

In order to address the finite memory constraint, consider the family of all learning algorithms of the type

$$T_n = f(T_{n-1}, X_n) \tag{2}$$

where X_n is the n^{th} observation, T_n is the state of the memory at time n, and f is a function (perhaps randomized), independent of n and the data. The algorithm is said to have <u>finite memory</u> of length m if T is m-valued (i.e., $T_n \in \{1,2,\ldots,m\}$ for $n = 1,2,\ldots$). For the classification problem it will also be necessary to specify a decision

rule $d: \{1,2,\ldots,m\} \to \{H_0, H_1\}$ which takes action $d(T_n)$ at time n. It will be seen that no randomization of d is required for the optimal procedures. However, it is generally true that an ϵ-optimal f must be random. By analogy to the previous example, T may be thought of as an m-valued statistic, and the problem is to find the algorithm for updating this statistic which loses the least amount of information.

It may also be seen that the pair (f,d) describes a finite-state machine with inputs X_n and outputs $d_n = d(T_n)$, $n = 1,2,\ldots$. The independence of the X_i's sets up a Markov process on the state space $S = \{1,2,\ldots,m\}$, as can be seen from the following recasting of the (f,d) description: The action of f may be prescribed by a (perhaps infinite) family of stochastic transition matrices indexed by x,

$$P(x) = [p_{ij}(x)], \; i,j = 1,2,\ldots,m \quad , \tag{3}$$

where $\sum_{j=1}^{m} p_{ij}(x) = 1$, and $p_{ij}(x) \geq 0$, $\forall$ i,j,x. Here $p_{ij}(x)$ is the probability that $T_n = j$ given that $T_{n-1} = i$ and that $X_n = x$ is observed. Taking the expectation over x, it is found that

$$P^{(0)} = \int P(x) \; d\mathcal{P}_0(x)$$

and

$$P^{(1)} = \int P(x) \; d\mathcal{P}_1(x) \tag{4}$$

are the state transition probability matrices under H_0 and H_1 respectively. The stationary or long-run probability distribution on the states may then be given by

$$\underline{\mu}^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_m^{(0)})$$

and

$$\underline{\mu}^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)}, \ldots, \mu_m^{(1)})$$

where $\underline{\mu}^{(0)}$, $\underline{\mu}^{(1)}$ satisfy the matrix equations

$$\begin{aligned} \underline{\mu}^{(0)} &= \underline{\mu}^{(0)} P^{(0)} \\ \text{and} \qquad & \qquad (5) \\ \underline{\mu}^{(1)} &= \underline{\mu}^{(1)} P^{(1)} \quad . \end{aligned}$$

The resulting long-run probability of error is now simply given by

$$P(e) = \pi_0 \sum_{i \in S_1} \mu_i^{(0)} + \pi_1 \sum_{i \in S_0} \mu_i^{(1)} \quad , \qquad (6)$$

where $S_j = \{i: \ d(i) = H_j\}$, $j = 0,1$, are the decision regions induced by the decision rule d. Note that the (f,d) description and the $(P(x),d)$ description are equivalent. In this paper $P^* = \inf_{(f,d)} P(e)$ will be found as a function of m, and an ϵ-optimal class of (f,d)'s will be demonstrated. (That is, it will be shown that for any $\epsilon > 0$ there exists an (f,d) in this class whose $P(e) \leq P^* + \epsilon$.) It will also be shown that, in general, no optimal (f,d) exists.

Using different methods the time-varying learning with finite memory algorithm

$$T_n = f(T_{n-1}, X_n, n), \quad T_n \in \{1,2,\ldots,m\} \qquad (7)$$

has been shown in Cover [2] to have $P^* = 0$, for a memory of size $m = 4$. Thus there exist learning rules for a time-varying finite memory which

yield asymptotically zero probability of error. No such hope exists in the time-invariant problem treated here.

Thus solutions exist if the automaton is allowed to be time-varying or adaptive.[3] However, a time-varying automaton requires a clock which can count without bound to keep track of n. Since this can only be approximated in practice, this paper will be restricted to time-invariant automata. Adaptive automata have additional memory in their variable structure and so will also be excluded. Thus in this paper the word automaton or machine will be understood to mean only a time-invariant, non-adaptive automaton.

Several authors in the Russian literature [4], [5], [6] have investigated the behavior of automata in random media. However, their work is primarily devoted to the analysis of the behavior of various ad hoc machine designs. Moreover, the problem formulations are more properly in the area of the sequential design of experiments (the so-called two-armed bandit problem) than in the area of hypothesis testing. This work is nonetheless interesting because of the similarity of the formalism to that of the problem considered here. Under an alternative definition of finite memory (in which the memory may consist solely of the last n observations) the two-armed bandit problem has been attacked by Robbins [7], Isbell [8], Smith and Pyke [9], Samuels [10], and Cover [11]. The hypothesis testing problem under the constraint that the memory be one dimensional (a single updatable real number) has been discussed by Spragins [12] and Fralick [13]. The latter work stimulated work presented here.

As has been mentioned, the states occupied by the automaton form a Markov chain. Several definitions similar to those used in the theory

of Markov chains (e.g., irreducible, ergodic, etc.) are needed and are given in the Appendix I.

A Lower Bound for P(e).

There always exists a dominating measure ν such that $\mathcal{P}_0$ and $\mathcal{P}_1$ are absolutely continuous with respect to ν. ($\mathcal{P}_0 + \mathcal{P}_1$ is such a measure.) Thus there exist densities $f_0(x)$ and $f_1(x)$ which are the respective Radon-Nikodym derivatives of $\mathcal{P}_0$ and $\mathcal{P}_1$ with respect to ν, and the likelihood ratio (l.r.), $\ell(x) = f_0(x)/f_1(x)$ is well defined on the extended real line, almost everywhere $\mathcal{P}_0 + \mathcal{P}_1$.

Definition: The almost everywhere (a.e.) least upper bound on the l.r., $\overline{\ell}$, is defined by

$$\overline{\ell} = \sup_{\mathcal{P}_0(A) + \mathcal{P}_1(A) > 0} \frac{\mathcal{P}_0(A)}{\mathcal{P}_1(A)} \tag{8}$$

Similarly, the a.e. greatest lower bound on the l.r., $\underline{\ell}$, is defined by

$$\underline{\ell} = \inf_{\mathcal{P}_0(A) + \mathcal{P}_1(A) > 0} \frac{\mathcal{P}_0(A)}{\mathcal{P}_1(A)} \tag{9}$$

Remark: For nicely behaved f_0, f_1, it is seen that $\overline{\ell}$ and $\underline{\ell}$ are merely the maximum and minimum values of the likelihood ratio.

Lemma 1: The ratio of the probability p_{ij}^0 of transition from state i to state j under H_0, to the probability p_{ij}^0 of the same transition under H_1, satisfies the inequality

$$\underline{\ell} \leq \frac{p_{ij}^0}{p_{ij}^1} \leq \overline{\ell} \tag{10}$$

Remark: If both p_{ij}^0 and p_{ij}^1 are zero, their ratio is undefined.

Proof: p_{ij}^0 is equal to $\int_{\mathcal{X}} p_{ij}(x)\, f_0(x)\, d\nu(x)$ and p_{ij}^1 is given by the same expression with f_1 substituted for f_0. Since $f_0(x) = \ell(x) f_1(x)$,

$$\frac{p_{ij}^0}{p_{ij}^1} = \frac{\int_{\mathcal{X}} p_{ij}(x)\, \ell(x)\, f_1(x)\, d\nu(x)}{\int_{\mathcal{X}} p_{ij}(x)\, f_1(x)\, d\nu(x)} \tag{11}$$

$$\leq \frac{\bar{\ell} \int_{\mathcal{X}} p_{ij}(x)\, f_1(x)\, d\nu(x)}{\int_{\mathcal{X}} p_{ij}(x)\, f_1(x)\, d\nu(x)} = \bar{\ell}$$

Similarly, replacing $\ell(x)$ by its a.e. lower bound $\underline{\ell}$, the other desired inequality is proved. Q.E.D.

Note that if $\bar{\ell}$ is infinite the proof requires slight modifications. This will also be true in later theorems, when the parameters involved are infinite. The modifications are straightforward and are thus deferred to Appendix IV.

It will be recalled that

$$P(e) = \pi_0 \sum_{i \in S_1} \mu_i^0 + \pi_1 \sum_{i \in S_0} \mu_i^1 \tag{6}$$

is to be minimized. For a state $i \in S_0$ is is desired that μ_i^0 be much larger than μ_i^1, and for $i \in S_1$ it is desired that μ_i^0 be much smaller than μ_i^1. Therefore to obtain a lower bound on P(e) it would be helpful to know just how large and how small the state likelihood ratio (s.l.r.) μ_i^0/μ_i^1 can be. The following lemma will be useful in solving that problem:

Lemma 2: For an ergodic automaton in which the s.l.r.'s, μ_i^0/μ_i^1 are arranged in non-decreasing order the following relation holds:

$$1 \leq \frac{\mu_{i+1}^0/\mu_{i+1}^1}{\mu_i^0/\mu_i^1} \leq (\bar{\ell}/\underline{\ell}) \tag{12}$$

Remark: Since the automaton is ergodic the s.l.r. is defined for all states.

Proof: The lower bound of (12) follows from the assumption that the state likelihood ratios have been arranged in non-decreasing order. To establish the upper bound, suppose that the lemma were false. Then for some $i \in S$,

$$\mu_j^0/\mu_j^1 \leq c, \ \forall_{j \leq i} \quad \text{and} \quad \mu_j^0/\mu_j^1 > c\bar{\ell}/\underline{\ell}, \ \forall_{j > i} \ , \tag{13}$$

where c is the value of the state likelihood ratio for state i.

Now the automaton is in the steady state (s.s.) so that if it is broken into two nonempty disjoint sets, C and C', where $C \cup C' = S$ then the "flow" of probability from C to C' must equal the flow from C' to C. Consequently the net flow is zero. The process is very similar to a diffusion process that is in dynamic equilibrium. The probability of occupation in state j may be thought of as the population of state j, and the probability of transiting from state j to state k is then analogous to the fraction of state j's population that flows to state k.

The flow from C to C' is the sum of the flows from each state in C to each state in C'. Similarly the flow from C' to C is the sum of the individual flows. The flow from a state j to a state k, given that H_0 is the true state of nature, is $\mu_j^0 p_{jk}^0$, and if H_1 is the true state, it is $\mu_j^1 p_{jk}^1$. Thus, if C is set equal to the first i states and C' is the last $n-i$ states and flows are equated, first under H_0 and then under H_1, the following equalities are obtained:

$$\sum_{j\epsilon C}\sum_{k\epsilon C'} \mu_j^0 p_{jk}^0 = \sum_{j\epsilon C'}\sum_{k\epsilon C} \mu_j^0 p_{jk}^0 \quad , \tag{14a}$$

$$\sum_{j\epsilon C}\sum_{k\epsilon C'} \mu_j^1 p_{jk}^1 = \sum_{j\epsilon C'}\sum_{k\epsilon C} \mu_j^1 p_{jk}^1 \quad . \tag{14b}$$

But using the inequalities (10) and (13)

$$\sum_{j\epsilon C}\sum_{k\epsilon C'} \mu_j^0 p_{jk}^0 \leq \sum_{j\epsilon C}\sum_{k\epsilon C'} (c\mu_j^1)(\bar{\ell} p_{jk}^1) \quad , \tag{15}$$

so that

$$\sum_{j\epsilon C}\sum_{k\epsilon C'} \mu_j^0 p_{jk}^0 \leq c\bar{\ell} \sum_{j\epsilon C}\sum_{k\epsilon C'} \mu_j^1 p_{jk}^1 \quad . \tag{16}$$

Similarly,

$$\sum_{j\epsilon C'}\sum_{k\epsilon C} \mu_j^0 p_{jk}^0 > (c\bar{\ell}/\underline{\ell})\underline{\ell} \sum_{j\epsilon C'}\sum_{k\epsilon C} \mu_j^1 p_{jk}^1 \quad . \tag{17}$$

But using (14a) the left sides of (16) and (17) are equal, and using (14b) the right sides are equal, a contradiction.

Note that the ergodicity of the machine is used in obtaining (17), for if the machine is not ergodic, there exists a partition of S into sets C and C' such that there is no flow from either one to the other. Then none of the ratios $\mu_j^0 p_{jk}^0 / \mu_j^1 p_{jk}^1$ used to obtain (17) are defined.

This is not to say that non-ergodic automata have no restrictions on how fast their s.l.r.'s increase. But a different sort of condition and proof are needed (see Appendix II).

Definition: The spread of an automaton is the ratio of its maximum s.l.r. to its minimum s.l.r.

Theorem 1: The spread of an m state automaton is less than or equal to γ^{m-1}, where $\gamma = \bar{\ell}/\underline{\ell}$.

Remark: γ is a measure of the "separation" between H_0 and H_1.

Proof: If the automaton is ergodic, m-1 aplications of lemma 2 yield the desired result. If the automaton is not ergodic it is shown in Appendix II that the theorem still holds. In fact, it is shown that except in certain degenerate cases the spread is strictly less than γ^{m-1} in the non-ergodic case. Since P* is a decreasing function of spread (see proof of Theorem 2, particularly the effect of changing (20)) non-ergodic automata can do no better than ergodic ones.

Theorem 2: For an m-state automaton

$$P^* = \begin{cases} \dfrac{2\sqrt{\pi_0\pi_1\gamma^{m-1}} - 1}{\gamma^{m-1}-1}, & \text{if } \gamma^{m-1} \geq \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_0}\right\} \\ \min\{\pi_0, \pi_1\}, & \text{otherwise} \end{cases} \tag{18}$$

is a lower bound on P(e), where

$$\gamma = \bar{\ell}/\underline{\ell} \quad . \tag{19}$$

Remark: Since $\Sigma\mu_i^0 = \Sigma\mu_i^1 = 1$, not all s.l.r.'s can be greater than one. Therefore, by Theorem 1, no s.l.r. can be greater than γ^{m-1}. Thus, if $\pi_0 > \pi_1$ and the a priori l.r., π_0/π_1, is greater than γ^{m-1}, no machine provides sufficient information to reverse the a priori decision. Similar remarks hold if $\pi_0 < \pi_1$ and π_1/π_0 is greater than γ^{m-1}. In either case, no machine is needed since the trivial rule of deciding

whichever hypothesis has the larger prior probability achieves the lower bound $\min\{\pi_0,\pi_1\}$. Note that $\gamma^{m-1} = \max\left\{\frac{\pi_0}{\pi_1}, \frac{\pi_1}{\pi_0}\right\}$ implies that $P^* = \min\{\pi_0,\pi_1\}$, in agreement with this heuristic discussion.

Remark: If $\pi_0 = \pi_1 = 1/2$, $\gamma^{m-1} > \max\left\{\frac{\pi_0}{\pi_1}, \frac{\pi_1}{\pi_0}\right\} = 1$, for $m \geq 2$. In this case using $(\gamma^{m-1}-1) = (\gamma^{\frac{m-1}{2}} + 1)(\gamma^{\frac{m-1}{2}} - 1)$, Equation (18) reduces to

$$P^* = \frac{1}{\gamma^{(m-1)/2} + 1} \tag{19}$$

Proof of Theorem 2: If k is the minimum s.l.r., then by Theorem 1

$$k \leq \mu_i^0/\mu_i^1 \leq k\,\gamma^{m-1} \quad \forall_{i\in S} \tag{20}$$

Using this equation and letting α be the P(e) under H_0 and β be the P(e) under H_1,

$$\alpha = \sum_{i\in S_1} \mu_i^0 \geq k \sum_{i\in S_1} \mu_i^1 = k(1-\beta) \ ,$$

or
$$\alpha \geq k(1-\beta) \ ; \tag{21}$$

and
$$\beta = \sum_{i\in S_0} \mu_i^1 \geq (1/k\gamma^{m-1}) \sum_{i\in S_0} \mu_i^0 \ ,$$

or
$$\beta \geq \frac{1}{k\,\gamma^{m-1}} (1-\alpha) \ . \tag{22}$$

Multiplying (21) and (22) one obtains

$$\alpha\beta \geq \frac{1}{\gamma^{m-1}} (1-\alpha)(1-\beta) \tag{23}$$

Equivalently

$$\frac{(1-\alpha)(1-\beta)}{\alpha\beta} \leq \gamma^{m-1}, \ \alpha > 0, \ \beta > 0 \tag{24}$$

Equation (24) gives a lower boundary for the operating characteristic (OC) of an automaton. [The OC is the region of achievable (α,β).] Thus the results of this analysis apply equally well to a Newman-Pearson formulation of the problem, since the machines that will be demonstrated in the next section can approach any point on this lower boundary for the OC, not only the point that yields minimum Bayes' risk.

To return to the Bayesian approach, note that the left side of (24) decreases in both α and β. Therefore, $\pi_0\alpha + \pi_1\beta$ will be a minimum when the inequality in (24) is replaced with an equality.

Given π_0, π_1, and γ, minimize $\pi_0\alpha + \pi_1\beta$ subject to the constraint given by (23) with equality used instead of the weak inequality. This is a problem in the calculus of variations and may be solved using Lagrangian multipliers. Expressed in a form useful for this method, (23) becomes

$$1 - \alpha - \beta + \alpha\beta(1-\gamma^{m-1}) = 0 \tag{25}$$

Defining the Lagrangian

$$J(\alpha,\beta) = \pi_0\alpha + \pi_1\beta + \lambda[1 - \alpha - \beta + \alpha\beta(1- \gamma^{m-1})] \tag{26}$$

and setting the partials with respect to α and β equal to zero,

$$\frac{\partial J}{\partial \alpha} = \pi_0 + \lambda[-1 + \beta(1-\gamma^{m-1})] = 0$$

$$\frac{\partial J}{\partial \beta} = \pi_1 + \lambda[-1 + \alpha(1-\gamma^{m-1})] = 0 \qquad (27)$$

are obtained.

Solving (27) for α and β yields

$$\alpha = \frac{\frac{\pi_1}{\lambda} - 1}{\gamma^{m-1}-1} \quad , \quad \beta = \frac{\frac{\pi_0}{\lambda} - 1}{\gamma^{m-1}-1} \qquad (28)$$

Then, substituting the expressions in (28) into the constraint equation, (24) results in

$$\lambda = \sqrt{\frac{\pi_0\pi_1}{\gamma^{m-1}}} \qquad (29)$$

Thus the values of α and β which extremize $\pi_o\alpha + \pi_1\beta$ are obtained by using this value of λ, and are given by

$$\alpha^* = \frac{\sqrt{\frac{\pi_1}{\pi_0}\gamma^{m-1}} - 1}{\gamma^{m-1}-1} \quad , \quad \beta^* = \frac{\sqrt{\frac{\pi_0}{\pi_1}\gamma^{m-1}} - 1}{\gamma^{m-1}-1} \quad , \qquad (30)$$

The resulting extreme is

$$P^*(e) = \frac{2\sqrt{\pi_0\pi_1\gamma^{m-1}} - 1}{\gamma^{m-1}-1} \qquad (31)$$

As in all such problems it must be checked that $P^*(e)$ is a local minimum not a local maximum. If $\gamma^{m-1} > \max\left\{\frac{\pi_0}{\pi_1}, \frac{\pi_1}{\pi_0}\right\}$, $P^*(e)$ is both a local and global minimum, whereas if $\gamma^{m-1} \leq \max\left\{\frac{\pi_0}{\pi_1}, \frac{\pi_1}{\pi_0}\right\}$ one of the endpoints (either $\alpha = 0, \beta = 1$ or $\beta = 0, \alpha = 1$) is the global minimum. Thus P^* as given by (18) is a lower bound on $P(e)$.

A Class of ϵ-Optimal Automata.

Now that a lower bound P^* exists on $P(e)$, the rather important question arises as to whether or not it is tight. That is, does there exist a machine which achieves the lower bound? If not, how closely can it be approached? As will be shown in the next section, in general no machine can actually achieve P^*. However, as will be shown below, P^* can be approached arbitrarily closely.

As a first step consider the special case where X is a Bernoulli random variable. In this case X can take on only two values, denoted by H (heads) and T (tails). Under the null hypothesis H_0, $\Pr\{X = H\} = p$, $0 \leq p \leq 1$. Under the alternative hypothesis H_1, $\Pr\{X = H\} = q = 1-p$. Without loss of generality it can be assumed that $p > \frac{1}{2}$. Thus a large number of H's tend to favor H_0, while a large number of T's favors H_1.

If, for the moment, equal priors are assumed, then knowledge of the difference between the number of heads and the number of tails is sufficient for an optimal decision. If the difference is positive, decide H_0; if the difference is negative, decide H_1; and if the difference is zero, the decision is arbitrary.

An infinite-state automaton could be used to implement such a scheme. Let the states be numbered with all the integers (both positive and negative). If at a particular time the automaton is in state j, it is interpreted as meaning that up to that time there have been j more H's than T's. Thus at time zero (before any observations have occurred) the machine is in state 0. If the first observation, X_1, is an H, the machine moves to state 1; if X_1 is a T, the machine

moves to state -1. It is easy to see that if the machine is in state i and the new observation is an H, it moves to state $i + 1$, whereas if the new observation is a T it moves to state $i - 1$.

The decision rule decides H_0 in states 1,2,... and H_1 in states -1, -2,.... In state 0 it may make either decision.

This automaton is essentially a counter, capable of counting to plus and minus infinity. One way to make the memory of the algorithm finite is to allow the counter to count only up to a fixed upper bound and down to a fixed lower bound. If the counter reaches its upper bound and another H is observed, let it stay in the same state (i.e., its upper bound). Similarly, if a T is observed while the automaton is in its lowest numbered state, let it stay there. The counter saturates, and hence will be called a saturable counter. It is depicted in Fig. 1, where an arrow from one state to another indicates an allowed transition and the letter H or T over the arrow indicates for which observation the transition occurs. It should also be noted that the states have been numbered from 1 through m.

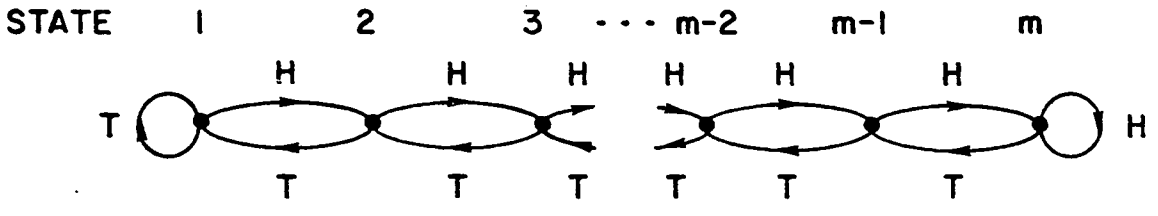

Fig. 1. SATURABLE COUNTER.

As was noted previously, it is possible to use the theory of Markov chains to solve for $\underline{\mu}^0$ and $\underline{\mu}^1$, the s.s. probability of occupation vectors under H_0 and H_1 respectively. Feller describes a method of solution using generating functions [14].

A simpler method will be used. This method makes an analogy to a diffusion process, as was done in the proof of Lemma 2. Since the Markov chain is assumed to be in the s.s., if it is broken into two subsets, then the flow from the first set to the second must equal the flow from the second to the first, so that the net flow is zero. If the first set is $\{1,2,\ldots,i\}$ and the second set is $\{i+1, i+2,\ldots,m\}$, then this condition results in the equations

$$\mu^0_{i+1} = (p/q)\mu^0_i \ , \ \mu^1_{i+1} = (q/p)\mu^1_i \ , \ i = 1,2,\ldots,m-1 \tag{32}$$

Thus the stationary distributions of the states are given by

$$\mu^0_i = a(p/q)^{i-1} \ , \ \mu^1_i = b(q/p)^{i-1} \ , \ i = 1,2,\ldots,m, \tag{32a}$$

where a is a normalizing constant such that the sum of the μ^0_i's equals one, and b is a similar normalizing constant for the μ^1_i's. These equations are depicted graphically in Fig. 2.

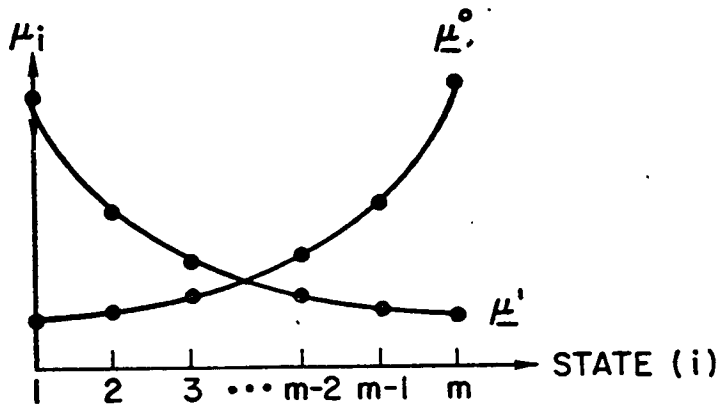

Fig. 2. DISTRIBUTION ON THE AUTOMATON'S STATES.

If $m = 2k$ and it is still assumed that the priors are equal, then the best partition of the states is to let S_0, the set of states in which the automaton decides H_0, be the set $\{k + 1, k + 2, \ldots, m\}$ and to let $S_1 = \{1,2,\ldots,k\}$. Since the priors are equal and the machine is symmetric in the two hypotheses (see Figs. 1 and 2), α and β are equal, so that the $P(e) = \alpha = \beta$. Using the fact that the sum of the μ_i^0's is one:

$$\alpha = \sum_{i \epsilon S_1} \mu_i^0 = \frac{\sum_{i=1}^{m/2} \mu_i^0}{\sum_{i=1}^{m} \mu_i^0} = \frac{a\left(\frac{1-(p/q)^{m/2}}{1-(p/q)}\right)}{a\left(\frac{1-(p/q)^{m}}{1-(p/q)}\right)} \tag{33}$$

and thus

$$P(e) = \frac{1}{(p/q)^{m/2} + 1} \tag{34}$$

Since it has been assumed (without loss of generality) that $p > q$, for large m the $P(e)$ decreases almost as $(q/p)^{m/2}$.

The saturable counter is the obvious restriction of the infinite counter when a finite memory constraint is imposed. However, with a slight modification the saturable counter can be improved, approximately doubling the effective memory. Looking at Fig. 2 it is seen that states 1 and m, the extreme states, are the ones in which the automaton is least likely to make an error. It would be nice if all probability were on these two states. That is, μ_1 (meaning both μ_1^0 and μ_1^1) and μ_m would be non-zero and all other μ_i would equal zero. This results in a non-ergodic automaton. However, it is possible, with an ergodic automaton, to approach this extreme as closely as desired by introducing

artificial randomization. The general form of the saturable counter is kept. However, if the automaton is in state 1 and the next observation is an H, let it pass to state 2 with small probability δ, and remain in state 1 with probability $1-\delta$ (instead of passing to state 2 with conditional probability one as it did in the original design). If the observation is a T which provides additional evidence that H_1 is correct, let the automaton remain in state 1 as it did before. It is seen that the effect of this randomization is to "trap" the automaton in state 1. Thus it will be referred to as a δ-trap.

Similarly, it is desired to trap the automaton in state m. So if the automaton is in state m and the new observation is a T, it transits to state m-1 with small probability $k\delta$. The reason for adding the factor of k is to allow the automaton to be unsymmetric if the priors are not equal, since at this point the assumption of equal priors will be dropped. The addition of this factor allows the machine to match its structure to the statistics of the problem. Thus k will be referred to as a matching section.

If the resultant Markov chain is solved, it is found that (32) still holds for $i = 2,3,\ldots,m-2$. since the chain is unchanged for states 2 through m-1. However, the effect of the δ traps in the end states is to increase μ_1 relative to μ_2 by a factor of $1/\delta$ and to increase μ_m by a factor of $1/k\delta$ relative to μ_{m-1}. Thus the following table exhibits $\underline{\mu}^0$ and $\underline{\mu}^1$:

State i

state	1	2	3	...	m-1	m
μ_i^0	a'/δ	$a'(p/q)$	$a'(p/q)^2$	...	$a'(p/q)^{m-2}$	$(a'/k\delta)(p/q)^{m-1}$
μ_i^1	b'/δ	$b'(q/p)$	$b'(q/p)^2$	...	$b'(q/p)^{m-2}$	$(b'/k\delta)(q/p)^{m-1}$

Again a' and b' are normalizing constants. Now if k is fixed and δ approaches zero, μ_1 and μ_m are much larger than all other μ_i. Thus the probability of being in state 1 or m approaches 1 as δ approaches 0. Of course δ can never reach zero or the chain becomes nonergodic, and then P* is not even approachable. In fact, when δ equals zero, Appendix II shows that P(e) is as large as when δ equals 1, the deterministic case considered previously. This example illustrates that $\underline{\mu}$ is not a continuous function of $[p_{ij}]$. Consequently the associated P(e) is not a continuous function of $[p_{ij}]$. Thus the P(e) of nonergodic machines is not obtainable as the limit of the P(e) or ergodic machines. This necessitates a separate argument in Appendix II to dispose of the nonergodic case. The ergodic machines are shown to be superior.

However, if δ is allowed to approach, but not reach, zero, P(e) will approach $\pi_0\mu_1^0 + \pi_1\mu_m^1$, since all other μ_i approach zero. This <u>limit</u> of P(e) will be denoted by P_{lim}. As P_{lim} is approached, it is seen that α approaches μ_1^0 and β approaches μ_m^1. Labelling these limits α_{lim} and β_{lim} and realizing that in the limit $\mu_1 + \mu_m = 1$ (since all other μ_i approaches zero), it is seen that

$$\alpha_{lim} = \frac{k}{k+(p/q)^{m-1}}, \qquad \beta_{lim} = \frac{1}{1+k(p/q)^{m-1}}, \tag{35}$$

so that

$$\frac{(1 - \alpha_{lim})(1 - \beta_{lim})}{\alpha_{lim}\,\beta_{lim}} = (p/q)^{2(m-1)} = \gamma^{m-1} \tag{36}$$

where in this special case $\bar{\ell} = (p/q)$ and $\underline{\ell} = (q/p)$. But (36) is just condition (24) which put a lower bound on α and β. Thus, for a given k, the α_{lim} and β_{lim} that result cannot be lowered. Since, as k varies from zero to infinity, α_{lim} varies from zero to one, the saturable counter with δ traps and matching section traces out the lower boundary of the allowable ROC's (in the limit as δ goes to zero). Thus in the limit as δ goes to zero, this automaton with k properly chosen approaches the lower bound on P(e) as derived in Theorem 2. The optimum value of k is obtained by differentiation and is given by

$$k^* = \begin{cases} \dfrac{\sqrt{\pi_1\,\gamma^{m-1}/\pi_0} - 1}{\sqrt{\gamma^{m-1}} - \sqrt{\pi_1/\pi_0}}, & \text{if } \gamma^{m-1} > \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_0}\right\} \\ 0, & \text{if } \gamma^{m-1} \le \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_0}\right\} \text{ and } \pi_0 > \pi_1 \\ \infty, & \text{if } \gamma^{m-1} \le \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_0}\right\} \text{ and } \pi_0 < \pi_1 \end{cases} \tag{37}$$

As has been noted, this automaton cannot achieve P^*, but rather can only approach P^* in the limit. That is, for any $\epsilon > 0$ there exists a $\delta > 0$ such that the saturable counter with δ traps of this value and with matching section $k = k^*$ has its $P(e) \le P^* + \epsilon$. Thus this class of machines is termed ϵ-optimal.

It has been shown that the saturable counter with δ traps and matching section can be made ϵ-optimal for the special case where X

is a Bernoulli random variable with parameter either p (under H_0) or $q = 1-p$ (under H_1). This result is easily extended to the case where X is still a Bernoulli random variable, but with $p = p_0$ (under H_0) and $p = p_1$ (under H_1). Without loss of generality it can be assumed that $p_0 > p_1$ so that more H's will result under H_0 than under H_1. However, it is possible for both p_0 and p_1 to be greater than $\frac{1}{2}$ so that at first the saturable counter would seem to be of little use since, under either hypothesis, the machine would be in the right half (states $(m/2) + 1$ through m) more often. A similar problem exists if p_0 and p_1 are both less than $\frac{1}{2}$. However, if the δ traps and matching sections are added, it is possible to overcome this apparent shortcoming. Again, in the limit as δ goes to zero, α approaches $\alpha_{lim} = \mu_1^0$, and β approaches $\beta_{lim} = \mu_m^1$. Solution of the Markov chain first under H_0 and then under H_1 yields

$$\alpha_{lim} = \frac{k}{k+(p_0/q_0)^{m-1}} \qquad \beta_{lim} = \frac{1}{1+k(q_1/p_1)^{m-1}} \tag{38}$$

Consequently

$$\frac{(1-\alpha_{lim})(1-\beta_{lim})}{(\alpha_{lim})(\beta_{lim})} = \left(\frac{p_0 q_1}{q_0 p_1}\right)^{m-1} = \gamma^{m-1} \tag{39}$$

since $\gamma = (\bar{\ell}/\underline{\ell})$, where, in this problem, $\bar{\ell} = (p_0/p_1)$ and $\underline{\ell} = (q_0/q_1)$.

As is seen by (38) as k is varied from zero to infinity α_{lim} and β_{lim} run from zero to one, and from one to zero, respectively. Here too the machines trace out the lower boundary of the ROC (again,

in the limit), so that for the proper value of k, this is an ϵ-optimal class of machines. The optimum value of k is given by:

$$k^* = \begin{cases} \dfrac{\pi_1 \gamma_0^{m-1} - \pi_0 \gamma_1^{m-1}}{(\pi_0 - \pi_1) + (\gamma_0^{m-1} - \gamma_1^{m-1}) \sqrt{\dfrac{\pi_0 \pi_1}{(\gamma_0 \gamma_1)^{m-1}}}}, & \text{if } \gamma^{m-1} > \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_0}\right\} \\ 0, & \text{if } \gamma^{m-1} \leq \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_0}\right\} \text{ and } \pi_0 > \pi_1 \\ \infty, & \text{if } \gamma^{m-1} \leq \max\left\{\dfrac{\pi_0}{\pi_1}, \dfrac{\pi_1}{\pi_1}\right\} \text{ and } \pi_0 < \pi_1 \end{cases} \tag{40}$$

where $\gamma_0 = (p_0/q_0)$ and $\gamma_1 = (p_1/q_1)$. The resultant P^* is given by:

$$P_{lim} = \frac{2\sqrt{\pi_0 \pi_1 \gamma^{m-1}} - 1}{\gamma^{m-1} - 1} = P^* \tag{41}$$

where $\gamma = \left(\dfrac{p_0\, q_1}{p_1\, q_0}\right)$.

Now it is possible by a simple extension to demonstrate a class of ϵ-optimal machines for the original, more general, problem. As is obvious from the method used to solve the Markov chains for $\underline{\mu}^0$ and $\underline{\mu}^1$, it is only the ratio of the probability of moving from state i to state $i+1$ to the probability of moving from state $i+1$ to i that determines $\underline{\mu}$. Thus if there are three possible outcomes for the experiment, say heads, tails and sides (abbreviated S), and if under H_0 the probabilities of these occurrences are p_0, q_0 and r_0 respectively $(p_0 + q_0 + r_0 = 1)$, and under H_1 they are p_1, q_1 and r_1 (these also sum to one) then if the automaton moves one state to the right on H,

one state to the left on T and remains in the same state on S, then P(e) is still given by (41).

Now returning to the general problem consider the following sets: $\mathcal{H} = \{x : \ell(x) = \bar{\ell}\}$, $\mathcal{T} = \{x : \ell(x) = \underline{\ell}\}$ and $\mathcal{S} = \{x : x \notin \mathcal{H} \text{ and } x \notin \mathcal{T}\}$. For the moment assume $\Pr\{\mathcal{H}\} > 0$ and $\Pr\{\mathcal{T}\} > 0$ (under either hypothesis). Later this assumption will be dropped. Define a new "saturable counter" as follows: whenever an $x \in \mathcal{H}$ is observed, the automaton moves up one state (unless it is in the highest state, in which case it stays there); whenever an $x \in \mathcal{T}$ is observed, the automaton moves down one state (unless it is in the lowest state); and if an $x \in \mathcal{S}$ is observed, the automation stays in its current state. Adding δ traps and a matching section this automation behaves exactly as if it were testing a coin that can show H, T or S, where p_0, q_0 and r_0 are $\mathcal{P}_0(\mathcal{H})$, $\mathcal{P}_0(\mathcal{T})$ and $\mathcal{P}_0(\mathcal{S})$. The quantities p_1, q_1, and r_1 are similarly defined. Thus the P(e) for the new saturable counter (in the general problem) is given by (41) where the parameters are defined above. Since $p_0/p_1 = \bar{\ell}$ and $q_0/q_1 = \underline{\ell}$, (41) becomes:

$$P_{lim} = \frac{2\sqrt{\pi_0\pi_1\,\gamma^{m-1}} - 1}{\gamma^{m-1} - 1}, \text{ where } \gamma = (\bar{\ell}/\underline{\ell}) \tag{42}$$

But this is just P^* for the general problem. Using the above substitutions, k^* is still given by (40).

Now drop the assumption that $\Pr\{\mathcal{H}\} > 0$ and $\Pr\{\mathcal{T}\} > 0$. Using the definitions of $\bar{\ell}$ and $\underline{\ell}$, it is seen that if $\bar{\ell} < \infty$, then for any $\epsilon > 0$ it is always possible to find sets $\mathcal{H}_\epsilon$ and $\mathcal{T}_\epsilon$ with nonzero probability measure such that for all $x \in \mathcal{H}_\epsilon$, $\ell(x) \geq \bar{\ell} - \epsilon$, and for

all $x \in \mathcal{T}_\epsilon$, $\ell(x) \leq \underline{\ell} + \epsilon$. If $\overline{\ell} = \infty$, then require that $1/\ell(x) \leq \epsilon$ for $x \in \mathcal{H}_\epsilon$. Thus by letting ϵ and δ approach zero, P(e) approaches P_{lim}. Hence it is seen that the saturable counter with only slight modifications, is an ϵ-optimal machine for the original problem. Equation (40) still gives k^*, if $p_0 = \mathcal{P}_0(\mathcal{H}_\epsilon)$, $q_0 = \mathcal{P}_0(\mathcal{T}_\epsilon)$, etc. The form of this machine is depicted in Fig. 3 where an arrow indicates an allowed transition and the event resulting in this transition is indicated over the arrow (for clarity the events which result in self loops have been deleted).

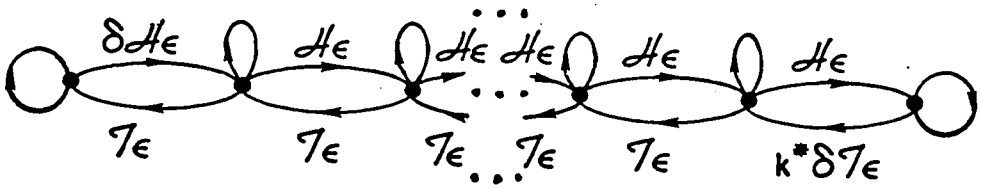

Fig. 3. THE CANONICAL FORM OF THE ϵ-OPTIMAL MACHINE.

It is of interest to note that if X has a continuous probability distribution, it is possible to achieve the randomization which is necessary for the operation of the δ traps and matching section by defining sets $\mathcal{H}'_\epsilon$ and $\mathcal{T}'_\epsilon$ whose elements have l.r.'s within ϵ of $\overline{\ell}$ and $\underline{\ell}$ respectively. Further, let $\Pr\{\mathcal{H}'_\epsilon\} \cong \delta \Pr\{\mathcal{H}_\epsilon\}$ and $\Pr\{\mathcal{T}'_\epsilon\} \cong k\delta \Pr\{\mathcal{T}_\epsilon\}$ (under both hypotheses). Then if the automaton leaves state 1 only when an $x \in \mathcal{H}'_\epsilon$ is observed and leaves state m only when an $x \in \mathcal{T}'_\epsilon$ is observed, the desired behavior is achieved, and no artificial randomization is required.

Examples:

Example 1: Let X be a Bernoulli random variable with distribution

$$X = \begin{cases} H, & p \\ T, & 1-p \end{cases}.$$

Consider the two-hypothesis testing problem $H_0 : p = p_0$ vs. $H_1 : p = p_1$, under equal priors $\pi_0 = \pi_1 = 1/2$. Recall in the case $\pi_0 = \pi_1$ that the ϵ-achievable lower bound on the probability of error reduces to $P^* = 1/[1+\gamma^{(m-1)/2}]$.

a) Let $p_0 = .99...99$ and $p_1 = .99...90$ (with the same number of 9's in between). This problem appears difficult because of the large number of trials necessary to obtain a significant test of the small difference between p_0 and p_1. Since an m-state automaton may only "count to m," it seems that memory will be exhausted before the test reaches an interesting level of significance. However, in this problem $\bar{\ell} = p_0/p_1 \cong 1$, $\underline{\ell} = q_0/q_1 = .1$, and $\gamma = \bar{\ell}/\underline{\ell} = p_0q_1/p_1q_0 \cong 10$. Thus for an $m = 5$ state memory, $P^* = 1/101 \cong .01$.

b) Now let $p_0 = 3/4$, $p_1 = 1/4$. Here $\gamma = p_0q_1/p_1q_0 = 9$, and $P^* = 1/82$ (for a 5-state automaton). This probability of error is actually higher than that of the previous example in which $p_0 = .99...99$ and $p_1 = .99...90$.

c) Of peculiar interest is the case $p_0 = .501$, $p_1 = .499$. Here $\gamma \cong 1.008$, which yields $P^* \cong .496$ for a 5-state automaton--little better than using no memory at all. In fact, it requires approximately 500 states to obtain $P^* = .01$. Clearly the difference $|p_0-p_1|$ is a poor measure of the resolvability of H_0 vs. H_1 in the finite-memory case.

The difference between examples a) and c) is that in example a) there is an event (the observation T) which occurs much more frequently under one hypothesis (H_1) than under the other. By essentially disregarding the other events, the high information content of the extreme event is well utilized. However, the similarity of examples a) and b) is lacking if only a finite number of observations is available. That is, the rate of convergence to the steady state is much slower in a) than in b).

Example 2: Let X be a univariate normal random variable with mean $\mu = +1$ (under H_0) and $\mu = -1$ (under H_1) and fixed variance $\sigma^2 = 1$. Let $\pi_0 = \pi_1 = 1/2$. In this case the likelihood ratio is given by $\ell(x) = \exp(2x)$. Therefore $\bar{\ell} = \infty$, $\underline{\ell} = 0$, and $\gamma = \infty$ --resulting in $P^* = 0$ for any memory at all ($m \geq 2$). To achieve this, let $\mathcal{H}_\epsilon = \{x : x \geq T\}$ and $\mathcal{J}_\epsilon = \{x : x \leq -T\}$. Move to state 1 for $x \in \mathcal{J}_\epsilon$, to state 2 for $x \in \mathcal{H}_\epsilon$, and remain in the current state otherwise. Then the asymptotic probability of error P(e) tends to zero as $T \to \infty$.

Example 3: X has a Cauchy distribution with pdf $f(x) = 1/\pi(1+(x-\mu)^2)$. Test $H_0 : \mu = 1$ vs. $H_1 : \mu = -1$ with $\pi_0 = \pi_1 = 1/2$. This example is of interest because the Cauchy and the normal distributions look similar and have comparable convergence rates for the probabilities of error in the infinite-memory case. However, calculation shows that $\bar{\ell} = \underline{\ell}^{-1} \cong 5.8$ and $\gamma \cong 33.6$. Thus a 2-state memory yields $P^* \cong .15$ for the Cauchy distribution, in marked contrast to the $P^* = 0$ obtainable in the normal case.

Uniqueness of the ϵ-optimal Class:

It has been shown that P^* is a lower bound on $P(e)$ and that an ϵ-optimal class of machines exists. However, might there not exist a machine for which $P(e) = P^*$? The following theorem shows this to be impossible except in certain degenerate cases.

Theorem 3: With the following exceptions, there exists no machine with $P(e) = P^*$.

Exceptions: 1. If the machine has two states ($m = 2$), then P^* is achievable. (No δ traps are needed.)

2. If $\gamma^{m-1} \leq \max\left\{\frac{\pi_0}{\pi_1}, \frac{\pi_1}{\pi_0}\right\}$ then P^* is achievable by the machine which always decides the hypothesis with the larger prior probability.

3. If $\bar{\ell}$ is infinite or $\underline{\ell}$ is zero and if there is non-zero probability of observing an X with this value of l.r., then $P^* = 0$ and is achievable. This case is degenerate since the support of $\mathcal{P}_0$ and $\mathcal{P}_1$ are different, and there exists a set of observation values which yield zero error.

Proof of Theorem: Referring to the proof of Theorem 2 it is seen that there are two necessary conditions for a machine to achieve $P(e)$ equal to P^*. First the automaton must have the maximum allowable spread of γ^{m-1} and secondly the probability of being in a state which does not have the maximum or minimum s.l.r. must be zero. These two conditions are contradictory (except in the degenerate cases mentioned) since the results of the Appendix show a nonergodic machine cannot achieve a spread of γ^{m-1}.

Now that it is known that no optimal machine exists, the question of uniqueness of the ϵ-optimal class arises. That is, might there exist a class of machines, much different in structure from the saturable counter, which is also ϵ-optimal? The following theorems show to what extent the answer to this question is no.

Theorem 4: Except in the degenerate cases listed below, the probabilities of error of a sequence of m-state automata approaches P^* under the following necessary and sufficient conditions.

1. The spread of the automata must approach the maximum allowable spread γ^{m-1}.

2. All μ_i, except μ_1 and μ_m, must approach zero (it is assumed that the states are numbered in order of increasing s.l.r.).

3. μ_1^0 must approach $\dfrac{\sqrt{\frac{\pi_1}{\pi_0}\gamma^{m-1}} - 1}{\gamma^{m-1} - 1}$.

Degenerate Exceptions:

1. $\bar{\ell} = \infty$ or $\underline{\ell} = 0$.

or

2. $\gamma^{m-1} \leq \max\{\pi_0/\pi_1, \ \pi_1/\pi_0\}$.

Proof: The sufficiency of the conditions will be proved first. Since all the μ_i's, except μ_1 and μ_m, approach zero, $\mu_1 + \mu_m$ approaches one. Thus in the limit

$$\mu_m^0 = 1 - \mu_1^0 = 1 - \frac{\sqrt{\frac{\pi_1}{\pi_0}\gamma^{m-1}} - 1}{\gamma^{m-1} - 1} \tag{43}$$

or

$$\mu_m^0 = \frac{\gamma^{m-1} - \sqrt{\frac{\pi_1}{\pi_0}}\,\gamma^{m-1}}{\gamma^{m-1} - 1} \quad . \tag{44}$$

Let C be the s.l.r of state 1. Then the s.l.r. of state m approaches $C\,\gamma^{m-1}$ so that in the limit

$$\mu_1^1 = \frac{1}{C}\,\mu_1^0 = \frac{1}{C}\left\{\frac{\sqrt{\frac{\pi_1}{\pi_0}}\,\gamma^{m-1} - 1}{\gamma^{m-1} - 1}\right\} \tag{45}$$

and

$$\mu_m^1 = \frac{1}{C\,\gamma^{m-1}}\,\mu_m^0 = \frac{1}{C\,\gamma^{m-1}}\left\{\frac{\gamma^{m-1} - \sqrt{\frac{\pi_1}{\pi_0}}\,\gamma^{m-1}}{\gamma^{m-1} - 1}\right\} . \tag{46}$$

But in the limit

$$\mu_1^1 + \mu_m^1 = 1 \, . \tag{47}$$

Therefore

$$C = \frac{\sqrt{\frac{\pi_1}{\pi_0}}\,\gamma^{m-1}}{\gamma^{m-1}} \quad . \tag{48}$$

Thus in the limit

$$P(e) = \pi_0 \mu_1^0 + \pi_1 \mu_m^1$$

$$= \pi_0 \left\{ \frac{\sqrt{\frac{\pi_1}{\pi_0} \gamma^{m-1}} - 1}{\gamma^{m-1} - 1} \right\} \tag{49}$$

$$+ \pi_1 \sqrt{\frac{\pi_0}{\pi_1 \gamma^{m-1}}} \left\{ \frac{\gamma^{m-1} - \sqrt{\frac{\pi_1}{\pi_0} \gamma^{m-1}}}{\gamma^{m-1} - 1} \right\}$$

$$= \frac{2\sqrt{\pi_0 \pi_1 \gamma^{m-1}} - 1}{\gamma^{m-1} - 1} = P^*$$

proving the sufficiency of the conditions.

To see that the conditions are necessary, refer to the proof of Theorem 2. For $P(e)$ to approach P^*, equations (21) and (22) must approach equality. But these are just conditions 1 and 2. Now if condition 2 is necessary, α necessarily approaches μ_1^0. But only one point on the lower boundary of the ROC yields $P(e) = P^*$ and at that point $\alpha = \alpha^*$. Thus μ_1^0 must approach α^* as given by (30), which is just condition 3. Q.E.D.

Conditions 2 and 3 merely state that a saturable counter must have its value of δ tend to zero and its value of k tend to k^*. Whether or not other types of machines can approach P^* really depends on condition 1. Thus the following two theorems, which give necessary and sufficient conditions for the spread of a sequence of automata to approach γ^{m-1}, demonstrate the essential uniqueness of the class of ϵ-optimal machines.

Theorem 5: If a sequence of m-state automata has its spread approach γ^{m-1} as a limit, then the following conditions must hold: (It is assumed that the states are arranged in order of increasing s.l.r.)

1. If the states of the automaton are partitioned into two sets $C = \{1,2,\ldots,k\}$ and $C' = \{k+1, k+2,\ldots,m\}$, then the fraction of the flow from C' to C which is *not from* state $k+1$ approaches zero. Similarly, the fraction of the flow from C to C' which is not from state k approaches zero.

2. The ratio of $\sum_{j=k+1}^{m} p^0_{k,j}$, the probability of moving from state k to a higher numbered state under H_0, to $\sum_{j=k+1}^{m} p^1_{k,j}$, the same probability under H_1, approaches $\bar{\ell}$. Similarly $\sum_{j=1}^{k-1} p^0_{k,j} \Big/ \sum_{j=1}^{k-1} p^1_{k,j}$ approaches $\underline{\ell}$.

The first part of condition 1 will be proved by contradiction. Let the automaton be partitioned into C and C' as in condition 1, but assume that the fraction of the flow from C' to C which is from $\{k+2, k+3,\ldots,m\}$ does not approach zero, but rather is always greater than some positive constant C_1 for a given subsequence of machines. Further assume that the spread of the sequence of automata approaches γ^{m-1} as a limit. But for the spread to approach γ^{m-1} it is necessary for the s.l.r. of any state i to approach $C_2 \gamma^{i-1}$, where C_2 is the minimum s.l.r. So for any $\epsilon > 0$, in the tail of the sequence

$$\mu^0_i/\mu^1_i > C_2(\gamma^{i-1} - \epsilon) \tag{50}$$

But equating flows between C and C'

$$\sum_{i\in C}\sum_{j\in C'} \mu_i^0 p_{ij}^0 = \sum_{i\in C'}\sum_{j\in C} \mu_i^0 p_{ij}^0 \equiv A \tag{51}$$

and

$$\sum_{i\in C}\sum_{j\in C'} \mu_i^1 p_{ij}^1 = \sum_{i\in C'}\sum_{j\in C} \mu_i^1 p_{ij}^1 \equiv B \tag{52}$$

are obtained. But by Lemmas 1 and 2

$$\underline{\ell} \leq p_{ij}^0/p_{ij}^1 \leq \bar{\ell} \tag{10}$$

and

$$\mu_i^0/\mu_i^1 \leq C_2\, \gamma^{i-1} \tag{53}$$

so that

$$A = \sum_{i\in C}\sum_{j\in C'} \mu_i^0 p_{ij}^0 \leq C_2\, \gamma^{k-1}\, \bar{\ell} \sum_{i\in C}\sum_{j\in C'} \mu_i^1 p_{ij}^1 \tag{54}$$

or

$$A \leq C_2\, \gamma^{k-1}\, \bar{\ell}\, B. \tag{55}$$

Now using (10) and (50)

$$A = \sum_{i\in C'}\sum_{j\in C} \mu_i^0 p_{ij}^0 > C_2\, \underline{\ell} \sum_{i\in C'}\sum_{j\in C} (\gamma^{(i-1)}-\epsilon)\, \mu_i^1 p_{ij}^1 \tag{56}$$

is obtained. Combining (55) and (56) yields

$$\gamma^k B > \sum_{i\in C'}\sum_{j\in C} (\gamma^{i-1}-\epsilon)\, \mu_i^1 p_{ij}^1 \tag{57}$$

since $\gamma = \bar{\ell}/\underline{\ell}$. But using the second expression for B

$$\sum_{i=k+1}^{m} \sum_{j=1}^{k} \mu_i^1 \, p_{ij}^1 > \sum_{i=k+1}^{m} \sum_{j=1}^{k} (\gamma^{i-(k+1)} - \epsilon') \mu_i^1 \, p_{ij}^1 \tag{58}$$

results, where $\epsilon' = \gamma^{-k}\epsilon$. Equivalently

$$\sum_{i=k+1}^{m} \sum_{j=1}^{m} (1-\gamma^{i-(k+1)} + \epsilon') \, \mu_i^1 \, p_{ij}^1 > 0 \tag{59}$$

or

$$\epsilon' \, \mu_{k+1}^1 \sum_{j=1}^{k} p_{k+1,j}^1 + \sum_{i=k+2}^{m} \sum_{j=1}^{k} (1-\gamma^{i-(k+1)} + \epsilon') \, \mu_i^1 \, p_{ij}^1 > 0 \tag{60}$$

or

$$\epsilon' \{\text{flow from state } k+1 \text{ to } C | H_1\} + (1-\gamma+\epsilon') \sum_{i=k+2}^{m} \sum_{j=1}^{k} \mu_i^1 \, p_{ij}^1 > 0 \tag{61}$$

But $\sum_{i=k+2}^{m} \sum_{j=1}^{k} \mu_i^1 \, p_{ij}^1$ is the flow from the states in C', except for $k+1$, to C. By assumption this is greater than C_1 times the flow from state $k+1$ to C. Further, it is always possible to make $(1-\gamma+\epsilon') \leq -C_3 < 0$ by going out far enough in the sequence. Thus

$$\epsilon' \{\text{flow from } k+1 \text{ to } C | H_1\} + (-C_3) \, C_1 \{\text{flow from } k+1 \text{ to } C | H_1\} > 0 \tag{62}$$

or

$$\epsilon' - C_1 C_3 > 0 \tag{63}$$

Since $C_1 > 0$, $C_3 > 0$ are fixed constants, far enough out in the sequence $\epsilon' < C_1 C_3$, a contradiction.

Thus the first part of condition 1 has been proved. A similar argument proves the second part.

Condition 2 will be proved by again partitioning the states into $C = \{1,2,\ldots,k\}$ and $C' = \{k+1, k+2, \ldots,m\}$ and equating flows under H_0 and H_1,

$$\sum_{i\epsilon C} \sum_{j\epsilon C'} \mu_i^0 p_{ij}^0 = \sum_{i\epsilon C'} \sum_{j\epsilon C} \mu_i^0 p_{ij}^0 \equiv A \tag{51}$$

and

$$\sum_{i\epsilon C} \sum_{j\epsilon C'} \mu_i^1 p_{ij}^1 = \sum_{i\epsilon C'} \sum_{j\epsilon C} \mu_i^1 p_{ij}^1 \equiv B \tag{52}$$

result. Now substitute $\{1,2,\ldots,k\}$ and $\{k+1, k+2,\ldots,m\}$ for C and C'. If ϵ_1 is the fraction of flow from C to C' which is not from state k, then the total flow is $\frac{1}{1-\epsilon_1}$ times the flow from state k, so

$$A = \mu_k^0 \sum_{j=k+1}^{m} p_{k,j}^0 + \sum_{i=1}^{k-1} \sum_{j=k+1}^{m} \mu_i^0 p_{ij}^0 = \left(\frac{1}{1-\epsilon_1}\right) \mu_k^0 \sum_{j=k+1}^{m} p_{k,j}^0 \tag{64}$$

is obtained. If C_2 is the minimum s.l.r, then as before (64) can lead to

$$A \leq \left(\frac{1}{1-\epsilon_1}\right) C_2 \gamma^{k-1} \left(\frac{\sum_{j=k+1}^{m} p_{k,j}^0}{\sum_{j=k+1}^{m} p_{k,j}^1} \right) \left(\mu_k^1 \sum_{j=k+1}^{m} p_{k,j}^1 \right) \tag{65}$$

Now

$$B = \sum_{i=1}^{k} \sum_{j=k+1}^{m} \mu_i^1 p_{ij}^1 \geq \mu_k^1 \sum_{j=k+1}^{m} p_{k,j}^1 \tag{66}$$

so

$$A \leq \left(\frac{1}{1-\epsilon_1}\right) C_2 \gamma^{k-1} \left(\frac{\sum_{j=k+1}^{m} p^0_{k,j}}{\sum_{j=k+1}^{m} p^1_{k,j}} \right) B. \tag{67}$$

Furthermore, since P(e) approaches P*, the s.l.r. of state i must approach γ^{i-1} so

$$A = \sum_{i=k+1}^{m} \sum_{j=1}^{k} \mu_i^0 p_{ij}^0 \geq C_2 (\gamma^k - \epsilon_2) \underline{\ell} \sum_{i=k+1}^{m} \sum_{j=1}^{k} \mu_i^1 p_{ij}^1 \tag{68}$$

where ϵ_2 can be made as small as desired by going out far enough in the sequence. But (68) is equivalent to

$$A \geq C_2(\gamma^k - \epsilon_2) \underline{\ell} B. \tag{69}$$

Therefore using (67) and (69)

$$\frac{\sum_{j=k+1}^{m} p^0_{k,j}}{\sum_{j=k+1}^{m} p^1_{k,j}} \geq (1-\epsilon_1)(\gamma - \epsilon_2') \underline{\ell} \tag{70}$$

results, $\epsilon_2' = \gamma^{-(k-1)} \epsilon_2$. But by condition 1, ϵ_1 approaches zero, and it has already been noted that ϵ_2 (and hence ϵ_2') approaches zero so the right side of (70) approaches $\gamma\underline{\ell}$. Since $\gamma = \bar{\ell}/\underline{\ell}$, the right side of (70) approaches $\bar{\ell}$, proving the first part of condition 2. A similar argument proves the second part.

Theorem 6: The conditions of Theorem 5 are not only necessary, but also sufficient for the spread of the automata to approach γ^{m-1}.

Proof: Partition the states of the automaton into $C = \{1,2,\ldots,k\}$ and $C' = \{k+1, k+2,\ldots,m\}$ and equate flows to obtain

$$\sum_{i\in C}\sum_{j\in C'} \mu_i^0 p_{ij}^0 = \sum_{i\in C'}\sum_{j\in C} \mu_i^0 p_{ij}^0 \equiv A \tag{51}$$

$$\sum_{i\in C}\sum_{j\in C'} \mu_i^1 p_{ij}^1 = \sum_{i\in C'}\sum_{j\in C} \mu_i^1 p_{ij}^1 \equiv B. \tag{52}$$

Now if ϵ_1 is the fraction of flow from C' to C which is not from state $k+1$ (under H_0), then

$$A = \frac{1}{1-\epsilon_1} \mu_{k+1}^0 \sum_{j=1}^{k} p_{k+1,j}^0 . \tag{71}$$

Now using the other expression for A

$$A = \mu_k^0 \sum_{j=k+1}^{m} p_{k,j}^0 + \sum_{i=1}^{k-1}\sum_{j=k+1}^{m} \mu_i^0 p_{ij}^0 \tag{72}$$

so that

$$A \geq \mu_k^0 \sum_{j=k+1}^{m} p_{k,j}^0 \tag{73}$$

and

$$\frac{1}{1-\epsilon_1} \mu_{k+1}^0 \sum_{j=1}^{k} p_{k+1,j}^0 \geq \mu_k^0 \sum_{j=k+1}^{m} p_{k,j}^0 . \tag{74}$$

Similarly using the expressions for B and letting ϵ_2 be the fraction of flow from C to C' which is not from state k

$$\frac{1}{1-\epsilon_2} \mu_k^1 \sum_{j=k+1}^{m} p_{k,j}^1 \geq \mu_{k+1}^1 \sum_{j=1}^{k} p_{k+1,j}^1 . \tag{75}$$

Combining (74) and (75) yields

$$\frac{\mu^0_{k+1}/\mu^1_{k+1}}{\mu^0_k/\mu^1_k} \geq (1-\epsilon_1)(1-\epsilon_2)\,\frac{\sum_{j=k+1}^{m} p^0_{k,j} \Big/ \sum_{j=k+1}^{m} p^1_{k,j}}{\sum_{j=1}^{k} p^0_{k+1,j} \Big/ \sum_{j=1}^{k} p^1_{k+1,j}} \tag{76}$$

But $\sum_{j=k+1}^{m} p^0_{k,j} \Big/ \sum_{j=k+1}^{m} p^1_{k,j}$ approaches $\bar{\ell}$ and $\sum_{j=1}^{k} p^0_{k+1,j} \Big/ \sum_{j=1}^{k} p^1_{k+1,j}$ approaches $\underline{\ell}$ by condition 2 so

$$\frac{\mu^0_{k+1}/\mu^1_{k+1}}{\mu^0_k/\mu^1_k} \geq (1-\epsilon_1)(1-\epsilon_2)\,\frac{\bar{\ell}-\epsilon_3}{\underline{\ell}-\epsilon_4} \tag{77}$$

where ϵ_3 and ϵ_4 approach zero. But ϵ_1 and ϵ_2 also approach zero (by condition 1) so that the ratio of the $k+1^{st}$ s.l.r. to the k^{th} s.l.r. approaches $\bar{\ell}/\underline{\ell} = \gamma$. Letting $k = 1,2,\ldots,m-1$ it is seen that the spread of the automaton must therefore approach γ^{m-1}, completing the proof.

These theorems show that the saturable counter is an essentially unique ϵ-optimal class.

Conclusions:

The form of the ϵ-optimal class provides insight into the optimal decision making process. Essentially the automaton waits for maximal or minimal l.r. events before changing state. Furthermore when it reaches an extreme state, the machine leaves with small probability. In the case of discrete distributions, this may require artificial randomization.

It is noticed that the automaton waits for extreme events before changing state. This shows that in many cases roundoff schemes are far from optimal, since they put emphasis on small changes. Thus taking a sufficient statistic and rounding it off to keep memory finite will in general not be close to an optimal strategy.

The automaton is able to wait for extreme events, even if they occur infrequently, since the number of trials is infinite. If the number of trials N is finite, the automaton will not be able to neglect events of moderate information. The problem of finding an optimal machine when N is finite is an interesting one, for, except in certain degenerate cases, as P(e) approaches P*, δ must approach zero. The resulting time to approach s.s. increases without bound. (See Appendix III.) Thus a machine which is close to optimal for an infinite number of samples is far from optimal in the small sample case. However, we conjecture that for finite N the optimal machine will still resemble the saturable counter in certain respects. We believe that high l.r. events will still cause upward transitions, and low l.r. events downwards transitions although the events need not be as extreme as before. Furthermore, we

believe that artificial randomization will still be needed; although the values of δ will not be near zero.

It would also be of interest to see whether human beings, in problems to which they have allotted finite memory (such as "like," "indifference" and "dislike") demonstrate an optimal randomized learning procedure similar to that suggested by this paper.

Appendix I: Definitions and Facts from the Theory of Markov Chains

As was noted in the body of the paper, for given H_o and H_1 the states occupied by an automaton form a Markov chain. Several definitions, similar to those of Markov chain theory, are needed.

Definition 1: A state j is said to be accessible from state i (abbreviated $i \rightarrow j$) if and only if (iff) it is possible to reach state j from state i in a finite number of steps. In more formal terms $i \rightarrow j$ iff $\sum_{n=1}^{\infty} \Pr\{\text{automaton is in state } j \text{ at time } n \mid \text{automaton is in state } i \text{ at time zero}\} > 0$.

Definition 2: Two states, i and j, communicate (written $i \leftrightarrow j$) iff $i \rightarrow j$ and $j \rightarrow i$. A set of states, $\mathcal{S}$ forms a communicating class iff $\forall_{i,j \in \mathcal{S}}\ i \leftrightarrow j$.

Definition 3: An automaton is irreducible iff the set of all its states forms a communicating class. This type of automaton is also known as ergodic (see Fact 7).

Definition 4: A state i is recurrent iff, given that the automaton starts in state i, it must eventually return there with probability one. A set of states $\mathcal{S}$ is said to be recurrent iff every state in $\mathcal{S}$ is recurrent. Note: Some authors use the word persistent instead of recurrent.

Definition 5: A state i is transient iff it is not recurrent. That is, given that the automaton starts in state i, there is non-zero probability that it will never return. A set of states $\mathcal{S}$ is transient iff every state in $\mathcal{S}$ is transient. Note: Some authors use the word non-recurrent instead of transient.

Definition 6: A set of states $\mathcal{A}$ is said to be closed iff once the automaton enters $\mathcal{A}$ it can never leave.

Some useful results from the theory of finite Markov chains will be stated without proof. The proof are either self-evident or may be found in introductory books on stochastic processes [15].

Fact 1: A recurrent communicating class is closed. Therefore once an automaton enters a recurrent communicating class it can never leave.

Fact 2: The set of states of an automaton can always be partitioned into an exhaustive collection of disjoint subsets $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_k$ and $\mathcal{T}$, such that each of the $\mathcal{R}_i$'s is a recurrent communicating class and $\mathcal{T}$ is transient.

Fact 3: An automaton always contains at least one recurrent communicating class.

Fact 4: An irreducible automaton contains only one communicating class. This class is necessarily the same as the set of all states in the automaton.

Fact 5: A recurrent state, if visited once, will, with probability one, be visited an infinite number of times. A transient state will, with probability one, be visited only a finite number of times. Therefore, with probability one, the automaton eventually reaches (and never leaves) a recurrent communicating class.

Fact 6: If an automaton is irreducible, regardless of the initial state, it will visit every state in finite time wpl.

Fact 7: An irreducible automaton is ergodic. That is, μ_i is equal, with probability one, to the limiting proportion of time the automaton spends in state i. More concisely, the time average is equal to the ensemble average.

Since an automaton must eventually reach, and never leave, one of its recurrent communicating classes, it is necessary to know the behavior of such classes. Since such a class, considered by itself, is an irreducible (or ergodic) automaton, it is seen that knowing the behavior of ergodic automata will greatly simplify the study of nonergodic automata.

Appendix II: Extensions of Theorem 1 to the Nonergodic Case

The following theorem will extend the proof of Theorem 1 to the nonergodic case. Note that since a transient state is visited only a finite number of times (with probability one) it follows that if i is transient, then μ_i must be zero. Thus if i is transient under both H_0 and H_1, the s.l.r. for state i is undefined.

Theorem A1: The spread of a nonergodic m state automaton is less than or equal to γ^{m-2}, provided γ is finite.

Proof: Case I: If the automaton has only one recurrent communicating class, $\mathcal{R}$ (with $m_1 < m$ states) and a set of transient states $\mathcal{T}$ (with $m_2 = m-m_1$ states), the automaton must eventually reach $\mathcal{R}$ independently of the initial state. Thus the machine effectively reduces to an m_1 state irreducible automaton. By lemma 2, the spread is less than or equal to γ^{m_1-1}, which is less than or equal to γ^{m-2}, since $m_1 \leq m-1$.

Case II: The automaton has no transient states, but several recurrent communicating classes $\mathcal{R}_1,\mathcal{R}_2,\ldots,\mathcal{R}_k$ having $m_1,m_2,\ldots,m_k$ states respectively. If the automaton starts in state $i \in \mathcal{R}_j$, it never leaves $\mathcal{R}_j$. Thus the machine is effectively irreducible with m_j states. Again by Lemma 2 the spread is less than or equal to $\gamma^{m_j-1} \leq \gamma^{m-2}$.

Case III: There are several recurrent communicating classes, $\mathcal{R}_1,\mathcal{R}_2,\ldots,\mathcal{R}_k$, having $m_1,m_2,\ldots,m_k$ states respectively. In addition there is a set of transient states $\mathcal{T}$ having m_t states. If the automaton starts in a recurrent state, Case II applies. If, however, the automaton starts in a state $i_0 \in \mathcal{T}$, several of the $\mathcal{R}_i$'s may be accessible.

If this is the case, there will be a set of probabilities $P(\mathcal{R}_1)$, $P(\mathcal{R}_2), \ldots, P(\mathcal{R}_k)$ denoting the respective probabilities of reaching $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_k$. (In this and future statements, the conditioning of starting in state i_0 is tacit.) Since the automaton must eventually reach one of the $\mathcal{R}_i$'s, these probabilities sum to one. As far as the automaton is concerned, all $\mathcal{R}_i$ such that $P(\mathcal{R}_i) = 0$ will never be reached and may be neglected. Similarly, if any state $i \in \mathcal{I}$ is not accessible from i_0, it may be neglected. Eliminating these states results in a smaller m' state automaton (or at least one which is no larger), so it can be assumed that all states in the automaton are accessible from i_0. (If such is not the case, eliminate the inaccessible states and apply the following proof to the smaller m' state machine. Since $\gamma > 1$, $\gamma^{m'-2} \leq \gamma^{m-2}$.)

Now for $i \in \mathcal{R}_1$, $\mu_i = P(\mathcal{R}_1)\, \mu_{1,i}$ where $\mu_{1,i}$ is the stationary probability of the automaton's being in state i given that it reaches $\mathcal{R}_1$. It is seen that the $\underline{\mu}_1$ vector is just the stationary probability of occupation vector for an irreducible m_1 state automaton. Similarly if $i \in \mathcal{R}_j$, $\mu_i = P(\mathcal{R}_j)\, \mu_{j,i}$ where $\underline{\mu}_j$ is the stationary probability of occupation vector for an irreducible m_j state automaton.

Now if the states with the maximum and the minimum s.l.r. (i and j respectively) occur in the same recurrent communicating class, say $\mathcal{R}_\ell$, then

$$\left(\frac{\mu_i^0}{\mu_i^1}\right) \Big/ \left(\frac{\mu_j^0}{\mu_j^1}\right) = \frac{\mu_i^0\, \mu_j^1}{\mu_i^1\, \mu_j^0}$$

$$= \frac{P^0(\mathcal{R}_\ell)\mu_{\ell,i}^0 \quad P^1(\mathcal{R}_\ell)\mu_{\ell,j}^1}{P^1(\mathcal{R}_\ell)\mu_{\ell,i}^1 \quad P^0(\mathcal{R}_\ell)\mu_{\ell,j}^0} \qquad \text{(A0)}$$

$$= \left(\frac{\mu^0_{\ell,i}}{\mu^1_{\ell,i}}\right) \Big/ \left(\frac{\mu^0_{\ell,j}}{\mu^1_{\ell,j}}\right) \leq \gamma^{m_\ell - 1} \leq \gamma^{m-2} \qquad \text{(A0)}$$

since $\underline{\mu}_\ell$ is the stationary probability of occupation vector for an m_ℓ state automaton, and $m_\ell < m$.

Next consider the possibility that state i (with the maximum s.l.r.) is in a different recurrent communicating class from state j (with the minimum s.l.r.). Without loss of generality let these classes be $\mathcal{R}_1$ and $\mathcal{R}_2$ respectively. Then

$$\frac{\mu^0_i}{\mu^1_i} \Big/ \frac{\mu^0_j}{\mu^1_j} = \left\{\frac{P^0(\mathcal{R}_1)\ P^1(\mathcal{R}_2)}{P^1(\mathcal{R}_1)\ P^0(\mathcal{R}_2)}\right\} \frac{\mu^0_{1,i}}{\mu^1_{1,i}} \frac{\mu^1_{2,j}}{\mu^0_{2,j}} \qquad \text{(A1)}$$

But $\forall_{i' \in \mathcal{R}_1}\ C \leq \dfrac{\mu^0_{1,i'}}{\mu^1_{1,i'}} \leq C\,\gamma^{m_1-1}$ by application of Lemma 2. Moreover $C \leq 1$. Otherwise all $\mu^0_{1,i'} > \mu^1_{1,i'}$ for all i' and it would be impossible for both vectors to sum to one. Similarly $C\gamma^{m_1-1} \geq 1$. Thus (A1) reduces to

$$\frac{\mu^0_i}{\mu^1_i} \Big/ \frac{\mu^0_j}{\mu^1_j} \leq \left\{\frac{P^0(\mathcal{R}_1)\ P^1(\mathcal{R}_2)}{P^1(\mathcal{R}_1)\ P^0(\mathcal{R}_2)}\right\} \gamma^{m_1-1}\ \gamma^{m_2-1} \qquad \text{(A2)}$$

It will be shown that the quantity in brackets is less than or equal to γ^{m_t}. Consequently (A2) becomes

$$\frac{\mu^0_i}{\mu^1_i} \Big/ \frac{\mu^0_j}{\mu^1_j} \leq \gamma^{m_1+m_2+m_t-2} \leq \gamma^{m-2} \qquad \text{(A3)}$$

since $(m_1 + m_2 + \ldots + m_k) + m_t = m$.

To show that the quantity in brackets is less than or equal to γ^{m_t}, consider the following experiment. The automaton is started in its initial state, $i_0 \in \mathcal{J}$, at $t = 0$. Eventually the automaton leaves $\mathcal{J}$ and enters one of the $\mathcal{R}_i$'s where it would then normally stay forever. However, whenever the automaton would normally exit to a recurrent state, force it instead to return to i_0, its initial state in $\mathcal{J}$. Define $N(t)$ to be the number of times up to time t that the automaton would normally have exited from $\mathcal{J}$ but instead was restarted in i_0. By the strong law of large numbers (SLLN)

$$\lim_{t \to \infty} [t/N(t)] = E\{\text{time to leave } \mathcal{J}\}, \text{ wpl} . \tag{A4}$$

It is seen that the effect of not allowing the automaton to leave $\mathcal{J}$ is the same as taking all paths leaving $\mathcal{J}$ and looping them back to i_0, making $\mathcal{J}$ into a <u>recurrent</u> communicating class with m_t states. It thus has a well-defined stationary probability of occupation vector $\underline{\sigma}$. Now by the SLLN

$$\lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} I\{\text{would have exited from } \mathcal{J} \text{ to } \mathcal{R}_1 \text{ at time } t = i\}$$

$$= \sum_{j \in \mathcal{J}} \sigma_j a_{j,1} , \text{ wpl.} \tag{A5}$$

where $a_{j,1}$ is the probability of going from $j \in \mathcal{J}$ to any state in $\mathcal{R}_1$ for the unmodified automaton, and I is an indicator function. By the SLLN

$$\lim_{T \to \infty} \frac{N(T)}{T} \left\{ \frac{1}{N(T)} \sum_{i=1}^{N(T)} I \left\{ \begin{matrix} \text{exited to } \mathcal{R}_1 \\ \text{when } N(t) = i \end{matrix} \right\} \right\}$$

$$= \frac{1}{E\{\text{time to leave } \quad\}} P(\mathcal{R}_1) \text{ , wpl.} \qquad \text{(A6)}$$

Equating (A5) and (A6) yields

$$P(\mathcal{R}_1) = E\{\text{time to leave } \mathcal{J}\} \sum_{j \in \mathcal{J}} \sigma_j \, a_{j,1} \qquad \text{(A7)}$$

But the same argument applies to all the other $\mathcal{R}_i$'s. Defining $a_{j,i}$ to be the probability of going from state $j \in \mathcal{T}$ to any state in $\mathcal{R}_i$ (again for the unmodified automaton) (A7) becomes

$$P(\mathcal{R}_i) = E\{\text{time to leave } \mathcal{J}\} \sum_{j \in \mathcal{J}} \sigma_j \, a_{j,i} \qquad \text{(A8)}$$

Furthermore, since $\underline{\sigma}$ is the probability of occupation vector for an irreducible m_t state automaton, there exists a constant C such that

$$\forall_{j \in \mathcal{T}} \; C \leq \frac{\sigma_j^0}{\sigma_j^1} \leq C \gamma^{m_t - 1} \; . \qquad \text{(A9)}$$

Since $a_{j,i}$ is just the probability of an X event, it follows from Lemma 1 that

$$\underline{\ell} \leq \frac{a_{j,i}^0}{a_{j,i}^1} \leq \bar{\ell} \; . \qquad \text{(A10)}$$

Therefore,

$$\frac{P^0(\mathcal{R}_1)P^1(\mathcal{R}_2)}{P^1(\mathcal{R}_1)P^0(\mathcal{R}_2)} = \left\{\frac{E^0\{\text{time to leave } \mathcal{I}\}\, E^1\{\text{time to leave } \mathcal{I}\}}{E^1\{\text{time to leave } \mathcal{I}\}\, E^0\{\text{time to leave } \mathcal{I}\}}\right\}$$

$$\times \frac{\sum_{j\in\mathcal{I}} \sigma_j^0 a_{j,1}^0 \sum_{j\in\mathcal{I}} \sigma_j^1 a_{j,2}^1}{\sum_{j\in\mathcal{I}} \sigma_j^1 a_{j,1}^1 \sum_{j\in\mathcal{I}} \sigma_j^0 a_{j,2}^0} \tag{A11}$$

$$\leq \frac{\left\{\sum_{j\in\mathcal{I}} (C\gamma^{m_t-1}\sigma_j^1)(\bar{\ell} a_{j,1}^1)\right\} \sum_{j\in\mathcal{I}} \sigma_j^1 a_{j,2}^1}{\sum_{j\in\mathcal{I}} \sigma_j^1 a_{j,1}^1 \sum_{j\in\mathcal{I}} (C\sigma_j^1)(\underline{\ell} a_{j,2}^1)}$$

$$= \frac{C\gamma^{m_t-1}\bar{\ell}}{C\underline{\ell}} = \gamma^{m_t} - .$$

completing the proof.

Since the spread of a nonergodic automaton cannot exceed γ^{m-2} and the maximum spread determines the lower bound on P(e), it is seen that, except for degenerate cases, an m state nonergodic automaton is at least one state worse than an optimal ergodic automaton.

Appendix III: Rate of Convergence

Our objective has been to minimize the asymptotic probability of error $P(e)$. In a practical situation, with only a finite sample size, it is important to know the rate of convergence of the probability of error at time n, $P_n(e)$, to $P(e)$. Initially, the rate of convergence depends on all eigenvalues and eigenvectors of the state transition matrix P. However, the asymptotic rate of convergence depends only on that eigenvalue of P which, in magnitude, is closest to one. Letting r be the magnitude of this eigenvalue, the time constant T is given by $T = 1/[\ell n(1/r)]$. Therefore the rate of convergence, defined to be $1/T$, is equal to $\ell n(1/r)$. Unfortunately, at this time, we do not have an analytical expression for r. However, it is intuitively obvious that as δ tends to zero (in the saturable counter with δ-traps and matching section) T tends to infinity and the rate of convergence tends to zero.

Although we cannot demonstrate the exact dependence of T on δ, we can show that a related parameter T' (defined below) is proportional to $1/\delta$, thus indicating the expected behavior.

Definition: The escape time T' of a saturable counter with δ-traps and matching section is defined to be the expected time to pass from the end state (1 or m) in which the wrong decision is made to the other end state (in which the correct decision is made). Therefore

$$T' = \pi_0 T_0 + \pi_1 T_1 \tag{A12}$$

where

$$T_o = E[\text{number of steps to reach state m from state 1} \mid H_o] \tag{A13}$$

$$T_1 = E[\text{number of steps to reach state 1 from state m} \mid H_1] \tag{A14}$$

To see why this parameter is of importance consider the special case where X is a Bernoulli random variable with parameter either p_o or p_1, and the automaton is started in state $\frac{m+1}{2}$ (m is assumed odd), the middle state. After a (random) time N_1 the automaton reaches one of the end states. But as will be shown below, for small values of P(e), $P_{N_1}(e)$ is much larger than P(e). This is because there is relatively high probability of the automaton's reaching the "wrong" end state. If it does, after an additional time N_2 (also random) it reaches the "correct" end state. Due to the ϵ-optimal design of the machine, once the correct end state is reached the automaton stays there for a "long" time. Thus for $n \geq N_1 + N_2$, $P_n(e)$ is close to P(e). Also due to the design $E[N_2] \gg E[N_1]$. Therefore as δ tends to zero $E[N_2]$ becomes the main factor limiting the convergence of $P_n(e)$ to P(e). Thus, as previously asserted, $T' = E[N_2]$ is related to T.

At this point it would be well to show that for small values of P(e), $P_{N_1}(e) \gg P(e)$ as was previously asserted. Essentially the problem reduces to a random walk with absorbing barriers at 1 and m and the initial state in the middle. If there is a drift to the right under both hypotheses ($p_o > 1/2, p_1 > 1/2$), then with high probability (greater than one half) the automaton will reach state m first, and decide H_0. But a priori there is probability π_1 that H_1 is the true state of nature. Therefore $P_{N_1}(e) > 1/2\ \pi_1 \gg P(e)$ if P(e) is reasonably small. Similar remarks hold if there is a constant drift to the left ($p_0 < 1/2$, $p_1 < 1/2$).

If under H_0 there is a drift to the right ($p_0 > 1/2$) and under H_1 there is a drift to the left ($p_1 < 1/2$) then the situation resembles $p_1 = 1 - p_0$, the first case considered in deriving the saturable counter.

In that case there is a one-to-one correspondence between sequences that result in reaching state 1 first and sequence that result in reaching state m first. Merely interchanging H and T in a sequence that reaches state 1 first causes state m to be reached instead and vice versa. By symmetry $P_{N_1}(e)$ is the same under H_0 and H_1, so assume H_0. A sequence which reaches state 1 has $(m-1)/2$ more T's than H's and a sequence which reaches state m has $(m-1)/2$ more H's than T's. Therefore any sequence which reaches state m first is $(p_0/q_0)^{(m-1)/2}$ times more probable than the corresponding one which reaches state 1 first. But then the probability of reaching state m first is $(p_0/q_0)^{(m-1)/2}$ times the probability of reaching state 1 first, and since the two probabilities sum to one

$$\Pr\{\text{reaching state 1 before reaching state m}|H_0\} = \frac{1}{1+\left(\frac{p_0}{q_0}\right)^{(m-1)/2}} \tag{A15}$$

therefore

$$P_{N_1}(e) = \frac{1}{1+\left(\frac{p_0}{q_0}\right)^{(m-1)/2}} \tag{A16}$$

Furthermore for optimal k and small δ

$$P(e) \doteq P^* \leq \frac{1}{1+\left(\frac{p_0}{q_0}\right)^{m-1}}$$

since

$$\gamma = \left(\frac{p_0 q_1}{q_0 p_1}\right) = (p_0/q_0)^2$$

and P^* is largest when $\pi_0 = \pi_1 = 1/2$. Thus for reasonably small values of $P(e)$

$$P_{N_1}(e) \gtrsim \sqrt{P(e)} \ . \tag{A17}$$

If $P(e) = .01$, $P_{N_1}(e) \doteq .1$! thus it is seen that $P_{N_1}(e) \gg P(e)$ when $P(e)$ is reasonably small. It should also be noted that $P_{N_1}(e)$ is $P(e)$ when $\delta = 0$. Therefore (A16) and (34) show that for $p_o = 1 - p_1$, $P(e)$ is the same when $\delta = 0$ or $\delta = 1$, as was noted previously. [Remember that m must be odd in (A16), whereas in (34) m must be even.]

Returning to the problem of calculating $T' = E[N_2]$, first consider T_0.

$$\begin{aligned} T_0 &= E[\text{number of steps (time) to reach state m from state 1} \,|\, H_0] \\ &= E[\text{time to reach 2 from 1}] \\ &\quad + E[\text{time to reach m from 2}] \end{aligned} \tag{A18}$$

where the conditioning on H_0 is tacit. But, given that the automaton is in state 1 it transits to state 2 with probability δp_0 and stays in state 1 with probability $1 - \delta p_0$. Thus using properties of the geometric distribution

$$E[\text{time to reach 2 from 1}] = 1/(\delta p_0) \ , \tag{A19}$$

Considering the second term in (A18),

$$\begin{aligned} &E[\text{time to reach m from 2}] \\ &= E[\text{time to reach m from 2} \,|\, \text{reach m before 1}] \times \Pr\{\text{reach m before 1 from 2}\} \\ &\quad + E[\text{time to reach m from 2} \,|\, \text{reach 1 before m}] \times \Pr\{\text{reach 1 before m from 2}\} \end{aligned} \tag{A20}$$

Further

$$\begin{aligned} &E[\text{time to reach m from 2} \mid \text{reach 1 before m}] \\ &= E[\text{time to reach 1 from 2} \mid \text{reach 1 before m}] + E[\text{time to reach m from 1}] \\ &= E[\text{time to reach 1 from 2} \mid \text{reach 1 before m}] + T_0 \end{aligned} \tag{A21}$$

Therefore

$$T_0 = \frac{\left\{ 1/(\delta p_0) + \left[E[\text{time to reach m from 2} \mid \text{reach m before 1}] \times \Pr\{\text{reach m before 1 from 2}\} \right] + \left[E[\text{time to reach 1 from 2} \mid \text{reach 1 before m}] \times \Pr\{\text{reach 1 before m from 2}\} \right] \right\}}{1 - \Pr\{\text{reach 1 before m from 2}\}} . \tag{A22}$$

There are three terms in the numerator of (A22). The first tends to infinity as δ tends to zero, but the other two terms do not involve δ and so are constant and finite. Therefore as δ tends to zero T_0 tends to $1/[\delta p_0(1 - \Pr\{\text{Reach 1 before m from 2}\})]$. A similar argument shows T_1 to be inversely proportional to δ (as δ tends to zero), so that $T' = \pi_0 T_0 + \pi_1 T_1$ has the same δ dependence. Thus $T' \to \infty$ as $\delta \to 0$.

Now, remove the restriction that X be a Bernoulli random variable. If there is non-zero probability of observing events in $\mathcal{H} = \{x: \ell(x) = \overline{\ell}$ and $\mathcal{T} = \{x: \ell(x) = \underline{\ell}\}$ then the previous analysis still applies. If however, $\mathcal{H}$ and $\mathcal{T}$ have zero probability measure than the rate of the convergence of the automaton will depend not only on δ, but also on the sets $\mathcal{H}_\epsilon$ and $\mathcal{T}_\epsilon$. For a given ϵ, the tail behavior of the probability measure induced on the l.r. determines the maximum rate of convergence. If ϵ is fixed then there is a modified lower bound P^*_ϵ on P(e) and the same analysis may be applied, so that then too T' is proportional to $1/\delta$.

Appendix IV: The Case of $\gamma = \infty$

As noted after the proof of lemma 1, certain proofs require modifications if $\overline{\ell} = \infty$ or $\underline{\ell} = 0$. For example, in lemma 1, if $\overline{\ell} = \infty$ the second inequality becomes $p_{ij}/p_{ij} \leq \infty$ which is trivially true. The real problem occurs later in the paper, when we assert that if $\gamma = \infty$, $p^* = 0$ is the greatest lower bound on $P(e)$. That it is a lower bound trivial, so the problem is to show that it is achievable (or at least approachable).

To see this, consider the case where $\overline{\ell} = \infty$ and $\mathcal{K} = \{x: \ell(x) = \overline{\ell}\}$ has non-zero probability (with respect to $\pi_0 p_0 + \pi_1 p_1$). Since $p_o(\mathcal{K}) \leq 1$ it follows that $p_1(\mathcal{K}) = 0$. But then $p_0(\mathcal{K}) > 0$ by assumption. Thus $\mathcal{K}$ occurs with non-zero probability under H_0 but with probability zero under H_1. Consider the two state machine, which decides H_1 in state 1 and H_0 in state 2. Start the machine in state 1 and let it transit to state 2 only if $\mathcal{K}$ is observed. If it ever reaches state 2 it stays there and never leaves.

Under H_1 the machine never leaves state 1. Thus it always makes the correct decision. Under H_0 it transits to state 2 in a finite time (with probability one) and from then on makes the correct decision. Thus, in either case, the asymptotic $P(e)$ is zero. Note that this is a degenerate case since P^* is actually achievable. Furthermore the machine which achieves P^* does not have but one communicating class.

Also note that in this case (40) predicts that $k^* = 0$ (since $\gamma_1 = 0$ and γ_0 is non-zero), in agreement with the lack of transitions from state 2 to state 1.

If $\mathcal{H}$ has zero probability measure, then as before $\mathcal{H}_\epsilon = \{x: [\ell(x)]^{-1} < \epsilon\}$ can be used as a suitable approximation to $\mathcal{H}$. However now k^* is near, but not equal to zero. Thus there will be low probability transitions from state 2 to state 1.

If $\underline{\ell} = 0$ similar remarks hold (merely replacing H_0 with H_1, etc.).

REFERENCES

1. Denny, J. L., "On Continuous Sufficient Statistics," Ann. Math. Stat. 35, pp. 1229-1233, 1964.

2. Cover, T. M., "Hypothesis Testing with Finite Statistics," submitted to Ann. Math. Stat.

3. Hellman, M. E. and T. M. Cover, " Comments on Automata in Random Media," to be published.

4. Tsetlin, M. L., "On the Behavior of Finite Automata in Random Media," Avtomatika i Telemakhanika 22, 10, pp. 1345-1354, Oct 1961; English translation is available in Automation and Remote Control.

5. Krylov, V. Y., "On One Automaton that is Asymptotically Optimal in a Random Medium," Avtomatika i Telemekhanika 24, 9, pp. 1226-1228, Sep 1963; English translation is available in Automation and Remote Control.

6. Varshavskii, V. I. and I. P. Vorontsova, "On the Behavior of Stochastic Automata with a Variable Structure," Avtomatika i Telemekhanika 24, 3, pp. 353-360, Mar 1963; English translation is available in Automation and Remote Control.

7. Robbins, H., "A Sequential Decision Problem with a Finite Memory," Proc. Natl. Acad. Sci. 42, pp. 920-923, 1956.

8. Isbell, J. R., "On a Problem of Robbins," Ann. Math. Stat. 30, pp. 606-610, 1959.

9. Smith, C. V. and R. Pyke, "The Robbins-Isbell Two Armed Bandit Problem with Finite Memory," Ann. Math. Stat. 36, pp. 1375-1386, 1965.

10. Samuels, S. M., "Randomized Rules for the Two Armed Bandit with Finite Memory," submitted to Ann. Math. Stat.

11. Cover, T. M., "A Note on the Two Armed Bandit Problem with Finite Memory," Inform. and Control, to appear Oct 1968.

12. Spragins, J., "Learning Without a Teacher," IEEE Trans. Inform. Theory IT-12, pp. 223-229, Apr 1966.

13. Fralick, S. C., "Learning to Recognize Patterns Without a Teacher," IEEE Trans. Inform. Theory IT-13, pp. 57-65, Jan 1967.

14. Feller, W., An Introduction to Probability Theory and its Applications: Vol. I, John Wiley and Sons, Inc., New York, p. 318, 1957.

15. Parzen, E., Stochastic Processes, Holden-Day, San Francisco, 1962.