Reprinted from INFORMATION AND CONTROL
All Rights Reserved by Academic Press, New York and London

Vol. 30, No. 1, January 1976
*Printed in Belgium*

# Optimal Finite Memory Learning Algorithms for the Finite Sample Problem

THOMAS M. COVER*

*Departments of Electrical Engineering and Statistics, Stanford University, Stanford, California*

MICHAEL A. FREEDMAN†

*Department of Mathematics, Georgia Institute of Technology, Atlanta, Georgia*

AND

MARTIN E. HELLMAN‡

*Department of Electrical Engineering, Stanford University, Stanford, California*

This paper explores the structure and performance of optimal finite state machines used to test between two simple hypotheses. It is shown that time-invariant algorithms can use knowledge of the sample size to obtain lower error rates than in the infinite sample problem. The existence of an optimal rule is established and its structure is found for optimal time-varying algorithms. The structure of optimal time-invariant rules is partially established. The particular problem of testing between two Gaussian distributions differing only by a shift is then examined. It is shown that the minimal error rate achievable after $N$ samples goes to zero like $\exp[-(\ln N)^{1/2}]$.

## 1. INTRODUCTION

Let $X_1, X_2, \ldots, X_N$ be independent observations, each distributed according to a probability measure $\mathscr{P}$ over the sample space $\mathscr{X}$. Consider the problem of testing between the two hypotheses:

$$\begin{aligned} H_0&: \ \mathscr{P} = \mathscr{P}_0 \\ H_1&: \ \mathscr{P} = \mathscr{P}_1 \end{aligned} \tag{1}$$

* Supported by Air Force Contract AF 44-620-69-C-0101.
† Supported under National Science Foundation Grant GK5800.
‡ Supported under National Science Foundation Grants GK5800 and GK33250.

where $\mathscr{P}_0$ and $\mathscr{P}_1$ are known measures. The a priori probabilities on the hypotheses $\pi_0$ and $\pi_1 = 1 - \pi_0$ are also assumed known.

With respect to this problem, an $m$-state time-invariant decision rule (Hellman and Cover, 1970) $\mathscr{A}$ is specified by a triple $(f, d, T_0)$, where $f$ is the state transition function, $d$ is the decision function, and $T_0$ is the initial state of memory. The memory state space is denoted by $S = \{1, 2,..., m\}$. At times $n = 1, 2,..., N$, transitions are made from state $T_{n-1}$ to state $T_n = f(T_{n-1}, X_n) \in S$, and at time $N$, a decision $d_N = d(T_N) \in \{H_0, H_1\}$ is made. Thus, the operation of the decision rule may be summarized by

$$T_n = f(T_{n-1}, X_n) \in S \qquad n = 1, 2,..., N \tag{2a}$$

$$d_N = d(T_N) \in \{H_0, H_1\}. \tag{2b}$$

An error is said to occur if $d_N \neq H_t$, where $H_t$ denotes the true hypothesis. The objective is, for given $m$, $N$, $\mathscr{P}_0$, $\mathscr{P}_1$, $\pi_0$, $\pi_1$, to find the rule $(f, d, T_0)$ which minimizes the probability of error $P_N(e) = \Pr(d_N \neq H_t)$.

If $f$ is a single-valued mapping, then $\mathscr{A}$ is said to be a deterministic rule. If $f$ is a randomized mapping, then $\mathscr{A}$ is called a randomized or stochastic rule. Elementary decision theoretic considerations show that the error probability cannot be lowered by randomization in $d$ or $T_0$. However, the optimal time-invariant algorithm usually involves randomization in the transition function $f$ (see Hellman and Cover, 1971; Hellman, 1972).

Let

$$P^*(m, N) = \inf_{\mathscr{A}_m} P_N(e) \tag{3a}$$

denote the infimum of $P_N(e)$ over all $m$-state, randomized, time-invariant decision rules $\mathscr{A}_m$, and let

$$P_d^*(m, N) = \inf_{\det \mathscr{A}_m} P_N(e) \tag{3b}$$

denote the infimum over all deterministic, $m$-state, time-invariant, decision rules, $\det \mathscr{A}_m$. Since deterministic rules are special cases of randomized rules,

$$P^*(m, N) \leqslant P_d^*(m, N) \qquad \forall_{m,N}. \tag{4}$$

In the infinite sample problem, definitions (3a) and (3b) extend in a natural way to

$$P_\infty(e) = \lim_{n \to \infty} P_n(e) \tag{5a}$$

$$P^*(m, \infty) = \inf_{\mathscr{A}_m} P_\infty(e) \tag{5b}$$

$$P_d^*(m, \infty) = \inf_{\det \mathscr{A}_m} P_\infty(e). \tag{5c}$$

Finite memory rules governed by (2) are time-invariant. Time-varying finite memory decision rules are described by

$$T_n = f(T_{n-1}, X_n, n) \in S \tag{6a}$$

$$d_N = d(T_N, N) \in \{H_0, H_1\}. \tag{6b}$$

Since time-varying rules include time-invariant rules

$$P^*_{tv}(m, N) \leqslant P^*(m, N), \tag{7}$$

where $P^*_{tv}(m, N)$ is the infimum of $P_N(e)$ over all $m$-state time-varying algorithms. We do not need to consider separate infima over deterministic and randomized time-varying rules, since, as shown later in this paper, they yield the same value.

In this paper, we examine certain aspects of the finite sample problem for both time-varying and time-invariant algorithms. First, in Section 3 we discuss the definition of $P^*(m, N)$ and demonstrate an unexpected behavior. Then, in light of the fact that an optimal rule does not exist for the infinite sample problem, we prove in Section 4 that an optimal rule does exist for any finite sample problem. We also prove that when $\mathscr{P}_0$ and $\mathscr{P}_1$ represent continuous distributions, there is a deterministic optimal rule for the time-invariant problem. There always is a deterministic optimal rule for the time-varying problem.

Attention is then focused on time-varying algorithms. In Section 5, we show that the optimal time-varying rule is deterministic and of a likelihood ratio form (as therein defined). In Section 6, we examine two-state time-invariant rules, and under certain assumptions, we show that the optimal time-invariant rule is also likelihood ratio.

The problem of testing between two Gaussian distributions differing only by a shift is dealt with in Section 7. It is shown that for $N$ large, $P^*(2, N) \sim \exp[-2(2 \ln N)^{1/2}]$. Previous work (Hellman and Cover, 1970) has shown that $P^*(2, \infty) = 0$, in agreement with the limit of this expression. However, the extremely slow rate of approach (slower than algebraic) was not hinted at by the infinite sample theory. The derivation of this behavior allows us to make a conjecture concerning the asymptotic behavior of $P^*(m, N)$ for this problem.

## 2. History

There are two distinct formulations of the finite memory hypothesis testing problem. One formulation has a finite time memory (Robbins, 1956; Isbell, 1959; Samuels, 1968), the other a finite state memory (Cover, 1969; Hellman

and Cover, 1970). The latter formulation will be used throughout this paper. It is the counterpart to minimal sufficient statistics for finitary decision algorithms. Decision-making by automata (Tsetlin, 1961; Krylov, 1963; Chandrasekaran and Shen, 1968; Fu and Li, 1968) is a closely related problem.

We will first deal with time-invariant algorithms. The infinite sample problem was solved by Hellman and Cover (1970). They showed that

$$P^*(m, \infty) = \min \left\{ \frac{2(\pi_0\pi_1\gamma^{m-1})^{1/2} - 1}{\gamma^{m-1} - 1}, \pi_0, \pi_1 \right\} \tag{8}$$

is the greatest lower bound on the asymptotic error probability achievable by an $m$-state algorithm. The parameter $\gamma$ depends only on $\mathscr{P}_0$ and $\mathscr{P}_1$, and will be defined shortly. If $P^*(m, \infty)$ equals $\pi_0$ or $\pi_1$, a degenerate situation exists in which the same decision is made in all states. In Hellman and Cover (1970), it is also shown that, in general, there is no optimal rule, but only an $\epsilon$-optimal class rules. That is, for any $\epsilon > 0$, there is a rule in the class which has $P_\infty(e) \leqslant P^*(m, \infty) + \epsilon$, yet no rule can be found for which $P_\infty(e) = P^*(m, \infty)$.

The class of $\epsilon$-optimal rules is characterized as follows: Let $l_{\max}$ be the essential supremum and $l_{\min}$ be the essential infimum of the likelihood ratio $l(x) = d\mathscr{P}_0/d\mathscr{P}_1$. For most problems, these are just the maximum and minimum values of $l(x)$. The parameter $\gamma$ in (8) is equal to $l_{\max}/l_{\min}$. Assuming that there is nonzero probability of observations with likelihood ratios of $l_{\max}$ and $l_{\min}$, the $\epsilon$-optimal class of rules is given by

$$\begin{aligned}
f(i, x) &= i + 1, && 2 \leqslant i \leqslant m - 1 \text{ and } l(x) = l_{\max} \\
&= i - 1, && 2 \leqslant i \leqslant m - 1 \text{ and } l(x) = l_{\min} \\
&= 2, && \text{with probability } \delta \text{ if } i = 1 \\
& && \quad \text{and } l(x) = l_{\max} \\
&= m - 1, && \text{with probability } k\delta \text{ if } i = m \\
& && \quad \text{and } l(x) = l_{\min} \\
&= i, && \text{otherwise}
\end{aligned} \tag{9a}$$

and

$$\begin{aligned}
d(i) &= H_0 \qquad i > m/2 \\
&= H_1 \qquad i \leqslant m/2.
\end{aligned} \tag{9b}$$

This rule is a saturable counter, adding $+1$ (if possible) to the state of memory on maximum likelihood ratio observations; adding $-1$ (if possible) on minimum likelihood ratio observations; and retaining the old state of memory on all other observations. In addition, transitions from state 1 to

state 2, and from state $m$ to $m-1$, are made with small probabilities $\delta$ and $k\delta$, respectively. This improves performance over the case, $\delta = 1$, $k = 1$, for two reasons. First, decisions made in states 1 and $m$ are the least likely to be in error. Making $\delta$ less than 1 increases the fraction of decisions made in these states, thereby lowering the error probability. Second, if there are asymmetries in the problem (e.g., $\pi_0 \neq \pi_1$), choosing $k$ to offset these asymmetries lowers the error probability. If $k$ is set to its optimal value [Hellman and Cover, 1970, Eq. (57)], the probability of error of the machines described above tends to $P^*$, as $\delta > 0$ tends to zero.

It is counter-intuitive that the $\epsilon$-optimal class requires randomization. If $\mathscr{P}_0$ and $\mathscr{P}_1$ represent continuous distributions, randomization is not needed (Hellman and Cover, 1970) for the $\epsilon$-optimal class. However, for discrete distributions there can be arbitrarily large discrepancies between the performance of randomized and deterministic rules (Hellman and Cover, 1971). To be precise, for any $m < \infty$ and $\epsilon > 0$, there exists a problem $(\mathscr{P}_0, \mathscr{P}_1, \pi_0, \pi_1)$ for which $P_d^*(m, \infty) > \frac{1}{2} - \epsilon$ and $P^*(2, \infty) < \epsilon$. Thus, the amount of memory saved by randomized rules can be arbitrarily large.

On the other hand, Hellman (1972) has shown that in a certain sense deterministic rules are asymptotically optimal for large memories. This may seem somewhat surprising, but, of course, there is no contradiction, as seen by a more precise statement: For any problem $(\mathscr{P}_0, \mathscr{P}_1, \pi_0, \pi_1)$, there exists a $B < \infty$ such that for all $b$ the optimal deterministic rule with $B + b$ bits in memory has a lower error probability than the optimal randomized rule with $b$ bits in memory. Thus, for large memories the fraction of bits lost by using a deterministic rule tends to zero.

Horos and Hellman (1972) use a slightly different model and find the $\epsilon$-optimal class of rules to be deterministic. This model allows a confidence to be associated with each decision, errors being weighted according to the confidence with which they are made.

Flower and Hellman (1972) examined the finite sample problem for Bernoulli observations. They found that most properties of the infinite sample solution carried over. For optimal designs, transitions were made only between adjacent states and randomization was needed. However, in the finite sample problem, randomization was needed on all transitions toward the center state (i.e., on transitions from states of low uncertainty to states with higher uncertainty). Samaniego (1974) proves that this structure is optimal for $m = 3$ when attention is restricted to symmetric machines and problems.

Flower and Hellman found that a symmetric problem (e.g., testing whether the bias of a coin is $\frac{3}{4}$ or $\frac{1}{4}$ with equal prior probabilities) did not necessarily have a symmetric solution. Within the class of rules they studied, the optimal

rule was asymmetric for $m$ odd, but was symmetric for $m$ even. From Section 7 of this paper, it will be seen that the initial distribution they used is not optimal for $m$ even and that the optimal machine is therefore not symmetric, even when $N$ is even.

Lynn and Boorstyn (1972) examined the finite sample problem for observations with continuous symmetric distributions. They calculated the probability of error for algorithms of a particular form, which they call finite memory linear detectors. For this type of detector, a transition occurs from state $i$ to $i+1$ if $i \leqslant m-1$ and $X_n > D$, a transition occurs from state $i$ to $i-1$ if $i \geqslant 2$ and $X_n < -D$, and the transition is from state $i$ to itself in all other cases. The threshold $D$ is optimized over the nonnegative real line. The authors note that this form of machine is somewhat restrictive, but that its simplicity makes it attractive. It resembles the $\epsilon$-optimal solution to the infinite sample problem in all but two respects. First, the $\delta$-randomization on transitions from the end states is missing. Lynn and Boorstyn found that using a larger threshold in transitions out of states 1 and 3 (equivalent to randomized transitions) lowered $P(e)$ by approximately a factor of 2. Second, moves are made on very large observations, not on large likelihood ratio observations. For the Gaussian example treated by Lynn and Boorstyn (and also Section 7 of this paper), there is no difference between $x$ being large and $l(x)$ being large since the likelihood ratio is monotone in $x$. However, for the problem of testing between two Cauchy distributions, one centered at $-1$, the other at $+1$, the likelihood ratio $l(x) \to 1$ (no information) as $|x| \to \infty$. This problem is easily eliminated by regarding $Y = \log l(X)$ as the observation. A problem which is symmetric in $X$ also will be symmetric in $Y$.

It is reasonable to expect multistate transitions to occur in finite sample problems. Lynn and Boorstyn tried this modification on a three-state machine for Gaussian statistics. However, they found that allowing multistate transitions decreased $P(e)$ by less than 10%, and for eight or more observations the decrease was less than 3%. Reasons for this behavior will be developed in Section 8 of this paper.

Shubert and Anderson (1973) studied a form of generalized saturable counter and found performance to be close to optimal. The simplicity of this class of rules makes it attractive for implementation on binary data. Shubert (1974) also studied an interesting variant of the Bernoulli hypothesis testing problem in which the machine observes not only $\{X_n\}$, but also two reference sequences $\{Y_n\}$ and $\{Z_n\}$ with biases $p_1$ and $p_2$, respectively. He showed that if memory is increased by one bit, then a deterministic machine can perform better than the original optimal randomized machine.

Chandrasekaran and Lam (1975) studied an interesting class of deterministic rules for the symmetric problem and conjectured that the optimal deterministic rules lies within this class.

Samaniego (1973) worked on the problem of estimating the parameter of a Bernoulli distribution and, restricting attention to a certain form of machine, found minimax solutions using a variant of the mean square error loss criterion. If $p$ is the true value of the parameter and $\hat{p}$ is the estimate, his loss function is $(p - \hat{p})^2/p(1 - p)$. The machine is restricted to make transitions only between adjacent states and to move up on heads and down on tails.

Hellman (1974) examined the infinite sample Gaussian estimation problem and showed that the problem can be reduced to a quantization problem. This result also applies to a larger class of infinite sample estimation problems.

All algorithms discussed thus far have been time-invariant, and while time-varying algorithms are less attractive to implement, their theory is sometimes simpler and provides insight into the design of time-invariant algorithms. Mullis and Roberts (1968) worked on a sequential decision problem with time-varying finite memory. The cost for an observation and the cost for each type of error were variable. They found necessary conditions for an optimal design and used an iterative technique to find an approximation to the optimal rule.

Cover (1969), concerned with the infinite sample time-varying problem, was able to show that a four-state memory (two bits) was sufficient to ensure that the probability of error tends to zero. One bit was used to remember the current favorite hypothesis and one bit was used to keep track of the success or failure of test blocks, which became increasingly larger. Koplowitz (1974) has recently shown that Cover's rule can be reduced to a three-state form. He also shows that for any $m - 1$ hypothesis problem, the optimal $m$-state time-varying rule has zero asymptotic error probability. Further, Koplowitz proves that, in general, $m$ states are necessary for this behavior. Hirschler and Cover (1975) have shown that eight states are sufficient to determine the rationality or irrationality (excluding a null set of irrationals) of the parameter of a coin, given independent coin flips.

Wagner (1972) uses rules similar to Cover's (1969) to estimate the mean of a distribution. For Bernoulli observations, Wagner's scheme is very close to optimal, since its maximum absolute error is at most $1/m$, with $m$-states in memory. Using Koplowitz's ideas, it can probably be shown that $1/(2(m - 1))$ is a lower bound on maximum absolute error.

Muise and Boorstyn (1972) showed that for the finite sample problem, the optimal time-varying rule essentially stores a quantized version of the likelihood ratio, although the quantization is time-varying and not of any

simple form. These optimal detectors result in the fastest decay of error probability with increasing sample size. In Section 5 of this paper, we develop simpler proofs of some of these results. The results of Cover (1969) that four states allow the error probability to tend to zero cannot be (or at least to date have not been) inferred from Muise and Boorstyn's work.

Roberts and Tooley (1970) attacked the problem of estimating a parameter with a time-varying finite memory. They restrict their rules to be of a special form which, although not optimal in general, does make sense (and is probably optimal) for many problems of interest. This restriction is that a transition take place from state $i$ to state $j$ at time $k$ if $X_k \in (\theta_{j-1}(i, k), \theta_j(i, k))$. Thus, larger observations cause transitions to higher numbered states. In some problems (such as those involving the Cauchy distribution), very large observations yield very little information, and such a rule seems to be distinctly suboptimal. However, for Gaussian statistics, large observations are very informative and the optimal unrestricted rule is probably of the given form. Koplowitz and Roberts (1973) unified and extended this work. In particular, their demonstration of necessary and sufficient conditions for the optimal state transition function should prove valuable.

Tooley and Roberts (1973) extended these ideas to estimating random processes with finite memory. Baxa and Nolte (1972) used rules similar to those of Roberts and Tooley (1970), except for the detection, as opposed to the estimation problem. Their rules, while suboptimal, show favorable performance for even three bits of memory.

Mullis and Roberts (1974) have formulated a more general finite memory model which includes control theory problems. While general results are probably not possible, Mullis and Roberts were able to establish limited results within this framework.

Work on the two-armed bandit problem with a finite memory constraint, also known as the problem of automata in random media, is of historical significance to this problem area and is discussed by Fu and Li (1968) and Cover and Hellman (1970).

## 3. Definition of $P^*(m, N)$

The definition used for error probability in the infinite sample time-invariant problem (Hellman and Cover, 1970) is

$$P_\infty(e) = E \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e_n , \tag{10}$$

where $e_n$ equals 1 or 0, accordingly, as the $n$th decision is or is not in error. The limit in (10) exists with probability one, but its value depends on which hypothesis is true and in which recurrent communicating class (Hellman and Cover, 1970) the machine becomes trapped. The expectation need only be taken over these possibilities.

An alternative definition

$$\tilde{P}_\infty(e) = \lim_{N\to\infty} Ee_N \tag{11}$$

is possible so long as attention is restricted to machines for which the indicated limit exists. When it does, the values of $\tilde{P}_\infty(e)$ and $P_\infty(e)$ will agree. Thus, the more complex definition (10) is only needed to ensure that periodic machines (where certain states can be occupied only at multiples of some period) are included in the class over which we minimize $P_\infty(e)$. From the $\epsilon$-optimal class derived by Hellman and Cover (1970) and described in the previous section, we see that some self transitions (i.e., from a state to itself) must occur with probability strictly between zero and one. Since a periodic Markov chain cannot have self-transitions (except those that occur with probability zero or one), periodic machines are excluded from being $\epsilon$-optimal. Thus restricting attention to machines for which the limit (11) exists would not change $P^*(m, \infty)$.

In the finite sample time-invariant problem, we might try to decide between using the definitions

$$P_N(e) \equiv Ee_N \tag{12}$$

and

$$P_N'(e) \equiv E\frac{1}{N}\sum_{n=1}^{N} e_n \tag{13}$$

as the definition of $P_N(e)$. However, it is doubtful that we would think of using

$$P_N''(e) \equiv \sup_{n\geqslant N} Ee_n\,. \tag{14}$$

In all future sections, we shall use the simplest definition (12), since it is the most tractable. However, doing so yields anomalous behavior, as indicated by the following.

FALSE THEOREM. *Using the definition* (12), $P^*(m, N) \geqslant P^*(m, \infty)$ *for all problems.*

COUNTEREXAMPLE (to positive statement of theorem). Let $X$ be a Bernoulli random variable with $P(X = 1 \mid H_0) = 10^{-10}$ and $P(X = 1 \mid H_1) = 10^{-20}$. Then $l_{\max} = 10^{-10}/10^{-20} = 10^{10}$, $l_{\min} \doteq 1$, $\gamma \doteq 10^{10}$, and for $\pi_0 = \pi_1 = \frac{1}{2}$, $P^*(2, \infty) \doteq 10^{-5}$. The two-state machine which achieves $P^*$, transits from state 1 to state 2 when $X = 1$ and from state 2 to state 1 with probability approximately $10^{-15}$ when $X = 0$. Note that $P^*$ is achievable (as opposed to $\epsilon$-achievable) because $m = 2$ (Hellman and Cover, 1970).

Now consider the machine which starts in state 1 and transits to state 2 when $X = 1$. Once in state 2, it stays there forever. Let $N = 10^{15}$. As defined in (12), the probability of error under $H_1$ is approximately $10^{-5}$, while under $H_0$ it is approximately $\exp(-10^5) \ll 10^{-5}$. Thus, $P_N(e) \doteq \frac{1}{2} \times 10^{-5} < P^*(2, \infty)$. If $N$ were $2 \times 10^{11}$, then this machine would have $P_N(e) \doteq 4 \times 10^{-9} \ll P^*(2, \infty)$.

An even more interesting anomaly is that using (12), $P^*(m, N)$ need not tend to $P^*(m, \infty)$ as $N \to \infty$. To see this, note that if the above machine transited from state 1 to state 2 with probability $\delta$ when $X = 1$, then its error probability at time $N/\delta$ would be approximately the same as for the original machine at time $N$. For example, if the sample size is $2 \times 10^{12}$, setting $\delta = 0.1$ would yield an error probability of approximately $4 \times 10^{-9}$. If $N = 2 \times 10^{41}$, setting $\delta = 10^{-30}$ would achieve $P_N(e) \doteq 4 \times 10^{-9}$, etc. Thus, we see that for all $N$ such that $2 \times 10^{11} \leqslant N < \infty$, $P^*(2, N) \leqslant 2 \times 10^{-9} \ll P^*(2, \infty) \doteq 10^{-5}$.

Note, however, that for any fixed $\delta$, as $N \to \infty$, $P_N(e) \to \frac{1}{2}$, so that no machine has a limiting error probability less than $P^*(2, \infty)$. If this were not so, the results of Hellman and Cover (1970) would be violated. It is necessary to know $N$ and then to match the machine to this value of $N$. If we observe fewer samples than anticipated, the resultant error probability may be higher than predicted, but that is to be expected. However, it is unexpectedly undesirable for the number of samples to be larger than anticipated. If our knowledge of $N$ is somewhat fuzzy, this is a poor model. Frequently, we only know that the number of samples will be large—at least as large as some integer $N$. In such a situation, (14) is a better definition of $P_N(e)$.

In contrast with the time-invariant behavior, time-varying rules have an error probability which is monotone in the sample size $N$. An $N + 1$ sample time-varying algorithm can neglect the $N + 1$st sample. Therefore, $P^*_{\text{TV}}(m, N)$ is nonincreasing in $N$, even using (12) as the definition of $P_N(e)$.

It should be noted that the above counterexample is a very asymmetric problem. We believe that for more symmetric problems, there is little if any difference between using (12) and (14) as the definition of $P_N(e)$. Our belief is supported by the following self-evident theorem.

THEOREM 1. *Let $\mathcal{A}$ minimize $P_N(e)$ as given in* (12) *with $N$ considered fixed. If $\mathcal{A}$'s error probability is nonincreasing in $n$ for $n \geqslant N$, then $\mathcal{A}$ is also optimal under definition* (14) *for this value of $N$.*

The optimal machines which we find for the Gaussian problem of Section 7 will have the above property and are thus optimal using either (12) or (14). For this reason, and also because of the intractabliity of (14), we use only the simpler definition (12) for the remainder of this paper. This section is intended only to post a warning sign, not to divert traffic.

## 4. EXISTENCE OF AN OPTIMAL MACHINE

As has been noted, an optimal $m$-state rule does not generally exist for the infinite sample problem. However, the reasons for this appear to be absent from the finite sample problem. The existence of optimal algorithms is established in the following theorem.

THEOREM 2. *For any problem $(\mathscr{P}_0, \mathscr{P}_1, \pi_0, \pi_1)$ and any $m, N < \infty$, there exists a time-varying $m$-state algorithm which achieves $P_N(e) = P_{tv}^*(m, N)$ and a time-invariant $m$-state algorithm which achieves $P_N(e) = P^*(m, N)$.*

*Proof.* We will prove the theorem for time-invariant rules, since the extension to time-varying rules is straightforward. The terminology in the remainder of this section is therefore that of the time-invariant problem, and any algorithms referred to are tacitly time-invariant.

The independence of the observations causes the memory states occupied by the machine to form a Markov chain under either hypothesis. The action of the state transition rule $f$ may be described equivalently by a family of $m$ by $m$ stochastic matrices $P(x)$, indexed by $x \in \mathscr{X}$. The entry in the $i$th row, $j$th column is

$$p_{ij}(x) \equiv \Pr(T_n = j \mid T_{n-1} = i, X_n = x). \tag{15}$$

Let $P^{(k)} = E(P(X) \mid H_k)$ be the state transition matrix of the Markov chain under $H_k$, $k = 0, 1$.

Similarly, the initial state of memory can be specified by an $m$-dimensional row vector $\mu(0)$ whose $i$th entry is the probability that $T_0 = i$. Then, if

$$\mu_i^k(n) = \Pr(T_n = i \mid H_k) \tag{16}$$

we see that

$$\mu^k(n) = \mu(0)[P^{(k)}]^n. \tag{17}$$

Let $S_k = \{i \in S: d(i) = H_k\}$, $k = 0, 1$, denote the decision regions specified by the decision-function $d$. Then,

$$P_N(e) = \pi_0 \sum_{i \in S_1} \mu_i^0(N) + \pi_1 \sum_{i \in S_0} \mu_i^1(N), \tag{18}$$

so that $P_N(e)$ depends on $f$ only through $(P^{(0)}, P^{(1)})$, that is, two rules whose state transition functions yield the same $(P^{(0)}, P^{(1)})$ pair will have the same error probability if their initial states and their decision regions are the same. Thus, we can minimize $P_N(e)$ over $(f, d, T_0)$ by fixing $T_0$ and $d$ and minimizing over all allowable $(P^{(0)}, P^{(1)})$ pairs and then minimizing over $T_0$ and $d$. Due to the fact that randomization over $T_0$ and $d$ is not necessary for optimality, we need only be concerned with $m$ values for $\mu(0)$ and $2^m$ decision functions $d$. Thus, if for each choice of $\mu(0)$ and $d$, the minimum of $P_N(e)$ is achievable (so it is a true minimum, not an infimum), then $P^*(m, N)$ is also achievable, since it is the minimum of the $m2^m$ minima thus found.

It can be seen that $P_N(e)$ is a continuous function (indeed an $N$th degree polynomial) of $P^{(0)}$ and $P^{(1)}$. If the region $\mathscr{R}$ of allowable $(P^{(0)}, P^{(1)})$ is closed and bounded (i.e., compact), then $P_N(e)$ takes on a minimum value in $\mathscr{R}$. But, $\mathscr{R} \subset [0, 1]^{2m^2}$, so that $\mathscr{R}$ is bounded. Thus, all that is needed is the following.

LEMMA 1. *$\mathscr{R}$ is closed.*

*Proof.* Since the regions inducing transitions from state $i$ can be chosen independently of the regions inducing transitions from state $j$, $\mathscr{R} = \mathscr{B}^m$ where $\mathscr{B} \subset [0, 1]^{2m}$ is the set of allowable values for the first (or any other) rows of $P^{(0)}$ and $P^{(1)}$. Thus, we must show $\mathscr{B}$ is closed. $\mathscr{B}$ (and hence $\mathscr{R}$) are convex because the mixture of two possibly randomized state transition functions is yet another. $\bar{\mathscr{B}}$, the closure of $\mathscr{B}$, is then both closed and convex. Let $\mathbf{p}^* = (\mathbf{p}_0^*, \mathbf{p}_1^*)$ be an extreme point of $\bar{\mathscr{B}}$. Then, there is a $2m$-vector $\mathbf{w} = (\mathbf{w}_0, \mathbf{w}_1)$ and a real number $L$ such that

$$\mathbf{w}^T\mathbf{p}^* = \sum_{i=1}^{m} (w_{0i} p_{0i}^* + w_{1i} p_{1i}^*) = L \tag{19}$$

and

$$\mathbf{w}^T\mathbf{p} > L \qquad \text{for all } \mathbf{p} \in \bar{\mathscr{B}},\ \mathbf{p} \neq \mathbf{p}^*. \tag{20}$$

But, the problem of minimizing $\mathbf{w}^T\mathbf{p}$ over $\mathbf{p} \in \mathscr{B}$ is equivalent to minimizing it over $\{p_i(x)\}_{i=1}^m$, where $p_{0i} = E(p_i(X) \mid H_0)$, $p_{1i} = E(p_i(X) \mid H_1)$, and $\sum_{i=1}^m p_i(x) = 1$. This is, in turn, equivalent to minimizing the cost in an

$m$-action two-hypothesis problem with a single observation $x$, where $p_i(x)$ is the probability of taking action $i$ when $x$ is observed and $w_{ki}$ is the cost of taking action $i$ when $H_k$ is the true hypothesis. It is well known that the minimum (Bayes') risk is achievable, completing the proof of Lemma 1 and hence of Theorem 2.

A further dividend is provided in the following lemma.

LEMMA 2. *The extreme points of $\mathscr{B}$ are generated by $p_{ij}(x)$ of the form*

$$\begin{aligned} p_{ij}(x) &= 1, \qquad a_{ij} \leqslant l(x) < b_{ij} \\ &= 0, \qquad \textit{otherwise.} \end{aligned} \tag{21}$$

*Proof.* This allows us to show that randomization is not needed for continuous distributions $(\mathscr{P}_0, \mathscr{P}_1)$. The proof is a bit involved, but rests on the fact that if, in an interval $A \leqslant l(x) < B$, action $i$ is taken with probability $\lambda$ and action $j$ is taken with probability $1 - \lambda$, then it is possible to find a partitioning of $[A, B)$ into $I_1 = [A, a_1) \cup [a_2, B)$ and $I_2 = [a_1, a_2)$ such that $\int_{I_1} d\mathscr{P}_k(x) = \lambda \int_{[A,B)} d\mathscr{P}_k(x)$, $k = 0, 1$. The remaining details of this proof are omitted.

If it can be shown that $P_N(e)$ achieves its minimum at an extreme point of the convex set of transition matrices $(P^0, P^1)$, then we would know that rules of the form (21) are optimal. We suspect that such is the case.

To make the preceding ideas more concrete, consider the problem for an $m = 2$ state machine. For simplicity, denote $p_{12}^k$ by $\alpha_k$ and $p_{21}^k$ by $\beta_k$. Thus, $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ specify the $(P^{(0)}, P^{(1)})$ matrices.

It is seen that the region of allowable $(\alpha_0, \alpha_1)$ coincides with the region of allowable $(\beta_0, \beta_1)$ and that the region of allowable $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ is just the Cartesian square of this region. Note that $0 \leqslant \alpha_0 \leqslant 1$, and for a fixed $\alpha_0$, maximizing or minimizing $\alpha_1$ is achieved by invoking the Neyman–Pearson lemma. Thus, for a given $\alpha_0$, it is seen that $\alpha_1$ is minimized by

$$\begin{aligned} p_{12}(x) &= 0 \qquad l(x) < L \\ &= c \qquad l(x) = L \\ &= 1 \qquad l(x) > L, \end{aligned} \tag{22}$$

where $L$ and $c$ are chosen to satisfy

$$\Pr\{l(x) < L\} + c \Pr\{l(x) = L\} = \alpha_0 . \tag{23}$$

Similarly, for a given $\alpha_0$, it is seen that $\alpha_1$ is maximized by taking

$$\begin{aligned} p_{12}(x) &= 0 \quad && l(x) > L \\ &= c \quad && l(x) = L \\ &= 1 \quad && l(x) < L, \end{aligned} \tag{24}$$

where now $L$ and $c$ are chosen to satisfy

$$\Pr\{l(x) > L\} + c \Pr\{l(x) = L\} = \alpha_0 . \tag{25}$$

By analogy to the operating characteristic of the Neyman–Pearson theory, we see that the curves of maximum achievable and minimum achievable $\alpha_1$ are, respectively, concave and convex functions of $\alpha_0$, and are continuous except perhaps at $\alpha_0 = 0$ and $\alpha_0 = 1$. This general behavior is shown in Fig. 1. As indicated in Fig. 1, the region is symmetrical in the sense that $(\alpha_0, \alpha_1)$ is in the allowable region if, and only if, $(1 - \alpha_0, 1 - \alpha_1)$ is in the allowable region. This symmetry is due to the fact that every measurable set has a measurable complement.

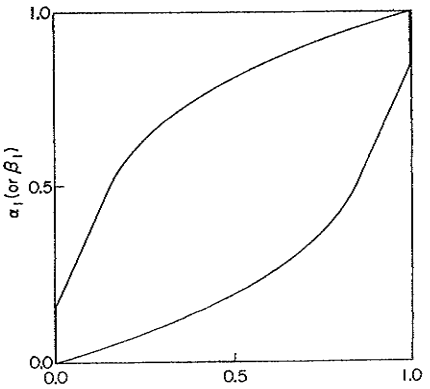


FIG. 1. Region of allowable $(\alpha_0, \beta_0)$ and $(\alpha_1\beta_1)$.

By using randomization, we can find a rule which achieves any point between the upper and lower bounds. Thus, the region of allowable $(\alpha_0, \alpha_1)$ is closed and bounded, and therefore, compact.

Suppose there are two state transition rules specified by $p_{ij}(x)$ and $q_{ij}(x)$ and such that $p_{ij}(x) + q_{ij}(x) \leqslant 1$ for all $x \in \mathcal{X}$. Then, there exists a "sum" state transition rule specified by $r_{ij}(x) = p_{ij}(x) + q_{ij}(x)$. It is seen that the values of $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ for the sum state transition rule are the sums of the

two original $(\alpha_0, \alpha_1, \beta_0, \beta_1)$'s. Thus, if a rule which corresponds to a fixed point on the lower boundary is added to all possible rules on the upper boundary, a curve is traced out, as shown in Fig. 2. Then, by varying the

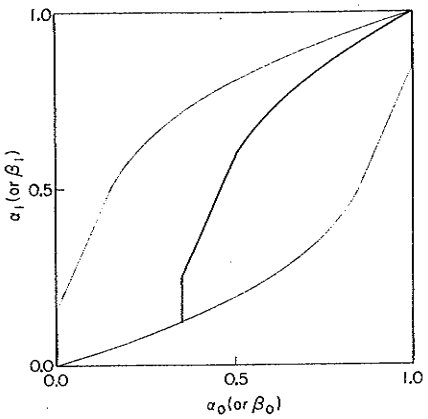


FIG. 2. Addition of two transition functions.

"fixed" point on the lower boundary, the entire region is swept out. Thus, any $(\alpha_0, \alpha_1)$ point in the allowable region can be obtained by using a rule of the form:

$$\begin{aligned} p_{ij}(x) &= 1 \qquad l(x) > L_1 \\ &= c_1 \qquad l(x) = L_1 \\ &= 0 \qquad L_2 < (x) < L_1 \\ &= c_2 \qquad l(x) = L_2 \\ &= 1 \qquad l(x) < L_2 \end{aligned} \tag{26}$$

for appropriate $L_1$, $L_2$, $c_1$, and $c_2$. By analogy to the Neyman–Pearson theory, it is seen that artificial randomization is not necessary for finite sample problems if the measures $\mathscr{P}_0$ and $\mathscr{P}_1$ represent continuous distributions. This is the same as in the infinite sample theory (Hellman and Cover, 1970).

The results of this section on the existence of optimal solutions are of use in several ways. First, it is interesting that finite sample problems possess optimal solutions, whereas infinite sample problems do not. Second, these results simplify proofs in the finite sample theory. For example, we now can deal with a single optimal rule, as opposed to an $\epsilon$-optimal class of rules. To show that a rule is optimal, we can show that all other rules perform more poorly. In the infinite sample problem, it was necessary to deal with an

infinite sequence (an $\epsilon$-optimal class) of $m$-state rules, each of which came closer to achieving $P^*(m, \infty)$. To show that a given class of rules was $\epsilon$-optimal, it was necessary to find a lower bound on error probability and show that for each $\epsilon > 0$ there was a rule in the class whose error probability was within $\epsilon$ of $P^*$. Third, in Section 6, we will use the concepts and intuition developed for the two-state problem to examine the structure of optimal two-state time-invariant rules.

## 5. Structure of the Optimal Time-Varying Rule

In this section, we prove that the optimal time-varying rule is deterministic and likelihood ratio in form. See Muise (1970) for an alternative formulation and proof.

Definition. A finite-memory time-varying rule is said to be likelihood ratio if

$$f(i, x, n) = j \quad \text{and} \quad l(x') > l(x) \quad \text{implies } f(i, x', n) \geqslant j \tag{27}$$

and

$$f(i, x, n) = j \quad \text{and} \quad k > i \quad \text{implies } f(k, x, n) \geqslant j. \tag{28}$$

*Remark.* Under this definition, transitions depend only on the likelihood ratio, and are upward for $l$ large and downward for $l$ small.

Theorem 3. *The optimal time-varying rule is deterministic.*

*Remark.* This is decidedly not true for infinite-sample time-invariant rules.

*Proof.* We first show that a deterministic initial state and decision rule are optimal. Suppose an algorithm starts in state $i$ with probability $p_i$. Then,

$$P_N(e) = \sum_{i=1}^{m} p_i P_N(e \mid T_0 = i). \tag{29}$$

But the error probability is minimized by making $T_0 = i_0$ for that value $i_0$ which minimizes $P_N(e \mid T_0 = i)$. Similarly, if $P(H_0 \mid T_N = i) > P(H_1 \mid T_N = i)$, then $d(i) = H_0$ is optimal. Only if $P_N(e \mid T_0 = i)$ is minimized

for two different values of $i$, or if $P(H_0 \mid T_N = i) = P(H_1 \mid T_N = i)$, can randomization be used without loss of optimality. However, even in these cases, there is always an optimal rule which is deterministic.

The real problem is to show that the optimal state transition function $f$ is deterministic.

Suppose we are given a randomized time-varying rule with state transition function $f_\omega(T_{n-1}, X_n, n)$, where we have explicitly included the dependence on the external random variable $\omega$, and where $\omega$ is required to be independent of the true hypothesis and the data. Note that this formulation includes as a special case the class of rules

$$T_n = f(T_{n-1}, X_n, n, Z_n) \tag{30}$$

where $\{Z_n\}_{n=1}^N$ are independent identically distributed random variables.

For a particular realization $\omega$, the original randomized rule is the deterministic time-varying rule

$$f_\omega(T_{n-1}, X_n, n).$$

Given $\omega$, the expected loss of $f_\omega$ is denoted $P_\omega$, while the expected loss of the original randomized rule is $EP_\omega = P$. But there must be at least one value $\omega = \omega_0$, such that $P_{\omega_0} \leqslant P$. Q.E.D.

The above theorem is perhaps best interpreted as saying that randomization, when independent of the data, can be regarded as a form of time-variation, and therefore, cannot increase the performance of time-varying rules.

We now outline a proof that the optimal time-varying rule is likelihood ratio. We do this by showing that any rule which is not likelihood ratio can be transformed into a likelihood ratio rule whose performance is better than that of the original rule.

DEFINITION. For $k = 0, 1$, $j \in S$, and $0 \leqslant n \leqslant N$, let $e_k(j, n)$ be the expected probability of error at time $N$, given $H_k$ and $T_n = j$.

It is seen that $e_k(j, n)$ depends only on $f(i, x, n')$ for $n' > n$ and on the decision rule $d(i)$. In particular, it does not depend on $f(i, x, n')$ for $n' \leqslant n$. Also, for $k = 0, 1, j \in S$, and $0 \leqslant n \leqslant N$ define

$$\pi_k(j, n) = \Pr(H_k \mid T_n = j). \tag{31}$$

It is easily seen that

$$\pi_0(j, n) = \pi_0 \mu_j{}^0(n) / (\pi_0 \mu_j{}^0(n) + \pi_1 \mu_j{}^1(n)), \tag{32}$$

so that $\pi_k(j, n)$ depends only on $f(i, x, n')$ for $n' \leqslant n$.

At time $n$, we will number the states in order of increasing $e_1(j, n)$, (i.e., $e_1(j, n) \leqslant e_1(j+1, n)$). If this produces a time-varying renumbering of the states, it is possible to incorporate the time-varying numbering into the state-transition function.

LEMMA 3. *Under the above assumption on state numbering,*

$$e_0(j, n) \geqslant e_0(j+1, n), \qquad \begin{matrix} 1 \leqslant n \leqslant N-1 \\ 1 \leqslant j \leqslant m-1. \end{matrix} \tag{33}$$

*Proof.* We must show that it is impossible for both $e_0(j, n) < e_0(j+1, n)$ and $e_1(j, n) \leqslant e_1(j+1, n)$. But this is clear, since then state $j$ would have a lower risk than state $j+1$ under either hypothesis. Modifying the algorithm by changing $f(j+1, x, n+1)$ to the same value as $f(j, x, n+1)$ for all $x$ would then cause $e_0(j+1, n)$ and $e_1(j+1, n)$ to take on the lower values $e_0(j, n)$ and $e_1(j, n)$. This can only lower $P_N(e)$. Q.E.D.

Now suppose we fix the state transition rule, except for transitions made at time $n$ and optimize the state transition rule only for these transitions. If the machine is in state $i$ at time $n$, $X_{n+1} = x$ is observed, and the rule causes a transition to state $j$, then the expected loss is

$$\begin{aligned} &\Pr(H_0 \mid T_n = i, X_{n+1} = x) e_0(j, n+1) + \Pr(H_1 \mid T_n = i, X_{n+1} = x) e_1(j, n+1) \\ &\quad = \frac{p_0(x)\, \pi_0(i, n)}{p_0(x)\, \pi_0(i, n) + p_1(x)\, \pi_1(i, n)}\, e_0(j, n+1) \\ &\qquad + \frac{p_1(x)\, \pi_1(i, n)}{p_0(x)\, \pi_0(i, n) + p_1(x)\, \pi_1(i, n)}\, e_1(j, n+1). \end{aligned} \tag{34}$$

Clearly, $T_{n+1} = f(i, x, n)$ should equal $j_0$, that state $j$ for which (34) is minimum. Equivalently, we must minimize the projection of the vector $(p_0(x)\, \pi_0(i, n), p_1(x)\, \pi_1(i, n))$ on the set $\{(e_0(j, n+1), e_1(j, n+1))\}_{j=1}^{m}$. This is depicted in Fig. 3. It is seen that increasing $l(x) = p_0(x)/p_1(x)$ causes the vector to lie closer to the horizontal, thereby causing $j_0$ to take on a higher value. Thus, (27) is satisfied by the optimal time-varying rule.

In a similar manner, it is seen that keeping $x$ constant, but changing $i$ so that $\pi_0(i, n)/\pi_1(i, n)$ is increased, also causes $j_0$ to increase. Therefore, to establish that the optimal rule satisfies (28), the second condition of the definition,

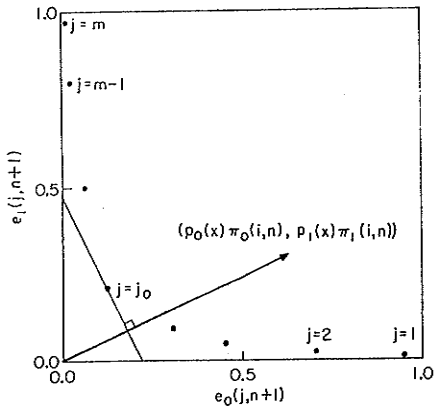


FIG. 3. Geometry of the time-varying minimization.

we must show that $e_1(i, n)$ (and hence $i$) is increasing in $\pi_0(i, n)/\pi_1(i, n)$. This can be inferred from the above picture, since

$$e_1(i, n, x) = e_1(f(i, n, x), n + 1), \tag{35}$$

where $e_1(i, n, x)$ is the expected risk under $H_1$ when $T_n = i$ and $X_n = x$.

We summarize the results of this section in the following theorem.

THEOREM 4. *There is an optimal time-varying rule with the following properties*:

(1) *The decision function* $d(i) \in \{H_0, H_1\}$ *is deterministic.*

(2) *The initial state of memory is nonrandom.*

(3) *The state transition rule is deterministic.*

(4) *The state transition rule is likelihood ratio, and therefore, is of the form*

$$f(i, n, x) = j, \qquad T_{i,j-1}^{(n)} \leqslant l(x) < T_{i,j}^{(n)} \tag{36}$$

*where, for* $1 \leqslant n \leqslant N$ *and* $1 \leqslant i, j \leqslant m$,

$$-\infty = T_{i,0}^{(n)} \leqslant T_{i,1}^{(n)} \leqslant \cdots \leqslant T_{i,m}^{(n)} = +\infty \tag{37}$$

*and*

$$T_{1,j}^{(n)} \geqslant T_{2,j}^{(n)} \geqslant \cdots \geqslant T_{m,j}^{(n)}. \tag{38}$$

## 6. Two-State Time-Invariant Rules

In the previous section, we were able to provide a short argument to prove that for all $m$ and $N$, the optimal time-varying rule is deterministic and likelihood ratio. By contrast, it is more difficult to establish the same degree of structure for the simplest of all time-invariant rules, those with only two states. Indeed, we have not yet been successful in generally establishing that the optimal two-state time-invariant rule is likelihood ratio.

Definition. A finite-memory time-invariant algorithm is likelihood ratio if it is likelihood ratio, as defined by Eqs. (27) and (28), with the dependence on $n$ deleted. This is equivalent to satisfying Eqs. (36), (37), and (38), with the dependence on $n$ deleted.

Reasoning as in Theorem 3, we find that randomizing the initial state of memory cannot lower $P_N(e)$, and therefore, we can restrict attention to the two deterministic choices $T_0 = 1$ and $T_0 = 2$. Using the notation of Section 4,

$$\begin{aligned} \alpha_0 &= p_{12}^0 \qquad \alpha_1 = p_{12}^1 \\ \beta_0 &= p_{21}^0 \qquad \beta_1 = p_{21}^1 \end{aligned} \tag{39}$$

and known formulas (Parzen, 1962) for the $N$-step transition matrix of a two-state Markov chain, we find

$$P_N(e) = \frac{\pi_0}{\alpha_0 + \beta_0}[\beta_0 + \alpha_0(1 - \alpha_0 - \beta_0)^N] + \frac{\pi_1}{\alpha_1 + \beta_1}\alpha_1[1 - (1 - \alpha_1 - \beta_1)^N], \qquad T_0 = 1 \tag{40a}$$

and

$$P_N(e) = \frac{\pi_0}{\alpha_0 + \beta_0}\beta_0[1 - (1 - \alpha_0 - \beta_0)^N] + \frac{\pi_1}{\alpha_1 + \beta_1}[\alpha_1 + \beta_1(1 - \alpha_1 - \beta_1)^N], \qquad T_0 = 2. \tag{40b}$$

Theorem 5. *Restricting attention to the class of rules for which* $\alpha_0 + \beta_0 \leqslant 1$ *or* $\alpha_1 + \beta_1 \leqslant 1$, *the optimal two-state time-invariant rule is of the likelihood ratio type.*

*Remarks.* (1) If we accept that the unrestricted optimal rule has disjoint transition regions (i.e., if no observation can cause transitions, both from

state 1 to 2 and from state 2 to 1), then the optimal rule lies in the class of restricted rules and is thus likelihood ratio.

(2) As noted below, Freedman (1971) has shown that for symmetric initial distributions [i.e., $\Pr(T_0 = 1) = \Pr(T_0 = 2) = \frac{1}{2}$] and $P^*(2, N) \leqslant \frac{1}{4}$, the optimal rule must have disjoint transition regions. Although a randomized initial state cannot lower the error probability, we believe that for symmetric problems, a symmetric distribution does not substantially increase the error probability.

*Proof.* Suppose we are given the optimal values of $\alpha_0$ and $\beta_0$ and we wish to find the optimal values of $\alpha_1$ and $\beta_1$. Since $\alpha_0$ and $\beta_0$ determine $P_N(e \mid H_0)$, our problem is to minimize $P_N(e \mid H_1)$. Without loss of generality, assume that $d(1) = H_1$, $d(2) = H_0$. Using the Markov dependence, we see that

$$P_N(e \mid H_1) = \mu_2^1(N) \tag{41a}$$

$$= \mu_2^1(N-1)(1-\beta_1) + [1 - \mu_2^1(N-1)]\,\alpha_1 \tag{41b}$$

$$= \alpha_1 + \mu_2^1(N-1)(1-\alpha_1-\beta_1), \tag{41c}$$

where $\mu_i^k(N) = \Pr(T_N = i \mid H_k)$. If we were allowed to choose $\alpha_1$ and $\beta_1$ differently at each time (which we are not), then (41b) implies that to minimize $P_N(e \mid H_1)$, the last value of $\beta_1$ should be the maximum possible, and the last value of $\alpha_1$ should be the minimum possible.

If the sum of $\alpha_1$ and $\beta_1$ is at most one, then $(1 - \alpha_1 - \beta_1) \geqslant 0$, and from (41c), minimizing $P_N(e \mid H_1)$ is then equivalent to minimizing $\mu_2^1(N-1)$. However, this is $P_{N-1}(e \mid H_1)$, and proceeding a step at a time, we conclude that even if we were allowed to choose different $(\alpha_1, \beta_1)$ pairs at each stage, the best strategy would be to always choose the maximum possible value of $\beta_1$ and the minimum possible value of $\alpha_1$. Since the minimum probability of error at time $N$ is no higher for the problem which allows a variable choice of $(\alpha_1, \beta_1)$, we conclude that the above strategy is also best for the original problem.

Note that if $\alpha_0 + \beta_0 < 1$, then the values of $\alpha_1$ and $\beta_1$ which result in the first step also will sum to less than one. This can be seen from Section 4. A similar argument can be made with the values of $\alpha_1$ and $\beta_1$ fixed, and $\alpha_0$ and $\beta_0$ being variable, completing the proof.

Using this line of reasoning also allows us to prove the following theorem, which will be useful in the next section.

THEOREM 6. *If $\gamma = \infty$, then there exists an integer $N_0$, such that for $N \geqslant N_0$, the optimal transition rule satisfies $\alpha_0^* + \beta_0^* \leqslant 1$ and is therefore a likelihood ratio rule.*

*Proof.* From (8) and (5), we have $\gamma = \infty \Rightarrow P^*(2, \infty) = 0 \Rightarrow \lim_{N\to\infty} P^*(2, N) = 0$. We shall consider the case $\Pr\{l(X) = 0 \text{ or } \infty\} = 0$. The cases $\Pr\{l(X) = 0\} > 0$, $\Pr\{l(X) = \infty\} > 0$, follow in a similar fashion.

Suppose $\alpha_0^* + \beta_0^* > 1$. Then, from the arguments of the preceding theorem, $\alpha_1^* + \beta_1^* > 1$, and from (41c),

$$\begin{aligned} P_N(e \mid H_1) &\geqslant \alpha_1^* + (1 - \alpha_1^* - \beta_1^*) \\ &= 1 - \beta_1^*. \end{aligned} \tag{42}$$

But, if $P^*(2, N) \to 0$, then $\beta_1^*(N) \to 1$, and therefore $\beta_0^*(N) \to 1$. Using a similar argument involving $P_N(e \mid H_0)$, we conclude that $\alpha_0^*(N), \alpha_1^*(N) \to 1$ also. Therefore, for any $\epsilon > 0$ and $N$ large enough,

$$\alpha_0^*, \alpha_1^*, \beta_0^*, \beta_1^* \geqslant 1 - \epsilon. \tag{43}$$

Then, if $T_0 = 1$ (similar reasoning applies if $T_0 = 2$),

$$\begin{aligned} P^*(e) &= \frac{\pi_0}{\alpha_0^* + \beta_0^*}\left[\beta_0^* + \alpha_0^*(1 - \alpha_0^* - \beta_0^*)^N\right] \\ &\quad + \frac{\pi_1\alpha_1^*}{\alpha_1^* + \beta_1^*}\left[1 - (1 - \alpha_1^* - \beta_1^*)^N\right] \\ &\geqslant \begin{cases} \frac{\pi_0}{2}(1 - \epsilon) & \text{if } N \text{ even} \\ \frac{\pi_1}{2}(1 - \epsilon) & \text{if } N \text{ odd} \end{cases} \\ &\geqslant \tfrac{1}{2}\min\{\pi_0, \pi_1\} - \epsilon \not\to 0. \end{aligned} \tag{44}$$

Q.E.D.

Another line of reasoning is also possible and yields the most general theorem we have been able to establish on the structure of optimal two-state time-invariant algorithms; namely, that the state transition function must be on the boundary of its allowable region. This almost implies that the optimal rule is likelihood ratio. First, we establish a lemma concerning the gradient of $P_N(e)$, considered as a function of $\alpha_0$, $\beta_0$, $\alpha_1$, and $\beta_1$.

LEMMA 4. (1) *If $N$ is odd and $T_0 = 1$, then $\partial P_N(e)/\partial\alpha_0 \leqslant 0$ and $\partial P_N(e)/\partial\alpha_1 \geqslant 0$.*

(2) *If $N$ is odd and $T_0 = 2$, then $\partial P_N(e)/\partial\beta_0 \geqslant 0$ and $\partial P_N(e)/\partial\beta_1 \leqslant 0$.*

(3) *If $N$ is even and $T_0 = 1$, then $\partial P_N(e)/\partial\beta_0 \geqslant 0$ and $\partial P_N(e)/\partial\beta_1 \leqslant 0$.*

(4) *If $N$ is even and $T_0 = 2$, then $\partial P_N(e)/\partial\alpha_0 \leqslant 0$ and $\partial P_N(e)/\partial\alpha_1 \geqslant 0$.*

(5) *If $0 < \alpha_0, \alpha_1, \beta_0, \beta_1 < 1$, then the weak inequalities in* (1) *to* (4) *above can be replaced by strict inequalities.*

*Proof.* Taking (40a) and interchanging $\pi_0$ and $\pi_1$, $\alpha_0$ and $\beta_1$, and $\alpha_1$ and $\beta_0$, gives (40b). Therefore, we need only prove (1), (3), and (5) above. Considering (1) first, we find from (40a) that

$$\partial P_N(e)/\partial\alpha_0 = (-\pi_0/\delta_0^2)\{\beta_0[1 - (1 - \delta_0)^N] + N\alpha_0\delta_0(1 - \delta_0)^{N-1}\}, \quad T_0 = 1 \tag{45}$$

where for ease of notation, we have defined $\delta_0 = \alpha_0 + \beta_0$. Since $0 \leqslant \delta_0 \leqslant 2$, the first term is clearly nonnegative, and unless $\beta_0 = 0$, it is positive. Because $N - 1$ is even, the second term is nonnegative, and it is positive unless $\alpha_0 = 0$ or $\delta_0 = 1$. We have thus established the first part of (1) above. We have also established (5) for this case since both terms in (45) are zero if and only if $\beta_0 = 0$ and $\alpha_0 = 0$ or 1.

Similarly, from (40a),

$$\partial P_N(e)/\partial\alpha_1 = (\pi_1/\delta_1^2)\{\beta_1[1 - (1 - \delta_1)^N] + N\alpha_1\delta_1(1 - \delta_1)^{N-1}\}, \quad T_0 = 1 \tag{46}$$

where $\delta_1 = \alpha_1 + \beta_1$. Reasoning in a similar manner, we find both terms to be nonnegative and the sum to be positive unless $\beta_1 = 0$ and $\alpha_1 = 0$ or 1. This completes the proof of (1) [and by analogy, the proof of (2)] and the portions of (5) relating thereto.

Proceeding to the proof of (3), we first differentiate (40a) to obtain

$$\partial P_N(e)/\partial\beta_0 = (-\pi_0\alpha_0/\delta_0^2)[(1 - \delta_0)^{N-1}(1 + (N - 1)\,\delta_0) - 1], \qquad T_0 = 1 \tag{47}$$

and

$$\partial P_N(e)/\partial\beta_1 = (\pi_1\alpha_1/\delta_1^2)[(1 - \delta_1)^{N-1}(1 + (N - 1)\,\delta_1) - 1], \qquad T_0 = 1. \tag{48}$$

To show that $\partial P_N(e)/\partial\beta_0 \geqslant 0$ for $N$ even, it suffices to show that

$$f_{N-1}(\delta_0) \equiv (1 - \delta_0)^{N-1}(1 + (N - 1)\,\delta_0) - 1 \tag{49}$$

is nonpositive for $N - 1$ odd and $0 \leqslant \delta_0 \leqslant 2$. Similarly, this would show that

$$f_{N-1}(\delta_1) = (1 - \delta_1)^{N-1}(1 + (N - 1)\,\delta_1) - 1 \tag{50}$$

in (48) is nonpositive for $N$ even and $0 \leqslant \delta_1 \leqslant 2$, thereby establishing $\partial P_N(e)/\partial\beta_1 \leqslant 0$. We find that

$$f_{N-1}(0) = 0 \tag{51}$$

and

$$df_{N-1}(x)/dx = -N(N-1)\,x(1-x)^{N-2}, \tag{52}$$

which imply

$$f_{N-1}(x) \leqslant 0 \qquad \text{for} \quad 0 \leqslant x \leqslant 2 \text{ and } N \text{ even}. \tag{53}$$

Furthermore, we see that (47) can equal zero only if $\alpha_0 = 0$ or $f_{N-1}(\delta_0) = 0$, which would imply $\alpha_0 = \beta_0 = 0$. Similarly, (48) can equal zero only if $\alpha_1 = 0$.

It is now possible to establish the following theorem.

THEOREM 7. *For all $N$, the optimal two-state time-invariant rule is either likelihood ratio [i.e., $(\alpha_0{}^*, \alpha_1{}^*)$ is on the lower boundary of Fig. 1 and $(\beta_0{}^*, \beta_1{}^*)$ is on the upper boundary]; or both $(\alpha_0{}^*, \alpha_1{}^*)$ and $(\beta_0{}^*, \beta_1{}^*)$ are on the lower boundary; or both $(\alpha_0{}^*, \alpha_1{}^*)$ and $(\beta_0{}^*, \beta_1{}^*)$ are on the upper boundary.*

*Remark.* Clearly, only the first of these conditions makes intuitive sense. Still, one must show the other two conditions preclude optimality. This is an open problem. This theorem shows that optimal transition rules lies on the boundary of the convex set of allowed transition rules.

*Proof.* First, consider $N$ odd. If the optimal initial state is $T_0{}^* = 1$, we fix $\beta_0$ and $\beta_1$ and optimize over $\alpha_0$ and $\alpha_1$. The region of allowable $(\alpha_0, \alpha_1)$ pairs is as shown in Fig. 1. From part (1) of Lemma 4, we find that $P_N(e)$ is minimized by $(\alpha_0, \alpha_1)$ along the lower boundary of Fig. 1. From (22), we see that this implies $p_{12}(x)$ corresponds to a likelihood ratio rule.

Now considering $p_{21}(x)$, or equivalently, $(\beta_0, \beta_1)$, we will show that $(\beta_0{}^*, \beta_1{}^*)$ cannot lie in the interior of the region shown in Fig. 1. If $(\beta_0{}^*, \beta_1{}^*)$ were in the interior, then necessarily

$$\frac{\partial P_N(e)}{\partial \beta_0} = \frac{\partial P_N(e)}{\partial \beta_1} = 0 \Big|_{(\alpha_0{}^*, \alpha_1{}^*, \beta_0{}^*, \beta_1{}^*)}. \tag{54}$$

From (47), (48), and (49), we find

$$\partial P_N(e)/\partial\beta_0 = (-\pi_0\alpha_0/\delta_0{}^2) f_{N-1}(\delta_0), \qquad T_0 = 1 \tag{55}$$

$$\partial P_N(e)/\partial\beta_1 = (\pi_1\alpha_1/\delta_1{}^2) f_{N-1}(\delta_1), \qquad T_0 = 1. \tag{56}$$

Since $\alpha_0^* > 0$ and $\alpha_1^* > 0$, (54) would imply

$$f_{N-1}(\delta_0^*) = f_{N-1}(\delta_1^*) = 0. \tag{57}$$

From (50), (51), and (52), we see that when $N$ is odd, $f_{N-1}(x)$ is

(1) negative and decreasing for $0 < x < 1$,
(2) increasing for $x > 1$, and
(3) $f_{N-1}(2) = 2(n-1)$.

This function therefore has a unique positive root between $x = 1$ and $x = 2$. Equation (57) would then imply $\delta_0^* = \delta_1^*$. However,

$$\frac{\partial^2 P_N(e)}{\partial \beta_0^2} = \frac{\pi_0 \alpha_0 \delta_0 N(N-1)(1-\delta_0)^N}{\delta_0^2(1-\delta_0)^2}, \qquad T_0 = 1 \tag{58}$$

$$\frac{\partial^2 P_N(e)}{\partial \beta_1^2} = \frac{-\pi_1 \alpha_1 \delta_1 N(N-1)(1-\delta_1)^N}{\delta_1^2(1-\delta_1)^2}, \qquad T_0 = 1 \tag{59}$$

so that these second partials would differ in sign at $\delta_0^* = \delta_1^*$, an impossibility for a local minimum. Thus, $(\beta_0^*, \beta_1^*)$ must lie on the boundary of its allowable region when $N$ is odd and $T_0^* = 1$.

When $N$ is odd and $T_0^* = 2$, we repeat the above argument, merely interchanging the roles of $\alpha$ and $\beta$ and of $H_0$ and $H_1$. We then find that $(\beta_0^*, \beta_1^*)$ must lie on the upper boundary, which is consistent with the optimal rule being likelihood ratio, and $(\alpha_0^*, \alpha_1^*)$ must lie on either the lower or the upper boundary. This completes the proof for $N$ odd.

We now turn to the proof when $N$ is even. First, if $T_0^* = 1$, we find from Lemma 4, part 3, that $\partial P_N(e)/\partial \beta_0 > 0$ and $\partial P_N(e)/\partial \beta_1 < 0$, which imply that $(\beta_0^*, \beta_1^*)$ must lie on the upper boundary.

If $(\alpha_0^*, \alpha_1^*)$ were to lie in the interior of its allowable region, we would have

$$\partial P_N(e)/\partial \alpha_0 = \partial P_N(e)/\partial \alpha_1 = 0 \,|_{(\alpha_0^*, \alpha_1^*, \beta_0^*, \beta_1^*)}.$$

This can be shown to imply that $\partial^2 P_N(e)/\partial \alpha_0^2$ and $\partial^2 P_N(e)/\partial \alpha_1^2$ differ in sign at $(\alpha_0^*, \alpha_1^*, \beta_0^*, \beta_1^*)$, a clear contradiction. We therefore know that for $N$ even and $T_0^* = 1$, $(\alpha_0^*, \alpha_1^*)$ must lie on the boundary of its allowable region.

Similarly, if $N$ is even and $T_0^* = 2$, we find that $(\alpha_0^*, \alpha_1^*)$ must lie on the lower boundary, and $(\beta_0^*, \beta_1^*)$ on the boundary of their allowable regions. Q.E.D.

Freedman (1971) has also considered the problem of finding the structure of the optimal two-state rule. He symmetrizes the problem in that $T_0$, the

initial state of memory, is chosen at random with both values being equally likely. Also, since $P^*(2, N) < \frac{1}{4}$ for his problem of interest [testing $\mathcal{N}(+1, 1)$ vs $\mathcal{N}(-1, 1)$], he assumes that $P^*(2, N) < \frac{1}{4}$. Under these assumptions, he shows that the optimal rule must be likelihood ratio. The proofs are somewhat involved, and thus, we will only give the reasoning which leads to the final result. The interested reader is referred to Freedman (1971) for details. The three lemmas and theorem that follow all assume symmetry in $T_0$ and that $P^*(2, N) < \frac{1}{4}$.

LEMMA 5.

$$\alpha_0^*/\beta_0^* > 1 \qquad \textit{and} \qquad \alpha_1^*/\beta_1^* < 1. \tag{60}$$

LEMMA 6. *The optimal rule must have disjoint transition regions. More precisely,*

$$\Pr\left\{\begin{matrix}X \textit{ causes transitions from state 1 to} \\ \textit{state 2 and from state 2 to state 1}\end{matrix}\right\} = 0. \tag{61}$$

LEMMA 7. *Considering $P_N(e)$ as a function of $\alpha_0$, $\beta_1$, $\beta_0/\alpha_0$, and $\alpha_1/\beta_1$, the optimal rule has $\partial P_N(e)/\partial(\beta_0/\alpha_0) > 0$, $\partial P_N(e)/\partial(\alpha_1/\beta_1) > 0$, $\partial P_N(e)/\partial\alpha_0 < 0$, and $\partial P_N(e)/\partial\beta_1 < 0$.*

THEOREM 8. *The optimal rule is likelihood ratio.*

In summary, we see that under certain mild assumptions or in special cases, we can show the optimal two-state time-invariant rule is likelihood ratio. This, together with the result that optimal time-varying $m$-state rules are likelihood ratio, lends credence to the assumption that even without restrictions, the optimal time-invariant rule is likelihood ratio.

## 7. A GAUSSIAN PROBLEM

Let us consider the special case where the two probability measures $\mathcal{P}_0$ and $\mathcal{P}_1$ are both Gaussian with variance one, but $\mathcal{P}_0$ has mean $+1$ and $\mathcal{P}_1$ has mean $-1$. Then, $l(x) = \exp(2x)$ so that $l_{\max} = \infty$, $l_{\min} = 0$, $\gamma = \infty$, and $P^*(2, \infty) = 0$. Theorem 6 of the preceding section ensures that for $N$ large enough, the optimal two-state time-invariant rule is a likelihood ratio rule. For $N = 1$, the optimal rule is also a likelihood ratio rule and corresponds to the Bayes decision rule.

For equal *a priori* probabilities and a symmetric random choice of initial state, Freeman (1971) has shown that the optimal rule is not only likelihood ratio, but also symmetric (i.e., $\alpha_0^* = \beta_1^*$, $\alpha_1^* = \beta_1^*$). This implies

$$\{x: p_{12}(x) = 1\} = \{x: x \geqslant M\}$$
$$\text{and} \tag{62}$$
$$\{x: p_{21}(x) = 1\} = \{x: x \leqslant -M\}$$

for some $0 < M < \infty$. This greatly reduces the problem of searching for the optimal two-state algorithm since $P_N(e \mid H_0) = P_N(e \mid H_1)$, and we need only minimize over $0 < M < \infty$.

As previously noted, a deterministic initial state is optimal, and it is therefore possible that $P_N(e)$ is increased by taking a symmetric initial distribution on $T_0$. However, the increased symmetry makes the problem tractable, and for reasons to be developed later, the increase in $P_N(e)$ should be minimal. Therefore, let us find $P_s^*(2, N)$, the minimum error probability achievable by a two-state time-invariant rule after $N$ observations when the initial distribution is symmetric.

Since $P_N(e \mid H_0) = P_N(e \mid H_1)$, these values will be denoted by $P_N(M)$

$$\begin{aligned} P_N(M) &= \mu_1{}^0(N) = \mu_2{}^1(N) \\ &= \frac{r}{1+r} + \frac{1}{2}\frac{1-r}{1+r}[1 - q(r+1)]^N, \end{aligned} \tag{63}$$

where

$$q = \alpha_0 = \beta_1 = Q(M-1) \qquad r = \frac{\alpha_1}{\alpha_0} = \frac{\beta_0}{\beta_1} = \frac{Q(M+1)}{Q(M-1)} \tag{64}$$

and

$$Q(\alpha) = \frac{1}{(2\pi)^{1/2}} \int_\alpha^\infty \exp[-x^2/2]\, dx.$$

The following bounds will be useful:

$$\left[1 - \frac{1}{(M-1)^2}\right] \leqslant [(2\pi)^{1/2}(M-1)\exp(\tfrac{1}{2}(M-1)^2)]q \leqslant 1 \qquad M \geqslant 1 \tag{65}$$

$$\left[1 - \frac{1}{(M+1)^2}\right] \leqslant \left[\frac{M+1}{M-1}\exp(2M)\right] r \leqslant \left(1 - \frac{1}{(M-1)^2}\right)^{-1} \quad M \geqslant 2 \tag{66}$$

$$r \leqslant \exp(-2M). \tag{67}$$

Bounds (65) and (66) are taken directly from Wozencraft and Jacobs (1965), while (67) holds because

$$r = \frac{\int_M^\infty dx \exp[-(x+1)^2/2]}{\int_M^\infty dx \exp[-(x-1)^2/2]}$$

$$= \frac{\int_M^\infty dx \exp[-(x-1)^2/2] \exp[-2x]}{\int_M^\infty dx \exp[-(x-1)^2/2]} \leqslant \exp(-2M). \tag{68}$$

To establish an upper bound on $P_s^*(2, N)$, from (63) and (67), we obtain

$$P_N(M) < r + \tfrac{1}{2}(1-q)^N < \exp(-2M) + \tfrac{1}{2}\exp(-Nq). \tag{69}$$

Now, (65) gives us

$$q = \frac{\theta(M)}{(2\pi)^{1/2}(M-1)} \exp[-\tfrac{1}{2}(M-1)^2], \tag{70}$$

where $\frac{3}{4} < \theta(M) < 1$ for $M \geqslant 3$. Let

$$M_\epsilon \triangleq (2 \ln N)^{1/2} + 1 - \epsilon \qquad \text{where} \qquad 0 < \epsilon < 1. \tag{71}$$

Then, setting $K = (2 \ln N)^{1/2}$, (69), (70), and (71) yield for $M_\epsilon \geqslant 3$

$$P_s^*(2, N) < P_N(M_\epsilon) < \exp[-2(K + 1 - \epsilon)] + \frac{1}{2}\exp\left[\frac{-\frac{3}{4}}{(2\pi)^{1/2}(K-\epsilon)} \exp\left(\epsilon K - \frac{\epsilon^2}{2}\right)\right]. \tag{72}$$

Now, for a fixed $\epsilon$, as $K$ grows large, the ratio of the second term to the first term in (72) tends to zero. Thus, for any $\epsilon > 0$ which leaves $M_\epsilon \geqslant 3$, there exists an $N_\epsilon$ such that if $N \geqslant N_\epsilon$, then

$$P_s^*(2, N) < (1+\epsilon) \exp -2((2 \ln N)^{1/2} + 1 - \epsilon) \tag{73}$$

[i.e., the ratio of the second term to the first term in (72) is less than $\epsilon$].

To obtain a lower bound to $P_s^*(2, N)$, we write

$$P_N(M) = A(M) + B(M), \tag{74}$$

where

$$A(M) = \frac{r}{1+r}, \qquad B(M) = \frac{1}{2}\frac{1-r}{1+r}[1 - q(r+1)]^N. \tag{75}$$

Since, as can be easily verified, $dr/dM$, $dq/dM < 0$, $A(M)$ is decreasing in $M$, and $B(M)$ is increasing in $M$. Therefore, we have

LEMMA 8. *For any* $M > 0$, $P_s^*(2, N) > \min\{A(M), B(M)\}$.

For large $N$, $B(M)$ is almost discontinuous at

$$M_0 = (2 \ln N)^{1/2} + 1. \tag{76}$$

This is because the maximum of $N$ independent $\mathcal{N}(0, 1)$ random variables converges to $(2 \ln N)^{1/2}$ in probability. $A(M)$, on the other hand, is relatively well behaved, varying like $\exp(-2M)$ for large $M$. This behavior is indicated in Fig. 4. It is seen that the minimum of $P_N(M) = A(M) + B(M)$ occurs

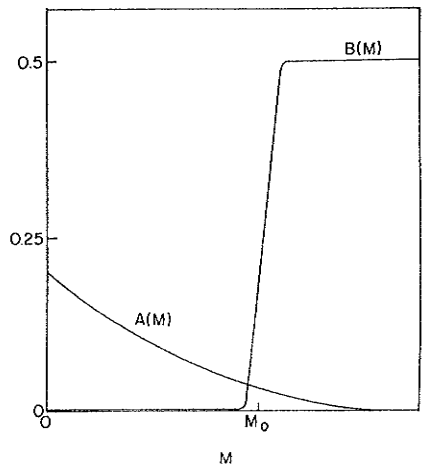


FIG. 4. Steady-state and transient terms.

slightly to the left of $M_0$ and is essentially $\exp(-2M_0)$. This explains why the upper bound (73), taken by choosing $M = M_0 - \epsilon$, will be tight as $N \to \infty$ and $\epsilon \to 0$.

LEMMA 9. *For* $N \geqslant 2$, $A(M_0) < B(M_0)$.

*Proof.* From (67) and (73),

$$A(M_0) = r/(1 + r) < r < \exp(-2M_0). \tag{77}$$

For $N \geqslant 1$, we have $M_0 \geqslant 1$ and $r \leqslant 0.05$. Thus, for $N \geqslant 1$,

$$B(M_0) = \frac{1}{2}\frac{1-r}{1+r}[1 - q(r+1)]^N \geqslant 0.45(1 - 1.05q)^N. \tag{78}$$

From (65),

$$q(M_0) < \exp[-(M_0 - 1)^2/2] = 1/N \tag{79}$$

so

$$B(M_0) \geqslant 0.45(1 - 1.05/N)^N. \tag{80}$$

Now,

$$(1 - \theta) > \exp(-2\theta) \qquad \text{for} \quad 0 < \theta < 0.79. \tag{81}$$

Thus, for $N \geqslant 2$,

$$B(M_0) \geqslant 0.45 \exp(-2.1) = 0.055. \tag{82}$$

Therefore, $A(M_0) < B(M_0)$, as long as $N \geqslant 2$ and

$$\exp(-2M_0) \leqslant 0.055. \tag{83}$$

Since $M_0 = (2 \ln N)^{1/2} + 1$, we see that (70) is satisfied for $N \geqslant 2$. Q.E.D.

From Lemmas 8 and 9, we conclude

$$P_s^*(2, N) > \frac{r(M_0)}{1 + r(M_0)} \qquad N \geqslant 2. \tag{84}$$

To put the lower bound in its final form, we use

$$r > \left(\frac{M-1}{M+1}\right)^2 \exp(-2M) \qquad M \geqslant 1. \tag{85}$$

This inequality follows from (53) and the fact that for $M > 1$,

$$\left(\frac{M-1}{M+1}\right)\left(1 - \frac{1}{(M+1)^2}\right) > \left(\frac{M-1}{M+1}\right)^2. \tag{86}$$

Finally, using (84), (85), and (76), we have for $N \geqslant 2$

$$P_s^*(2, N) > \frac{(2 \ln N)/(2 + (2 \ln N)^{1/2})^2}{1 + \exp(-2((2 \ln N)^{1/2} + 1))} \exp(-2((2 \ln N)^{1/2} + 1)). \tag{87}$$

Note that as $N \to \infty$, the ratio of the upper and lower bounds, (73) and (87), tends to one.

It is possible to extend the above reasoning and to obtain thereby:

THEOREM 9. *If, under $H_0$ the distribution on $X$ is $\mathcal{N}(\mu_1, \sigma^2)$, while under $H_1$ it is $\mathcal{N}(\mu_2, \sigma^2)$; and, if the hypotheses are equally likely, then, letting* $2\mu = |\mu_2 - \mu_1|$,

$$\lim_{N\to\infty} \frac{P_s^*(2, N)}{\exp[-(2\mu/\sigma)((2 \ln N)^{1/2} + \mu/\sigma)]} = 1. \tag{88}$$

A program was written to find numerical values of $P_s^*(2, N)$ for various values of $N$ in the range $1 \leqslant N \leqslant 10^{64}$. Figure 5 shows a plot of $P_s^*(2, N)$

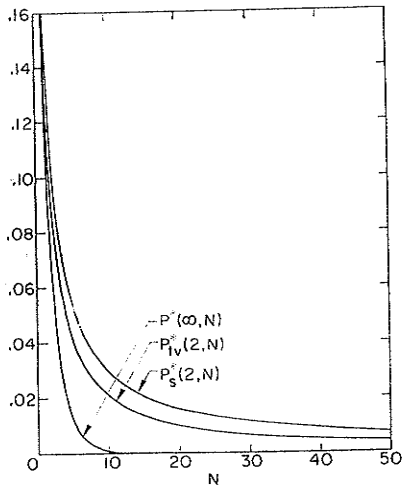


FIG. 5. Comparison of $P_s^*(2, N)$, $P_{tv}^*(2, N)$, and $P^*(\infty, N)$ for $1 \leqslant N \leqslant 50$.

and also of $P^*(\infty, N)$ for $1 \leqslant N \leqslant 50$. Although both have asymptotic value zero, a marked difference in rate of approach is evident. Table I further contrasts this difference by showing the number of observations required by both two-state and infinite-state memories to achieve certain error probabilities.

Figure 5 also shows the minimum error probability achievable by a two-state time-varying rule. This curve is strictly below the time-invariant curve, as it must be. However, for a large number of observations ($N \approx 50$) the ordinates differ by less than a factor of 2. Although not shown in this figure, approximately the same ratio holds at $N = 1000$. This behavior is due to the fact that the time-varying rule has time-varying thresholds

TABLE I

| Desired error probability | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-15}$ |
|---|---|---|---|---|---|
| Number of observations required by two-state memory | 33 | ~660 | ~24,000 | $\sim 6.3 \times 10^{6}$ | $\sim 6.3 \times 10^{54}$ |
| Number of observations required by $\infty$-state memory | 6 | 10 | 14 | 19 | 69 |

$\{M_n\}_{n=1}^{N}$. Initially, these thresholds are small (e.g., $M_1 = 0.0$, $M_2 = 0.834$), but they increase with $n$. Some thought shows that $M_n^*$ must increase like $(2 \ln n)^{1/2} + 1$. If it increases any more quickly, very few transitions occur, while if it increases any more slowly, time-varying rules would be inferior to time-variant rules, an impossibility. The slow rate of increase of $(2 \ln n)^{1/2}$ ensures that for a fraction of time tending to one, $M_n^*$ is essentially the same as $M_N^*$. Since the transient term $B(M)$ is negligible at $M = M_N^*$, the superior "initial distribution" achieved by the time-varying rule during its early stages is of negligible value.

Similarly, because of the negligible value of the transient term, we conjecture that $P^*(2, N)$ and $P_s^*(2, N)$ are asymptotically equal. Figure 6 plots the ratio $P_s^*(2, N)/P^*(2, N)$ for $1 \leqslant N \leqslant 1000$ and supports the conjecture.

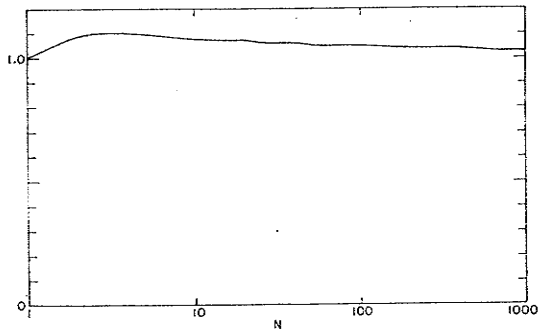


FIG. 6. Plot of $P_s^*(2, N)/P^*(2, N)$ for $1 \leqslant N \leqslant 1000$.

Figure 7 compares $P^*(2, N)$ with its asymptotic form $\exp[-2((2 \ln N)^{1/2}+1)]$. It is seen that for $(2 \ln N)^{1/2} + 1 > 7$, $P^*(2, N)$ is within a factor of 3 of its

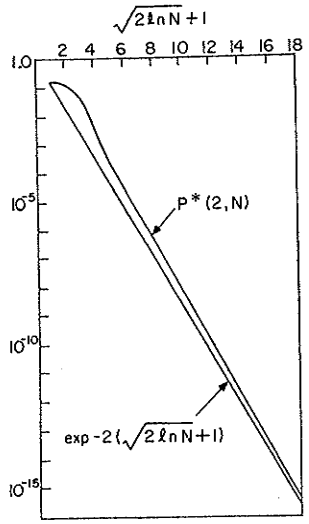


FIG. 7. Plot of $P_s^*(2, N)$ and its asymptote.

asymptote. However, this corresponds to $N > e^{18} \sim 10^8$. Thus, for small values of $N$, the asymptotic expression gives only a general idea as to behavior.

## 8. CONJECTURED BEHAVIOR OF $P^*(m, N)$

For the hypothesis test $\mathscr{P}_0 = \mathscr{N}(1, 1)$ vs $\mathscr{P}_1 = \mathscr{N}(-1, 1)$, we conjecture that for all $m < \infty$,

$$\lim_{N\to\infty} \frac{P^*(m, N)}{\exp[-2(m-1)((2 \ln N)^{1/2} + 1)]} = 1. \tag{89}$$

Our reasoning is based on the assumption that the optimal rule is likelihood ratio for all $m$ and $N$. If a transition is made from state $i$ to state $i + 1$ when $X = x_1$ and from state $i + 1$ to state $i + 2$ when $X = x_2$, one would expect that when $X = x_3 = x_1 + x_2$, a transition would occur from state $i$ to state $i + 2$, since the information gained (the likelihood ratio) is the same for observing $x_1$, $x_2$ successively as for observing $x_3$ by itself. Thus, the optimal $m$-state rule will have multistate transitions. If the minimum value of $X$ causing a transition between adjacent states is $x_1$, then the minimum value to cause double level transitions should be approximately $2x_1$; the minimum value to cause triple-level transitions should be approximately $3x_1$; etc.

If $x_1$ is chosen close to $(2 \ln N)^{1/2}$, then for large $N$, multilevel transitions occur with negligible probability. If $x_1$ is chosen to be less than $0.5\,(2 \ln N)^{1/2}$, then multilevel transitions cannot be neglected. However, the maximum information (log likelihood ratio), which can be accumulated in favor of either hypothesis, is now less than one-half of what it would be if $x_1 \approx (2 \ln N)^{1/2}$. And, since, for the two-state machine, choosing $x_1 \approx (2 \ln N)^{1/2}$ causes the transient term to be unimportant (i.e., the machine is essentially in steady state), we believe the same behavior will apply to $m$-state machines. This has to do with the fact that for any $\epsilon > 0$, the expected number of observations which exceed $(2 \ln N)^{1/2} + E(X) + \epsilon$ tends to zero, and yet, the expected number of observations which exceed $(2 \ln N)^{1/2} + E(X) - \epsilon$ tends to infinity.

Therefore, the optimal choice of $x_1$ will be slightly less than $(2 \ln N)^{1/2} + 1$ (i.e., approximately the same as $M_N{}^*$). Any larger value causes the machine never to change state, with probability close to one. Any smaller value causes the machine to be in essentially its steady-state mode, but since the steady-state error rate is decreasing in $x_1$, we want $x_1$ to be as large as possible.

Choosing $x_1 = (2 \ln N)^{1/2} + 1 - \epsilon_N$ allows us to neglect multilevel transitions. Adding the "$\delta$-traps" [see Eq. (9a) and Hellman and Cover (1970)], the resultant Markov chain is solved easily for its steady-state occupation probabilities, and the associated error rate is found to be $\exp\{-2(m-1)[(2 \ln N)^{1/2} + 1 - \epsilon_N]\}$. By letting $\epsilon_N \to 0$ as $N \to \infty$, we obtain (89).

The above reasoning shows why Lynn and Boorstyn (1972) found that multilevel transitions were of little use in lowering the error rate, even for moderate values of $N$. By noting that deletion of the $\delta$-traps causes the error rate to behave approximately as $\exp\{-m[(2 \ln N)^{1/2} + 1]\}$, we see why they found that adding $\delta$-traps was quite effective in lowering the error rate. Actually, as noted, Lynn and Boorstyn did not use randomized $\delta$-traps, but achieved the same effect by using higher thresholds for transitions out of states 1 and $m$.

When $N$ is large, the trade-off between memory size and sample size is in favor of increased memory. Using the asymptotic formulas we find that, for testing between $\mathcal{N}(+1, 1)$ and $\mathcal{N}(-1, 1)$, approximately $10^7$ or $10^8$ samples are needed to obtain $P_N(e) = 10^{-6}$ when $m = 2$. By adding one state to memory, we obtain $m = 3$ and find that only 20 samples are needed. This latter result must be interpreted carefully, since, even with no memory limitation, an error rate of $3.4 \times 10^{-6}$ is all that is achievable with $N = 20$. Clearly, the asymptotic formula is not applicable for this choice of $m$ and $N$. However, this example points out an interesting fact: For small $N$, the

asymptotic expression for $P^*(m, N)$ is below the true curve for $P^*(\infty, N)$. By plotting the two (for fixed $m$), we can obtain an idea as to where the asymptotic formula is and is not applicable.

## 9. Summary

The structure of the $\epsilon$-optimal infinite-sample finite-memory rule is known from previous work for both the time-varying and time-invariant problems. This paper establishes the existence of optimal finite-memory rules for the finite sample problem. It further conjectures, and partially supports, that the optimal time-invariant rule is likelihood ratio. On the other hand, it is completely established that the structure of optimal time-varying rules is deterministic and likelihood ratio. Optimal time-varying and time-invariant rules share a deterministic initial state and decision function.

Received: August 27, 1973; revised: June 20, 1975

## References

Baxa, E. G., and Nolte, L. W. (1972), Adaptive signal detection with finite memory, *IEEE Trans. Syst., Man, Cybern.* **SMC-2**, 42–49.

Chandrasekaran, B., and Shen, D. W. S. (1968), On expediency and convergence in variable-structure automata, *IEEE Trans. Sys. Sci. and Cybern.* **SSC-4**.

Chandrasekaran, B., and Lam, C. C. (1975), A finite-memory deterministic algorithm for the symmetric hypothesis testing problem, *IEEE Trans. on Info. Theory* **IT-21**, 40–46.

Cover, T. M. (1969), Hypothesis testing with finite statistics, *Ann. Math. Stat.* **40**, 828–835.

Cover, T. M., and Hellman, M. E. (1970), The two-armed bandit problem with time invariant finite memory, *IEEE Trans. on Info. Theory* **IT-16**, 185–195.

Flower, R. A., and Hellman, M. E. (1972), Hypothesis testing with finite memory in finite time, *IEEE Trans. on Info. Theory* **IT-18**, 429–431 (Correspondence).

Freedman, M. (1971), "A Finite Memory, Finite Time, Gaussian Hypothesis Testing Problem," M.S. Thesis, Electrical Engineering Department, Mass. Inst. of Tech., Cambridge, Mass.

Fu, K. S., and Li, T. J. (1968), "On the Behavior of Learning Automata and its Applications," Purdue University Technical Report No. TR-EE 68-20.

Hellman, M. E., and Cover, T. M. (1971), On memory saved by randomization, *Ann. Math. Stat.* **42**, 1075–1078.

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

HELLMAN, M. E. (1972), The effects of randomization on finite memory decision schemes, *IEEE Trans. on Info. Theory* **IT-18**, 499–502 (Correspondence).

HELLMAN, M. E. (1974), Finite memory algorithms for estimating the mean of a Gaussian distribution, *IEEE Trans. on Info. Theory* **IT-20**, 382–384 (Correspondence).

HIRSCHLER, P., AND COVER, T. M. (1975), A finite memory test of the irrationality of the parameter of a coin, *Ann. Statistics*, to appear.

HOROS, J. A., AND HELLMAN, M. E. (1972), A confidence model for finite-memory learning systems, *IEEE Trans. on Info. Theory* **IT-18**, 811–813 (Correspondence).

ISBELL, J. R. (1959), On a problem of robbins, *Ann. Math. Stat.* **30**, 606–610.

KOPLOWITZ, J., AND ROBERTS, R. (1973), Sequential estimation with a finite statistic, *IEEE Trans. on Info. Theory* **IT-19**, 631–635.

KOPLOWITZ, J. (1974), A note on hypothesis testing with a finite statistic, *IEEE Trans. on Info. Theory*, to appear.

KRYLOV, V. Y. (1963), On one automaton that is asymptotically optimal in a random medium, *Avtomatika i Telemekhanika* **24**, 9, 1226–1228.

LYNN, P. F. (1971), "Finite Memory Detectors," Ph.D. Thesis, Poly. Inst. of Brooklyn, New York.

LYNN, P. F., AND BOORSTYN (1972), "Bounds on Finite Memory Detectors," presented at 1972 *IEEE International Symposium on Information Theory*, Asilomar, Calif., Jan. 31–Feb. 3.

MUISE, R. (1971), "Optimum, Time-Varying, Finite-Memory Detection," Ph.D. Thesis, Poly. Inst. of Brooklyn, New York.

MUISE, R. W., AND BOORSTYN, R. R. (1972), Detection with time-varying finite-memory receivers, *Abstracts of Papers*, 1972 *IEEE International Symposium on Information Theory*, 35–36.

MULLIS, C. T. (1968), "A Class of Finite Memory Decision Processes," M.S. Thesis, University of Colorado.

MULLIS, C. T., AND ROBERTS, R. A. (1968), Memory limitation and multistage decision processes, *IEEE Trans. on Sys. Sci. and Cybern.* **SSC-4**, 307–316.

MULLIS, C. T., AND ROBERTS, R. A. (1974), Finite memory problems and algorithms, *IEEE Trans. on Info. Theory* **IT-20**, 440–455.

PARZEN, E. (1962), "Stochastic Processes," p. 197, Holden-Day, San Francisco.

ROBBINS, H. (1956), A sequential decision problem with a finite memory, *Proc. Nat. Acad. Sci. USA* **42**, 920–933.

ROBERTS, R. A., AND TOOLEY, J. R. (1970), Estimation with finite memory, *IEEE Trans. on Info. Theory* **IT-16**, 685–691.

SAMANIEGO, F. (1973), Estimating a binomial parameter with finite memory, *IEEE Trans. on Info. Theory* **IT-19**, 636–643.

SAMANIEGO, F. (1974), On tests with finite memory in finite time, *IEEE Trans. on Info. Theory* **IT-20**, 387–388.

SAMUELS, S. M. (1968), Randomized rules for the two-armed bandit with finite memory, *Ann. Math. Stat.* **39**, 2103–2107.

SHUBERT, B., AND ANDERSON, C. (1973), Testing a simple symmetric hypothesis by a finite-memory deterministic algorithm, *IEEE Trans. on Info. Theory* **IT-19**, 644–647.

SHUBERT, B. (1974), Finite-memory classification of Bernoulli sequences using reference samples, *IEEE Trans. on Info. Theory* **IT-20**, 384–387.

TOOLEY, J. R., AND ROBERTS, R. A. (1973), On estimating random processes with finite memory, *IEEE Trans. Syst., Man, Cybern.* **SMC-3**, 294–299.

TSETLIN, M. L. (1961), On the behavior of finite automata in random media, *Avtomatika i Telemekhanika* **22**, 10, 1345–1354; available in English translation.

WAGNER, T. J. (1972), Estimation of the mean with time-varying finite memory, *IEEE Trans. on Info. Theory* **IT-18**, 523–525.

WOZENCRAFT, J., AND JACOBS, J. (1965), "Principles of Communication Engineering," Wiley, New York.

Printed by the St Catherine Press Ltd., Tempelhof 37, Bruges, Belgium.