The Two-Armed-Bandit Problem With Time-Invariant Finite Memory

THOMAS M. COVER AND MARTIN E. HELLMAN

Abstract—This paper solves the classical two-armed-bandit problem under the finite-memory constraint described below. Given are probability densities p_0 and p_1, and two experiments A and B. It is not known which density is associated with which experiment. Thus the experimental outcome Y of experiment A is as likely to be distributed according to p_0 as it is to be distributed according to p_1. It is desired to sequentially choose an experiment to be performed on the basis of past observations according to the algorithm $T_n = f(T_{n-1}, e_n, Y_n)$, $e_n = e(T_{n-1})$, where $T_n \in \{1, 2, \cdots, m\}$ is the state of memory at time n, $e_n \in \{A, B\}$ is the choice of experiment, and Y_n is the random variable observation. The goal is to maximize the asymptotic proportion r of uses of the experiment associated with density p_0.

Let $l(y) = p_0(y)/p_1(y)$, and let $\bar{l}$ and $\bar{\bar{l}}$ denote the almost everywhere greatest lower bound and least upper bound on $l(y)$. Let $l = \max\{\bar{\bar{l}}, 1/\bar{l}\}$. Then the optimal value of r, over all m-state algorithms (f, e), will be shown to be $l^{m-1}/(l^{m-1} + 1)$. An ϵ-optimal family of m-state algorithms will be demonstrated. In general, optimal algorithms do not exist, and ϵ-optimal algorithms require artificial randomization.

I. Introduction

SUPPOSE one is given two coins, labeled A and B. Suppose also that it is known that one of the coins has bias p_0 towards heads and the other has bias p_1 towards heads, but it is not known which coin has which bias. At each trial a coin is to be selected and tossed, and it is desired to maximize the proportion of heads (successes) achieved in the limit as the number of trials tends to infinity. An equivalent objective is to maximize the proportion of tosses using the coin with the larger bias. How should the choice of coin at trial n depend on the previous outcomes, in order to achieve this goal? This problem is commonly referred to as the sequential design of experiments or the two-armed-bandit problem (TABP) [1]–[3].

Note that this problem combines hypothesis testing (which coin has which bias?) with the added degree of freedom that the experimenter may select his experiment (A or B) at each toss. The experimenter must utilize his information to maximize the proportion of successes.

This paper will be concerned with a generalized TABP in which the "coins" may have an infinite number of sides. A further generalization of the TABP to an infinite number of coins will be provided in Section VI. These problems will be solved under a finite-memory constraint, i.e., the experimenter is not allowed to remember the outcomes of all previous trials, but only a finite-valued statistic. On the basis of this statistic, the next coin must be chosen.

Stated more precisely, the experimenter is provided two experiments, A and B. Also given are two probability measures $\mathcal{P}_0$ and $\mathcal{P}_1$ defined on the arbitrary probability space $(\mathcal{Y}, \mathcal{B})$, where $\mathcal{Y}$ is the experimental outcome space and $\mathcal{B}$ is a σ-field of subsets over $\mathcal{Y}$. There are two hypotheses concerning the probability distribution of the experimental outcome Y:

$$
\begin{aligned}
H_0 &: \begin{cases} Y \sim \mathcal{P}_0 \text{ under experiment } A \\ Y \sim \mathcal{P}_1 \text{ under experiment } B \end{cases} \\
H_1 &: \begin{cases} Y \sim \mathcal{P}_1 \text{ under experiment } A \\ Y \sim \mathcal{P}_0 \text{ under experiment } B. \end{cases}
\end{aligned}
\tag{1}
$$

Let the a priori probabilities of H_0 and H_1 be π_0 and π_1 respectively, where $\pi_0 + \pi_1 = 1$. This seemingly Bayesian formulation, in which the priors are specified, is not restrictive since the set of all admissible algorithms (or the set of all optimal algorithms with respect to the Neyman–Pearson formulation) may be generated by letting π_0 take on all values in the unit interval.

Let $e_i \in \{A, B\}$ denote the ith experiment performed and $Y_i \in \mathcal{Y}$ denote the ith experimental outcome. It is assumed that the experimental outcomes are independent in the sense that

$$P(Y_i, Y_j \mid e_i, e_j, H) = P(Y_i \mid e_i, H)P(Y_j \mid e_j, H) \qquad i \neq j,$$

where H is the true hypothesis.

A success is said to occur if the experiment associated with $\mathcal{P}_0$ is performed. At times $n = 1, 2, 3, \cdots$ a choice of experiment e_n is made. Letting

$$s_n = \begin{cases} 1, & \text{if success occurs at time } n, \\ 0, & \text{if failure occurs at time } n, \end{cases} \tag{2}$$

the objective is to maximize

$$r = E\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} s_i\right\} \tag{3}$$

where the expectation is taken with respect to the distribution on the two hypotheses and the distribution on $\{s_i\}$ induced by the experiment selection algorithm. Therefore, r is the expected long run proportion of successes.

Let the data be summarized by an m-valued statistic T that is updated according to the rule

$$T_n = f(T_{n-1}, X_n) \qquad T_n \in \{1, 2, \cdots, m\} \tag{4}$$

Manuscript received May 19, 1969; revised September 29, 1969. This work was supported in part under Contract AF 49(638) 1517 and under the NSF Graduate Fellowship Program.

T. M. Cover is with the Department of Electrical Engineering, Stanford University, Stanford, Calif. 94305.

M. E. Hellman was with the Watson IBM Research Center, Yorktown Heights, N. Y. He is now with the Massachusetts Institute of Technology, Cambridge, Mass. 02139.

where T_n is the value of T after n observations, $X_n = (e_n, Y_n)$ is the nth observation (note the difference between an observation $X = (e, Y)$ and an experimental outcome Y) and f is a stochastic function. Further, let e_n be constrained to depend on the past outcomes $X_1, X_2, \cdots, X_{n-1}$ only through T_{n-1}, according to the function

$$e_n = e(T_{n-1}) \qquad n = 1, 2, \cdots \tag{5}$$

where e: $\{1, 2, \cdots, m\} \rightarrow \{A, B\}$ is again allowed to be a stochastic function. (The randomization in the functions f and e must, to avoid cheating, be independent of the data.) The size of memory is defined to be m.

The objective is now to find the pair (f, e) that maximizes r for given m, π_0, $\mathcal{P}_0$, and $\mathcal{P}_1$. For a reformulation in terms of optimal finite-state machines see Section III.

As was previously mentioned, it is not only necessary to test H_0 versus H_1, but also to use the result of the test in an attempt to obtain successes. This produces a conflict. The experimenter may believe H_0 (in which case he should perform A) and yet he may wish to perform B if it would yield more information, thereby increasing the probability of success on future trials. The conflict is between a desire for immediate success and a desire to gather information.

Another conflict exists. A good test requires large memory, but, as mentioned, hypothesis testing may not yield a high proportion of successes. Thus, once the test is completed, a large number of experiments that use the result of the test is desired. However, an m-valued statistic can only "count to m." There is a problem in deciding how much memory to allocate to testing and how much to allocate to using the information gathered by testing. Fortunately, the optimal solution that we shall present suggests an interpretation answering this question. The surprising answer is that all of the states of memory may be devoted to hypothesis testing, and the information so gathered may be used to gain successes in a manner that does not interfere with the hypothesis testing.

II. History of the Problem

The TABP was introduced by Robbins [1] in 1952. In that paper there was no constraint on memory and the experiments were restricted to be binary-valued (coin tosses). Robbins argued that a scheme that sampled the "inferior" coin infinitely often, but with density of sampling tending to zero, yielded $r = 1$. Here, at a particular time, the "inferior" coin is defined to be the coin yielding the lower cumulative proportion of heads. Subsequently, Bradt, Johnson, and Karlin [2] and Bradt and Karlin [3] examined generalizations of the TABP in which it was desired to maximize the number of successes in a finite number of trials. This problem remains open in the case where the coin biases (p_1, p_2) have an arbitrary known joint distribution. However, Feldman [4] has solved the generalized version of the TABP corresponding to (1) (with known a priori probabilities) in the infinite-memory case.

The idea of adding a finite-memory constraint is due to Robbins [5]. Robbins defines memory to be of length k if the choice of coin at each trial is allowed to depend only on the outcomes of the k previous trials. Letting $\mathfrak{X} = \{A, B\} \times \mathcal{Y}$ denote the observation space, the problem becomes one of determining the function e: $\mathfrak{X}^k \rightarrow \{A, B\}$ for which the algorithm

$$e_{n+1} = e(X_n, X_{n-1}, \cdots, X_{n-k+1}) \tag{6}$$

$$X_n = (e_n, Y_n) \tag{7}$$

maximizes r. Since $\mathfrak{X}$ has but four members in Robbins's problem, memory is still finite according to the definition of Section I, with $m = 4^k$. However, if the experimental outcome space $\mathcal{Y}$ is infinite, an infinite-state memory is needed to recall the last k experimental outcomes.

Although Robbins's original algorithm has been successively improved by Isbell [6], Smith and Pyke [7], and Samuels [8], an optimal scheme has not been established. However, if the choice of coin may also depend on time, the problem has been solved by Cover [9]. A memory $k = 2$ is sufficient, i.e., there exists an algorithm e for which the scheme

$$e_{n+1} = e(X_n, X_{n-1}, n) \tag{8}$$

achieves an asymptotic proportion of successes $r = 1$. The algorithm is independent of the biases p_1 and p_2 on the two coins, and thus is optimal (achieves $r = 1$) for the more general problem of maximizing the asymptotic proportion of heads with two coins having arbitrary unknown biases. This work also implies that, with the definition of memory given in Section I, a memory of $m = 4$ states is sufficient [10] for a time-varying algorithm to achieve $r = 1$.

A series of publications following the work of Tsetlin has appeared in the Russian literature [11]–[21] on the behavior of automata in random media in an attempt to model adaptive or learning systems. In many cases the algorithms considered are similar to the TABP with finite memory of the type defined in Section I. A series of *ad hoc* "expedient" automata (i.e., automata that perform better than simply alternating coins at each trial) is examined, but no optimal automata are found. Subsequent work by Fu and Li [22], [23] and Chandrasekaran and Shen [24]–[27] has enlarged the set of algorithms for which the asymptotic behavior has been found. The fundamental problem implicit in [11]–[27] is presented in Section I and solved in this paper. It should be mentioned that the motivation of the previous papers is different from ours in the respect that previous work centered on modeling learning processes by finite-state automata. For this reason, the number of states m was frequently allowed to tend to infinity in the analysis, and the emphasis on optimal m-state automata was lost.

Note this one word of caution. Memory size has been defined to be the number of states of the automaton. This seems to us to be natural. However, we have not included any measure of the complexity of the computation of the state transition function f and the choice of experiment function e. Fortunately, the optimal function f is rather simple to implement, as can be seen from the example at the end of Section IV. Moreover, if an auxiliary stream of random variables is available, the calculation of f and e may be performed by "hard-wired" circuitry without memory elements.

III. Finite-State Machines

The two-armed-bandit problem that will be solved in this section has the form

$$\begin{array}{lll} & \text{experiment } A & \text{experiment } B \\ H_0 : & Y \sim \mathcal{P}_0 & Y \sim \mathcal{P}_1 \\ H_1 : & Y \sim \mathcal{P}_1 & Y \sim \mathcal{P}_0 \end{array} \tag{9}$$

where $\mathcal{P}_0$ and $\mathcal{P}_1$ are arbitrary known probability measures. Thus Y is not restricted to be a binary-valued random variable as in previous work [1], [5]. In Section VI, the solution will be generalized to the form

$$\begin{array}{lll} & A & B \\ H_0 : & Y \sim \mathcal{P}_0 & Y \sim \mathcal{P}_1 \\ H_1 : & Y \sim \mathcal{P}_2 & Y \sim \mathcal{P}_3. \end{array} \tag{10}$$

Attention will be restricted to the algorithm

$$T_n = f(T_{n-1}, X_n) \qquad T_n \in \{1, 2, \cdots, m\} \tag{11}$$

$$e_n = e(T_{n-1}) \qquad e_n \in \{A, B\} \tag{12}$$

$$X_n = (e_n, Y_n) \tag{13}$$

where T is the state of memory, e the choice of experiment, bnd Y the resulting observation. A reformulation of this algorithm in the terminology of finite-state machines will ae convenient. X and Y will denote random variables, and x and y their outcomes.

Consider a finite-state stochastic sequential machine with state space $\mathfrak{T} = \{1, 2, \cdots, m\}$, input space $\mathfrak{X} = \{A, B\} \times \mathfrak{Y}$ and output space $\{A, B\}$. Let the state transition behavior of this machine be specified by a family of $m \times m$ stochastic matrices $[p_{ij}(x)]$, defined for $x = (e, y)$, $e \in \{A, B\}$, $y \in \mathfrak{Y}$, and $i, j \in \{1, 2, \cdots, m\}$. Then

$$p_{ij}(e, y) = \Pr\{T_n = j \mid T_{n-1} = i, X_n = (e, y)\} \tag{14}$$

is the conditional probability of transition from memory state i to j under the observation of experiment e with outcome y.

Let the output function be described by the sequence α_i, $0 \leq \alpha_i \leq 1$, $i = 1, 2, \cdots, m$, with the interpretation that

$$\alpha_i = \Pr\{e_{n+1} = A \mid T_n = i\}. \tag{15}$$

Thus, the next experiment chosen is a random variable depending solely on the past experience as summarized in the current state of memory T_n. The automaton is depicted in Fig. 1.

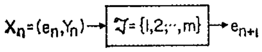


Fig. 1. Decision process viewed as a finite state automaton.

The state transition matrices conditioned on H_0 and H_1 are given by

$$p^0_{ij} = \underset{x}{E}\{p_{ij}(x) \mid H_0\} \tag{16}$$

and

$$p^1_{ij} = \underset{x}{E}\{p_{ij}(x) \mid H_1\}. \tag{17}$$

As will be shown in the proof of Theorem 1, these expectations may be explicitly expressed as follows:

$$p^0_{ij} = \int (p_{ij}(A, y)\alpha_i f_0(y) + p_{ij}(B, y)(1 - \alpha_i) f_1(y))\, d\nu\, (y) \tag{18}$$

$$p^1_{ij} = \int (p_{ij}(A, y)\alpha_i f_1(y) + p_{ij}(B, y)(1 - \alpha_i) f_0(y))\, d\nu\, (y) \tag{19}$$

where f_0 and f_1 are the Radon–Nikodym derivatives (densities) of $\mathcal{P}_0$ and $\mathcal{P}_1$ with respect to some dominating measure ν. Define the $m \times m$ matrices $\boldsymbol{P}^0 = [p^0_{ij}]$ and $\boldsymbol{P}^1 = [p^1_{ij}]$ and let $\boldsymbol{\mu}^0$ and $\boldsymbol{\mu}^1$ be the stationary probability distributions on the state space $\mathfrak{T}$ under H_0 and H_1. The stationary probability distributions are solutions of the matrix equations

$$\boldsymbol{\mu}^0 = \boldsymbol{\mu}^0 \boldsymbol{P}^0 \tag{20}$$

$$\boldsymbol{\mu}^1 = \boldsymbol{\mu}^1 \boldsymbol{P}^1. \tag{21}$$

Note that if $\boldsymbol{P}^k$ is irreducible, μ^k_i is the asymptotic proportion of time spent in state i, conditioned on H_k. Parallel work on hypothesis testing with finite memory [28] establishes that irreducible automata are at least "one state better" than reducible automata. The same argument applies to the current formulation of the TABP so that, here too, attention will be restricted to irreducible automata.

Letting r_0 and r_1 be the asymptotic proportion of successes under H_0 and H_1, it is seen that

$$r_0 = \sum_{i=1}^{m} \mu^0_i \alpha_i \tag{22}$$

$$r_1 = \sum_{i=1}^{m} \mu^1_i (1 - \alpha_i) \tag{23}$$

where the α_i are defined by (15).

If a Bayesian approach is taken and a priori probabilities π_0 and π_1 ($\pi_0 + \pi_1 = 1$) are assigned to H_0 and H_1 then

$$r = \pi_0 r_0 + \pi_1 r_1. \tag{24}$$

Although the Bayesian approach will be taken, the results will apply to the Neyman–Pearson formulation as well. In the Neyman–Pearson formulation the problem is to maximize r_1 subject to the constraint $r_0 \leq \alpha$, for a given level α.

Returning to the Bayesian formulation, the goal is to maximize r over all $p_{ij}(x)$, and α_i; $i, j = 1, 2, \cdots, m$. Designate this maximum value of r by r^*.

In order to place an upper bound on r it is necessary to relate the parameters of the automaton to the statistics of the problem. The following definitions and theorems will prove useful.

Definitions

Let the measure $\nu = \mathcal{P}_0 + \mathcal{P}_1$. Thus $\mathcal{P}_0$ and $\mathcal{P}_1$ are both absolutely continuous with respect to ν. Define $f_0(y)$ and $f_1(y)$ to be the Radon–Nikodym derivatives of $\mathcal{P}_0$ and $\mathcal{P}_1$ with respect to ν (f_0 and f_1 are the probability density functions of $\mathcal{P}_0$ and $\mathcal{P}_1$ with respect to ν). Let

$$\begin{aligned} l_A(y) &= f_0(y)/f_1(y) \\ l_B(y) &= f_1(y)/f_0(y) = 1/l_A(y). \end{aligned} \tag{25}$$

It is seen that l_A and l_B are the likelihood ratios for an experimental outcome y that results from experiments A and B, respectively.

Further define (for $C \subseteq \mathcal{Y}$)

$$\bar{\bar{l}}_A = \sup_{\nu(C)>0} \left[\frac{\mathcal{P}_0(C)}{\mathcal{P}_1(C)}\right] \qquad \bar{l}_A = \inf_{\nu(C)>0} \left[\frac{\mathcal{P}_0(C)}{\mathcal{P}_1(C)}\right] \tag{26}$$

$$\bar{\bar{l}}_B = \sup_{\nu(C)>0} \left[\frac{\mathcal{P}_1(C)}{\mathcal{P}_0(C)}\right] \qquad \bar{l}_B = \inf_{\nu(C)>0} \left[\frac{\mathcal{P}_1(C)}{\mathcal{P}_0(C)}\right]. \tag{27}$$

Thus $\bar{\bar{l}}_A$ is the almost everywhere (a.e.) maximum likelihood ratio (l.r.) for experiment A; and $\bar{l}_A$ is the a.e. minimum l.r. for experiment A. Similarly, $\bar{\bar{l}}_B$ and $\bar{l}_B$ are the a.e. maximum and minimum l.r.'s for experiment B. Clearly, from the definitions, $\bar{\bar{l}}_A = 1/\bar{l}_B$ and $\bar{l}_A = 1/\bar{\bar{l}}_B$. Thus defining

$$\begin{aligned} \bar{\bar{l}} &= \max\{\bar{\bar{l}}_A, \bar{\bar{l}}_B\} \\ \bar{l} &= \min\{\bar{l}_A, \bar{l}_B\} \end{aligned} \tag{28}$$

it is seen that

$$\bar{\bar{l}} = \max\{\bar{\bar{l}}_A, 1/\bar{l}_A\} \tag{29}$$

$$\bar{l} = \min\{\bar{l}_A, 1/\bar{\bar{l}}_A\} \tag{30}$$

and

$$\bar{\bar{l}} = 1/\bar{l}. \tag{31}$$

Definition

The likelihood ratio $l(x)$ of an observation $x = (e, y)$, $e \in \{A, B\}$, is defined to be

$$l(x) = l_e(y). \tag{32}$$

Obviously, $\bar{l} \leq l(x) \leq \bar{\bar{l}}$.

For example, if two unlabeled coins of biases $p_0 = 0.7$ and $p_1 = 0.8$ are given, the possible events C are heads and tails, and

$$\bar{\bar{l}}_A = \max\left\{\frac{0.7}{0.8}, \frac{0.3}{0.2}\right\} = \frac{3}{2} \tag{33a}$$

$$\bar{\bar{l}}_B = \max\left\{\frac{0.8}{0.7}, \frac{0.2}{0.3}\right\} = \frac{8}{7} \tag{33b}$$

$$\bar{\bar{l}} = \max\{\tfrac{3}{2}, \tfrac{8}{7}\} = \tfrac{3}{2}. \tag{33c}$$

Since $l_A(T) = \bar{\bar{l}} = \frac{3}{2}$; $l_B(T) = \bar{l} = \frac{2}{3}$, the maximum and minimum likelihood ratio events are given by "tails on coin A" and "tails on coin B," respectively.

Theorem 1

For all $i, j \in \{1, 2, \cdots, m\}$,

$$1/\bar{\bar{l}} \leq p_{ij}^0/p_{ij}^1 \leq \bar{\bar{l}}. \tag{34}$$

Proof: From (16) it is seen that

$$p_{ij}^0 = \Pr\{T_n = j \mid T_{n-1} = i, H_0\}. \tag{35}$$

Equivalently

$$\begin{aligned} p_{ij}^0 = \Pr\{T_n = j \mid T_{n-1} = i, H_0, e_n = A\} \\ \cdot \Pr\{e_n = A \mid T_{n-1} = i, H_0\} \\ + \Pr\{T_n = j \mid T_{n-1} = i, H_0, e_n = B\} \\ \cdot \Pr\{e_n = B \mid T_{n-1} = i, H_0\}. \end{aligned} \tag{36}$$

But, since the choice of e_n is a (randomized) function of T_{n-1} alone,

$$\begin{aligned} \Pr\{e_n = A \mid T_{n-1} = i, H_0\} \\ &= \Pr\{e_n = A \mid T_{n-1} = i\} \\ &= \alpha_i. \end{aligned} \tag{37}$$

Similarly

$$\Pr\{e_n = B \mid T_{n-1} = i, H_0\} = 1 - \alpha_i. \tag{38}$$

From (14),

$$\begin{aligned} \Pr\{T_n = j \mid T_{n-1} = i, H_0, e_n = A\} \\ = \int p_{ij}(A, y) f_0(y)\, d\nu(y), \end{aligned} \tag{39}$$

since under H_0 the experimental outcome Y has f_0 as its density function when A is performed. Similarly,

$$\begin{aligned} \Pr\{T_n = j \mid T_{n-1} = i, H_0, e_n = B\} \\ = \int p_{ij}(B, y) f_1(y)\, d\nu(y). \end{aligned} \tag{40}$$

Then (36) becomes

$$\begin{aligned} p_{ij}^0 = \alpha_i \int p_{ij}(A, y) f_0(y)\, d\nu(y) \\ + (1 - \alpha_i) \int p_{ij}(B, y) f_1(y)\, d\nu(y). \end{aligned} \tag{41}$$

By definition

$$f_0(y) = l_A(y) f_1(y) \tag{42}$$

$$f_1(y) = l_B(y) f_0(y) \tag{43}$$

so that

$$p_{ii}^0 = \alpha_i \int p_{ij}(A, y) l_A(y) f_1(y)\, d\nu(y) + (1 - \alpha_i) \int p_{ij}(B, y) l_B(y) f_0(y)\, d\nu(y). \tag{44}$$

Furthermore, $l_A(y) \le \bar{\bar{l}}_A$ and $l_B(y) \le \bar{\bar{l}}_B$ a.e. ν, and by (29) $\bar{\bar{l}} = \max\{\bar{\bar{l}}_A, \bar{\bar{l}}_B\}$. Hence

$$p_{ii}^0 \le \bar{\bar{l}} \left[\alpha_i \int p_{ij}(A, y) f_1(y)\, d\nu(y) + (1 - \alpha_i) \int p_{ij}(B, y) f_0(y)\, d\nu(y) \right]. \tag{45}$$

Proceeding similarly it is found that

$$p_{ij}^1 = \alpha_i \int p_{ij}(A, y) f_1(y)\, d\nu(y) + (1 - \alpha_i) \int p_{ij}(B, y) f_0(y)\, d\nu(y). \tag{46}$$

Combining (45) and (46) yields

$$p_{ij}^0 / p_{ij}^1 \le \bar{\bar{l}} \tag{47}$$

thus proving half of the theorem. The other half follows in an analogous manner.

Definition

The state likelihood ratio vector $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_m)$ is defined by

$$\lambda_i = \mu_i^0 / \mu_i^1 \qquad i = 1, 2, \cdots, m. \tag{48}$$

Theorem 2

For an irreducible automaton in which the λ_i's are arranged in nondecreasing order the following relation holds:

$$\lambda_{i+1} / \lambda_i \le (\bar{\bar{l}})^2. \tag{49}$$

Remark

Since it has been noted that irreducible automata can do at least as well as reducible ones, the irreducibility restriction is of no consequence.

Proof: The proof of this theorem follows from Theorem 1 using arguments contained in Lemma 2 of [28]. The reader is referred there for details.

Theorem 3

For an m-state automaton r is bounded above by r^* where

$$r^* = \max\left\{ \frac{(\bar{\bar{l}})^{2(m-1)} - 2(\pi_0 \pi_1 (\bar{\bar{l}})^{2(m-1)})^{1/2}}{(\bar{\bar{l}})^{2(m-1)} - 1}, \pi_0, \pi_1 \right\}. \tag{50}$$

In the special case $\pi_0 = \pi_1 = \frac{1}{2}$,

$$r^* = \bar{\bar{l}}^{(m-1)} / (\bar{\bar{l}}^{(m-1)} + 1). \tag{51}$$

Remark 1

If $r^* = \pi_0$ (or π_1), a degenerate situation exists in which the machine that always chooses experiment A (or B) is optimal. In this case memory is not large enough to gather sufficient information to offset the a priori bias [28].

Remark 2

The larger the value of $\bar{\bar{l}}$, the larger the resultant proportion of successes r^*. Thus, $\bar{\bar{l}}$ is a measure of the separation between H_0 and H_1.

Example

Before proceeding with the proof of the theorem, an example will be given. Consider the coin-flipping TABP

	A	B
$\pi_0 = \frac{1}{2}$	$H_0 : p_0 = 0.9$	$p_1 = 0.8$
$\pi_1 = \frac{1}{2}$	$H_1 : p_1 = 0.8$	$p_0 = 0.9$

where p_0 and p_1 are the probabilities of the event heads (H) under the appropriate conditions. Thus, for example, if coin A is flipped and H_0 is true, then $\Pr\{\text{heads}\} = p_0 = 0.9$. Calculation shows that

$$\bar{\bar{l}} = \max\{\bar{\bar{l}}_A, \bar{\bar{l}}_B\} = \max\{\tfrac{9}{8}, \tfrac{8}{9}, \tfrac{1}{2}, \tfrac{2}{1}\} = 2. \tag{52}$$

Thus, for an m-state memory the best possible limiting proportion of uses of the "best" coin (in this case, coin A) is given by

$$r^* = \frac{2^{m-1}}{2^{m-1} + 1}. \tag{53}$$

In the next section an automaton will be exhibited that achieves r^* arbitrarily closely.

Example

If $p_0 = 0.5$, $p_1 = 0.501$, the situation is quite different. Here $\bar{\bar{l}} \cong 1.002$ and

$$r^* \cong (1.002)^{m-1} / ((1.002)^{m-1} + 1). \tag{54}$$

Thus, even $m = 500$ states yields only a proportion of successes $r^* \approx e/(e + 1)$. No 500-state machine can do better.

Proof: By Theorem 2, $\lambda_2 \le \lambda_1 (\bar{\bar{l}})^2$, $\lambda_3 \le \lambda_1 (\bar{\bar{l}})^4$, $\cdots$, $\lambda_m \le \lambda_1 (\bar{\bar{l}})^{2(m-1)}$. Thus, for all $i \in \mathfrak{J} = \{1, 2, \cdots, m\}$

$$\lambda_1 \le \mu_i^0 / \mu_i^1 \le \lambda_1 (\bar{\bar{l}})^{2(m-1)}. \tag{55}$$

Hence

$$r_0 = \sum_{i \in \mathfrak{J}} \mu_i^0 \alpha_i \le \lambda_1 (\bar{\bar{l}})^{2(m-1)} \sum_{i \in \mathfrak{J}} \mu_i^1 \alpha_i. \tag{56}$$

But

$$r_1 = \sum_{i \in \mathfrak{J}} \mu_i^1 (1 - \alpha_i) = 1 - \sum_{i \in \mathfrak{J}} \mu_i^1 \alpha_i. \tag{57}$$

So

$$r_0 \le \lambda_1 (\bar{\bar{l}})^{2(m-1)} (1 - r_1). \tag{58}$$

Similarly

$$r_1 \leq (1/\lambda_1)(1 - r_0). \tag{59}$$

Since r_0 and r_1 are nonnegative, multiplying (58) and (59) yields the fundamental inequality

$$r_0 r_1 \leq (\bar{\bar{l}})^{2(m-1)}(1 - r_0)(1 - r_1). \tag{60}$$

Now

$$r = \pi_0 r_0 + \pi_1 r_1 \tag{61}$$

achieves its maximum value when the weak inequality in (56) is met with equality. Thus,

$$r_0 r_1 = (\bar{\bar{l}})^{2(m-1)}(1 - r_0)(1 - r_1) \tag{62}$$

and the problem reduces to maximizing (61) subject to (62), a straightforward problem in the calculus of variations. The end result is that the maximum allowable value of r is r^* as given by (50), the desired result. The same type of problem arises in [28] and the reader is again referred there for details.

Note that the constraint equation (62) gives an upper bound for the operating characteristic of the automaton. That is, given r_0, the value of r_1 that satisfies (62) is the maximum possible. Thus, a bound is placed on the behavior of the automaton, in the sense of Neyman and Pearson.

This section has placed an upper bound on r. The next section will demonstrate a class of machines that can approach that bound.

IV. An ϵ-Optimal Class of Automata

An ϵ-optimal class of automata will now be demonstrated. That is for any $\epsilon > 0$ there will exist a machine in this class with $r \geq r^* - \epsilon$. Thus r^* can be approached as closely as desired.

Assume there exists an observation $\bar{\bar{x}} = (e_0, y_0)$ such that

$$l_{e_0}(y_0) = \bar{\bar{l}} = \max \{\bar{\bar{l}}_A, \bar{\bar{l}}_B\}. \tag{63}$$

Thus the observation of $\bar{\bar{x}}$ yields maximum information favoring H_0. Without loss of generality, assume $\bar{\bar{l}}_A > \bar{\bar{l}}_B$, which implies $e_0 = A$. Thus

$$l_A(y_0) = \bar{\bar{l}} \tag{64}$$

and, from (26), (64), and (31)

$$l_B(y_0) = 1/l_A(y_0) = 1/\bar{\bar{l}} = \bar{l}. \tag{65}$$

Hence $\bar{\bar{x}} = (A, y_0)$ and $\bar{x} = (B, y_0)$ are the maximum and minimum likelihood-ratio observations. The experimental outcome $Y = y_0$ yields maximum information for testing H_0 versus H_1, regardless of the coin flipped.

This is in distinction to the single-coin problem in which the observations $\bar{\bar{y}}$ and $\bar{y}$ achieving $\bar{\bar{l}}_A$ and $\bar{l}_A$ may be unrelated. The resulting "spread" is $\bar{\bar{l}}_A/\bar{l}_A$ for the one-armed bandit (see [28]) and is given by $\bar{\bar{l}}/\bar{l} = \max \{\bar{\bar{l}}_A^2, 1/\bar{l}_A^2\}$ for the TABP. Thus, two coins are better than one unless $\bar{\bar{l}}_A = 1/\bar{l}_A$.

Returning to the TABP, if A results in y_0 (the maximum-likelihood event) hypothesis H_0 is supported, whereas if B results in y_0 (the minimum-likelihood event) hypothesis H_1 is supported. For the moment assume that y_0 occurs with nonzero probability. That is, let

$$\Pr \{X = \bar{\bar{x}} \mid A, H_1\} = p > 0. \tag{66}$$

Then, by the symmetry of the hypotheses, and the definition of $\bar{\bar{x}}$,

$$\begin{aligned}\Pr \{X = \bar{x} \mid B, H_1\} &= \Pr \{X = \bar{\bar{x}} \mid A, H_0\} \\ &= \bar{\bar{l}} \Pr \{X = \bar{\bar{x}} \mid A, H_1\} \\ &= \bar{\bar{l}}p.\end{aligned} \tag{67}$$

Similarly,

$$\Pr \{X = \bar{x} \mid B, H_0\} = p. \tag{68}$$

Consider the m-state machine, which, in states 2 through $m - 1$, uses A and B with equal probability

$$\alpha_i = \tfrac{1}{2} \qquad i = 2, 3, 4, \cdots, m - 1 \tag{69}$$

and in states 1 and m uses A with probabilities

$$\alpha_1 = \delta/2 \quad \text{and} \quad \alpha_m = 1 - (k\,\delta/2), \tag{70}$$

where $\delta > 0$.

Furthermore, when the machine changes state, let it move at most one state at a time, moving to a higher state (from i to $i + 1$) only when $X = \bar{\bar{x}}$ is observed, and to a lower state (from i to $i - 1$) only when $X = \bar{x}$ is observed. For all other observations let the automaton remain in the same state. Also, if the automaton is in state 1 (or m) and $X = \bar{x}$ ($X = \bar{\bar{x}}$) is observed let the automaton stay in state 1 (m).

This machine is depicted in Fig. 2, where an arrow indicates an allowed transition, and the event over the arrow indicates the observations for which that transition occurs. The probability of using experiment A in state i is denoted by α_i. (For clarity the events over self-loops are omitted.)

To solve for $\mathbf{\mu}^0$ and $\mathbf{\mu}^1$, it is easiest to use the following method (see [29] for details). Partition S, the set of states in the automaton, into $C = \{1, 2, \cdots, i\}$ and $C' = \{i + 1, i + 2, \cdots, m\}$. In the steady state, the probability of a transition from C to C' must equal the probability of a transition from C' to C. But the only allowed transition from C to C' is from i to $i + 1$ and the only allowed transition from C' to C is from $i + 1$ to i. Thus, taking the case where H_0 is the true state of nature,

$$\begin{aligned}&\mu_i^0 \Pr \{X = \bar{\bar{x}} \mid \text{automaton is in state } i, H_0\} \\ &= \mu_{i+1}^0 \Pr \{X = \bar{x} \mid \text{automaton is in state } i + 1, H_0\}.\end{aligned} \tag{71}$$

Now $\bar{\bar{x}} = (A, y_0)$ and $\bar{x} = (B, y_0)$; so for $i = 1, 2, \cdots, m - 1$,

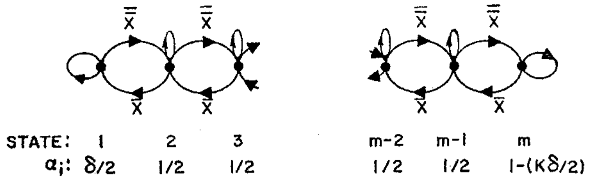


Fig. 2. The canonical form of the ϵ-optimal class of automata. $\bar{\bar{x}} = (A, y_0)$ is maximum likelihood ratio observation. $\bar{x} = (B, y_0)$ is minimum likelihood ratio observation.

$$\Pr\{X = \bar{\bar{x}} \mid \text{automaton is in state } i, H_0\}$$
$$= \alpha_i \Pr\{X = \bar{\bar{x}} \mid A, H_0\} = \alpha_i \bar{\bar{l}} p \qquad (72)$$

and

$$\Pr\{X = \bar{x} \mid \text{automaton is in state } i+1, H_0\}$$
$$= (1 - \alpha_{i+1}) \Pr\{X = \bar{x} \mid B, H_0\} = (1 - \alpha_{i+1})p. \qquad (73)$$

From (71), (72), and (73) it is seen that

$$\mu^0_{i+1} = \frac{\alpha_i \bar{\bar{l}}}{(1 - \alpha_{i+1})} \mu^0_i. \qquad (74)$$

Proceeding in an analogous manner

$$\mu^1_{i+1} = \frac{\alpha_i \bar{l}}{1 - \alpha_{i+1}} \mu^1_i \qquad (75)$$

is obtained. Using the expressions (69) and (70) for α_i and letting $\mu^0_i = a/\delta$, $\mu^1_i = b/\delta$ results in the following table.

State	1	2	3	$\cdots$	$m-2$	$m-1$	m
$\mu_i{}^0$	a/δ	$a(\bar{\bar{l}})$	$a(\bar{\bar{l}})^2$		$a(\bar{\bar{l}})^{(m-3)}$	$a(\bar{\bar{l}})^{(m-2)}$	$(a/k\delta)\ (\bar{\bar{l}})^{(m-1)}$
$\mu_i{}^1$	b/δ	$b(\bar{l})$	$b(\bar{l})^2$		$b(\bar{l})^{(m-3)}$	$b(\bar{l})^{(m-2)}$	$(b/k\delta)\ (\bar{l})^{(m-1)}$

The normalizing constants a and b are implicitly defined by

$$\sum \mu^0_i = 1, \qquad \sum \mu^1_i = 1. \qquad (76)$$

By inspection, the steady-state probability that the automaton is in state 1 or m approaches 1, as $\delta \to 0$, $\delta > 0$. That is, as $\delta \to 0$, $\mu^0_1 + \mu^0_m \to 1$ and $\mu^1_1 + \mu^1_m \to 1$, and for all $i \neq 1$ or m, $\mu^0_i \to 0$ and $\mu^1_i \to 0$. Further, $\alpha_1 \to 0$ and $\alpha_m \to 1$. Taking the limit as $\delta \to 0$, $\delta > 0$, results in

$$r_0 = \sum_{i=1}^{m} \mu^0_i \alpha_i = \mu^0_m \qquad (77)$$

and

$$r_1 = \sum_{i=1}^{m} \mu^1_i (1 - \alpha_i) = \mu^1_1. \qquad (78)$$

Therefore

$$r_0 r_1 = \mu^0_m \mu^1_1 = \frac{ab}{k\delta^2} (\bar{\bar{l}})^{m-1} \qquad (79)$$

and

$$(1 - r_0)(1 - r_1) = \mu^0_1 \mu^1_m = \frac{ab}{k\delta^2} (\bar{l})^{m-1}. \qquad (80)$$

Combining these last two equations and recalling that $\bar{\bar{l}} = 1/\bar{l}$ yields

$$r_0 r_1 = (1 - r_0)(1 - r_1)(\bar{\bar{l}})^{2(m-1)}. \qquad (81)$$

But (81) is just the constraint equation (62), which placed an upper bound on the values of r_0 and r_1. Thus the values of r_0 and r_1 achieved (in the limit as $\delta \to 0$) by this class of machines are the largest possible, and hence define the optimal operating characteristic. Any point on the optimal operating characteristic can be reached in the limit as $\delta \to 0$ for suitable choice of k. (The constraint that δ be nonzero is crucial; letting $\delta = 0$ results in vastly inferior performance.)

In the Neyman–Pearson formulation, k is set to that value for which r_0 equals its desired value. In the Bayesian formulation k is set to that value k^* that maximizes r. This is accomplished by minimizing

$$r = \pi_0 \mu^0_m + \pi_1 \mu^1_1 \qquad (82)$$

subject to (79), (80), and

$$\mu^0_1 + \mu^0_m = 1 \qquad (83)$$

$$\mu^1_1 + \mu^1_m = 1 \qquad (84)$$

yielding

$$k^* = \frac{(\bar{\bar{l}})^{m-1} \sqrt{\pi_0 \pi_1} - \pi_0}{(\bar{\bar{l}})^{m-1} \pi_0 - \sqrt{\pi_0 \pi_1}} \qquad (85)$$

as the optimum value of k, under the nondegeneracy condition $r^* > \max\{\pi_0, \pi_1\}$.

Remark

The standard case $\pi_0 = \pi_1 = \frac{1}{2}$ results in $k^* = 1$. Also note that this analysis was done under the assumption $\bar{\bar{l}}_A > \bar{\bar{l}}_B$. If $\bar{\bar{l}}_B > \bar{\bar{l}}_A$, merely exchange the roles of A and B as well as the roles of H_0 and H_1 to reduce to the previous case.

Thus, under the assumption that $\bar{\bar{x}}$ and $\bar{x}$ occur with nonzero probability, an ϵ-optimal class of automata has been shown to exist. Now examine the case where $\bar{\bar{x}}$ occurs with probability zero. By the definition of $\bar{\bar{l}}$, for ($\bar{\bar{l}} < \infty$), for any $\epsilon > 0$, there must exist a set of observations $\mathfrak{X}_1(\epsilon) \subseteq \mathfrak{X}$ such that $\mathfrak{X}_1(\epsilon)$ has nonzero probability measure and

$$l[\mathfrak{X}_1(\epsilon)] = \frac{\Pr\{\mathfrak{X}_1(\epsilon) \mid H_0\}}{\Pr\{\mathfrak{X}_1(\epsilon) \mid H_1\}} \geq \bar{\bar{l}} - \epsilon. \qquad (86)$$

Similarly, if $\bar{x}$ occurs with zero probability, for any $\epsilon > 0$, there exists a set $\mathfrak{X}_2(\epsilon) \subseteq \mathfrak{X}$ such that $\mathfrak{X}_2(\epsilon)$ has nonzero probability measure and

$$l[\mathfrak{X}_2(\epsilon)] \leq \bar{l} + \epsilon. \qquad (87)$$

Now replacing $\bar{\bar{x}}$ with $\mathfrak{X}_1(\epsilon)$ and $\bar{x}$ with $\mathfrak{X}_2(\epsilon)$ in the preceding development it is easily shown that the resultant r_0 and r_1 satisfy

$$r_0 r_1 \geq [\gamma(\epsilon)]^{m-1} (1 - r_0)(1 - r_1) \qquad (88)$$

where

$$\gamma(\epsilon) = l[\mathfrak{X}_1(\epsilon)]/l[\mathfrak{X}_2(\epsilon)]. \qquad (89)$$

Again, by varying k, the entire operating characteristic determined by (88) is generated. Furthermore as ϵ approaches zero, $\gamma(\epsilon)$ approaches $(\bar{\bar{l}})^2$. Therefore, for $k = k^*$, by letting both $\delta > 0$ and $\epsilon > 0$ approach zero, r approaches r^*. Thus an ϵ-optimal class of automata has been demonstrated for the general problem.

Example

Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be univariate normal distributions with variance $\sigma^2 = 1$ and means $\mu_0 = +1$ and $\mu_1 = -1$, respectively. We wish to maximize $\lim_{n\to\infty} (1/n) \sum_{i=1}^{n} Y_i$. This is equivalent to maximizing the limiting proportion of uses of the experiment corresponding to $\mathcal{P}_0$. We find that $l_A(y) = e^{4y}$. Thus $\bar{\bar{l}} = \infty$ and $r^* = 1$, for all $m \geq 2$. A two-state memory ϵ-achieving $r^* = 1$ can be defined as follows.

Move to state 2, under (B, y) if $y < -R$, and move to state 1, under (A, y) if $y < -R$. In state 1 use experiment B; in state 2 use experiment A. No randomization is required in this example. Here the proportion of successes can be made arbitrarily close to 1 by choosing R sufficiently large.

V. Composite Hypotheses

In this section it will be shown that the ϵ-optimal m-state machine, for the problem in which two coins of unknown bias are presented, is almost independent of the exact biases of the coins.

Consider the problem in which two coins A and B are available. In this case Y assumes one of only two values: H(heads) and T(tails). Two hypotheses exist concerning p_A and p_B, the respective biases of the coins toward heads:

$$\begin{array}{llll} & & \text{coin } A & \text{coin } B \\ \pi_0 = \frac{1}{2} & H_0: & p_A = p_1 & p_B = p_2 \\ \pi_1 = \frac{1}{2} & H_1: & p_A = p_2 & p_B = p_1. \end{array} \tag{90}$$

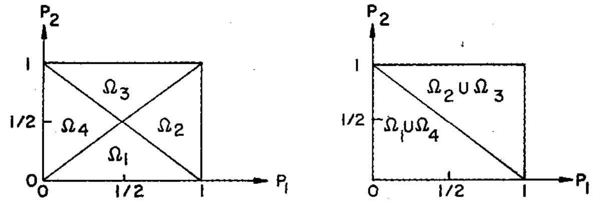
It is desired to maximize the long-run proportion of heads achieved. This is equivalent to the compound hypothesis test

$$\begin{aligned} H_0' &: p_A > p_B \\ H_1' &: p_A < p_B \end{aligned} \tag{91}$$

followed by the utilization of the coin deemed to have the highest probability of heads. The m-state machine achieving the highest limiting proportion of heads will be designated as the optimal machine. Let (p_1, p_2) be constrained to be in the region Ω_1 depicted in Fig. 3(a). Under this constraint, H_0 and H_1 are composite hypotheses. However, since $\pi_0 = \pi_1 = \frac{1}{2}$, the optimum value of k is $k^* = 1$ for all $(p_1, p_2) \in \Omega_1$. Furthermore, since in this region $p_1 > p_2$

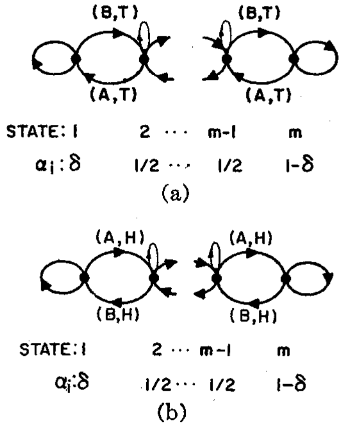
$$\begin{aligned} l_A(H) &= p_1/p_2 = \bar{\bar{l}}_A \qquad & l_A(T) &= q_1/q_2 = \bar{l}_A \\ l_B(H) &= p_2/p_1 = \bar{l}_B \qquad & l_B(T) &= q_2/q_1 = \bar{\bar{l}}_B \end{aligned} \tag{92}$$

and since in Ω_1 it is also true that $p_1q_1 > p_2q_2$, it follows that $\bar{\bar{l}}_A > \bar{\bar{l}}_B$. Therefore, $\bar{\bar{l}} = \bar{\bar{l}}_A$. Thus $\bar{\bar{x}} = (A, H)$ and $\bar{x} = (B, H)$ are the maximum-information events. This is true for all $(p_1, p_2) \in \Omega_1$. The ϵ-optimal m-state machine flips coin A with probability $\alpha_1 = \delta$ in state 1, with probability $\alpha_m = 1 - \delta$ in state m, and with probability $\alpha_i = \frac{1}{2}$ in states $i = 2, 3, \cdots, m - 1$, where δ is a small positive real number.



(a) (b)

Fig. 3. Equivalence regions for the compound problem.



(a)

(b)

Fig. 4. (a) ϵ-optimal automaton for $p_1 + p_2 \geq 1$. (b) ϵ-optimal automaton for $p_1 + p_2 \leq 1$.

It can also be shown that the regions Ω_2, Ω_3, and Ω_4 of Fig. 3(a) each have but one class of ϵ-optimal machines. Thus, it is only necessary to know in which region (p_1, p_2) lies to design an ϵ-optimal class of machines. The values of $\bar{\bar{x}}$, $\bar{x}$, α_1, and α_m for each region are summarized below.

$$\begin{array}{llll} \Omega_1: \bar{\bar{x}} = (A, H) & \bar{x} = (B, H), \alpha_1 = \delta, & \alpha_m = 1 - \delta \\ \Omega_2: \bar{\bar{x}} = (B, T) & \bar{x} = (A, T), \alpha_1 = \delta, & \alpha_m = 1 - \delta \\ \Omega_3: \bar{\bar{x}} = (A, T) & \bar{x} = (B, T), \alpha_1 = 1 - \delta, & \alpha_m = \delta \\ \Omega_4: \bar{\bar{x}} = (B, H) & \bar{x} = (A, H), \alpha_1 = 1 - \delta, & \alpha_m = \delta. \end{array} \tag{93}$$

It is readily verified by indexing the states in reverse order that the ϵ-optimal m-state machine for $(p_1, p_2) \in \Omega_4$ is identical to that for $(p_1, p_2) \in \Omega_1$. Similarly, the ϵ-optimal machine for Ω_2 and Ω_3 are identical. These equivalence regions are shown in Fig. 3(b).

Therefore, only two machines are needed for all (p_1, p_2). Equivalently, it is only necessary to know whether $p_1 + p_2 > 1$ or $p_1 + p_2 < 1$ in order to design an ϵ-optimal machine for this problem.

The design of the optimal machine does not depend on exact knowledge of p_1 and p_2. It is only necessary to know whether heads (H) or tails (T) is the maximal-information observation. If $p_1 + p_2 \geq 1$, then tails yields maximum information; if $p_1 + p_2 \leq 1$, then heads yields maximum information. The corresponding optimal machines are shown in Fig. 4. Note that if $p_1 = 1 - p_2$, i.e., the bias of one coin is the complement of the bias of the other, then either machine is optimal.

VI. Generalization to an Infinite Family of Experiments

In the foregoing, the experimenter had two experiments (coins A and B) and two hypotheses concerning the distributions governing the outcome Y under each experiment. In this section an outline will be provided of an ϵ-optimal m-state memory solution under the generalization that the experimenter has at his command an arbitrary known class of experiments Θ.

Let $\{\mathcal{P}_\theta^{(0)}\}$, $\{\mathcal{P}_\theta^{(1)}\}$ be a collection of probability measures indexed by $\theta \in \Theta$.

Consider the two hypotheses

$$\begin{aligned} &H_0 : Y \sim \mathcal{P}_\theta^{(0)}, \text{ if } \theta \text{ is the experiment performed} \\ &H_1 : Y \sim \mathcal{P}_\theta^{(1)}, \text{ if } \theta \text{ is the experiment performed} \end{aligned} \tag{94}$$

where the experiment $\theta \in \Theta$ may be freely chosen. Suppose that it is desired to maximize

$$E\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} J(Y_i)\right\}, \tag{95}$$

where the expectation is taken over H_0 and H_1 (with probabilities π_0 and π_1), where $Y_1 \sim \mathcal{P}_{\theta_i}^{(H)}$ and J is a real valued function. The two-coin problem of Section I, for example, is a special case of this problem, obtained by setting $J(Y_i) = 1$ or 0 accordingly as $Y_i = H$ or T, and letting $\Theta = \{A, B\}$, $p_A^0 = p_B^1 = p_0$, $p_B^0 = p_A^1 = p_1$. The finite-memory algorithms to be considered are of the form

$$\begin{aligned} T_n &= f(T_{n-1}, X_n) \\ \theta_n &= \theta(T_{n-1}) \\ X_n &= (\theta_n, Y_n) \end{aligned} \tag{96}$$

where $T_n \in \{1, 2, \cdots, m\}$ is the state of the m-state memory at time n; θ: $\{1, 2, \cdots, m\} \to \Theta$ is the choice of experiment as a function of the state, and $X_n = (\theta_n, Y_n)$ is the observation at time n. As before, the transition function f and decision function θ may be stochastic.

The usual conflict exists—the best experiments for the resolution of the hypotheses may not be the best for the maximization of $E[J(Y)]$. However, the form of the ϵ-optimal algorithm derived in Section IV suggests that the conflicting problems may be treated separately.

Let $\mathcal{A}_\theta$ denote the collection of subsets $A \subseteq \mathcal{Y}$ for which $\mathcal{P}_\theta^{(0)}(A) + \mathcal{P}_\theta^{(1)}(A) > 0$. Define

$$\bar{\bar{l}} = \sup_{\theta\in\Theta} \sup_{A\in\mathcal{A}_\theta} \left[\frac{\mathcal{P}_\theta^{(0)}(A)}{\mathcal{P}_\theta^{(1)}(A)}\right], \tag{97}$$

$$\bar{l} = \inf_{\theta\in\Theta} \inf_{A\in\mathcal{A}_\theta} \left[\frac{\mathcal{P}_\theta^{(0)}(A)}{\mathcal{P}_\theta^{(1)}(A)}\right], \tag{98}$$

and $\gamma = \bar{\bar{l}}/\bar{l}$. As a consequence of the work in [28], it can be shown that the probabilities of error (α and β) under each hypothesis (H_0 and H_1, respectively) must satisfy

$$\gamma^{m-1}\alpha\beta \geq (1 - \alpha)(1 - \beta) \tag{99}$$

for any m-state algorithm.

Let $\bar{\bar{\theta}}^*$ and $\bar{\theta}^*$ be experiments that ϵ-achieve $\bar{\bar{l}}$ and $\bar{l}$, respectively. That is (in the case $\bar{\bar{l}} < \infty$),

$$\inf_{A\in\mathcal{A}_{\bar{\theta}^*}} \frac{\mathcal{P}_{\bar{\theta}^*}^{(0)}(A)}{\mathcal{P}_{\bar{\theta}^*}^{(1)}(A)} \leq \bar{l} + \epsilon \tag{100}$$

$$\sup_{A\in\mathcal{A}_{\bar{\bar{\theta}}^*}} \frac{\mathcal{P}_{\bar{\bar{\theta}}^*}^{(0)}(A)}{\mathcal{P}_{\bar{\bar{\theta}}^*}^{(1)}(A)} \geq \bar{\bar{l}} - \epsilon. \tag{101}$$

Now define

$$J_0(\theta) = \int J(y)\, d\mathcal{P}_\theta^{(0)}(y) \tag{102}$$

$$J_1(\theta) = \int J(y)\, d\mathcal{P}_\theta^{(1)}(y) \tag{103}$$

as the expected payoffs under each hypothesis for the experiment θ.

If $\Pr\{H_0\} = t$, the use of experiment θ would incur expected payoff $tJ_0(\theta) + (1 - t)J_1(\theta)$. Let the maximum expected payoff over Θ be defined by

$$J(t) = \sup_{\theta\in\Theta} \{tJ_0(\theta) + (1 - t)J_1(\theta)\} \qquad 0 \leq t \leq 1. \tag{104}$$

(Note that $J(t)$, being the supremum of a collection of convex functions of t, is therefore a convex function of t.) Let $\theta^*(t)$ denote an experiment ϵ-achieving $J(t)$. That is,

$$tJ_0(\theta^*(t)) + (1 - t)J_1(\theta^*(t)) \geq J(t) - \epsilon. \tag{105}$$

Suppose next that an automaton with δ-traps is used in which all but ϵ of the probability is concentrated in states 1 and m. Since the probabilities of error of each kind are α and β,

$$\Pr\{H_0 \mid \text{state } 1\} = \frac{\pi_0\alpha}{\pi_0\alpha + \pi_1(1 - \beta)} = t_1 \tag{106}$$

$$\Pr\{H_0 \mid \text{state } m\} = \frac{\pi_0(1 - \alpha)}{\pi_0(1 - \alpha) + \pi_1\beta} = t_m. \tag{107}$$

Using the optimal experiments $\theta_1^* = \theta^*(t_1)$ and $\theta_m^* = \theta^*(t_m)$ in states 1 and m results in an expected payoff

$$\begin{aligned} \tilde{J}(\alpha, \beta) = {} & (\pi_0(1 - \alpha) + \pi_1\beta) \\ & \cdot J(\pi_0(1 - \alpha)/(\pi_0(1 - \alpha) + \pi_1\beta)) \\ & + (\pi_0\alpha + \pi_1(1 - \beta))J(\pi_0\alpha/(\pi_0\alpha + \pi_1(1 - \beta))). \end{aligned} \tag{108}$$

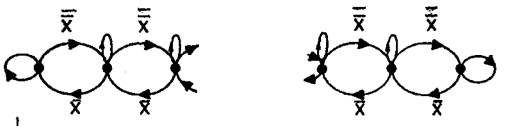
Finally, let

$$J^* = \sup_{(\alpha,\beta)} \tilde{J}(\alpha, \beta) \tag{109}$$

where the supremum is taken over all α, β satisfying (99) and $\alpha, \beta \in [0, 1]$.

We now establish that J^* is the least upper bound on

$$E\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} J(Y_i)\right\}$$



STATE	1	2	3	· · · · · ·	m-2	m-1	m
$\Pr\{\bar{\bar{\theta}}^*\|i\}$	$\delta/2$	1/2	1/2		1/2	1/2	0
$\Pr\{\bar{\theta}^*\|i\}$	0	1/2	1/2		1/2	1/2	$K^*\delta/2$
$\Pr\{\theta_1^*\|i\}$	$(1-\delta/2)$	0	0		0	0	0
$\Pr\{\theta_m^*\|i\}$	0	0	0		0	0	$1-(K^*\delta/2)$

$\bar{\bar{x}}$ ϵ-ACHIEVES $\bar{\bar{\ell}}$; $\bar{\bar{x}} = (\bar{\bar{\theta}}^*, y_1)$

$\bar{x}$ ϵ-ACHIEVES $\bar{\ell}$; $\bar{x} = (\bar{\theta}^*, y_2)$

Fig. 5. The automaton that ϵ-achieves J^*.

achievable by an m-state automaton. An automaton that ϵ-achieves J^* (for $\delta \approx 0$) is specified in Fig. 5, where k^* is chosen to yield the α, β achieving J^* in (109). Again, the self-loop events have been deleted. Note that only four experiments in Θ need be utilized: the experiments $\bar{\theta}^*$ and $\bar{\bar{\theta}}^*$, which yield most efficient resolution of the hypotheses; and θ_1^* and θ_m^*, which yield optimal payoffs. States 2, 3, $\cdots$, $m - 1$ are the bookkeeping states in which $\bar{\bar{\theta}}^*$ is used with probability one half and $\bar{\theta}^*$ is used with probability one half. For $\delta \approx 0$, the automaton is in states 1 or m with probability ≈ 1. Thus the maximal payoff experiments θ_1^* and θ_m^* are performed with probability arbitrarily close to one.

Note that if $\Theta = \{A, B\}$, the problem specified by (10) is solved. The generalization to more than two hypotheses does not appear easy.

VII. Conclusions

Inspection of the solution of the TABP indicates that optimal finite-memory learning is reasonably far from human intuition and practice. However, the heuristics garnered from inspection of the solution are easily assimilated for future application. Some interesting properties of the solution are the following.

1) The solution is only ϵ-optimal. In general, optimal solutions do not exist.

2) Artificial randomization is required in order to select the experiment to be performed at each stage. The state transition function f is deterministic. This differs from what we might call the one-armed-bandit problem [29] in which the experiment to be performed at each stage is unchanged, but for which the ϵ-optimal state transition function f involves randomization.

3) The conflict between data gathering and success gathering factors out. If we were given two experiments and were interested solely in hypothesis testing, and ignored the goal of using the best experiment a large proportion of the time, we could do no better than the probability of error currently obtained as reflected in r^*. This conflict does not generally disappear in the infinite-memory version of the TABP, with the exception of the symmetric version considered by Feldman [4].

4) Two coins are better than one. If the experimenter has in his possession two coins (with scrambled labels), he can always achieve a lower probability of error of determining which coin is which (with finite memory) than if he were provided only one of the coins. The probabilities of error are equal only if the coin biases are complementary.

5) The interior states can be considered to be allotted to hypothesis testing—the terminal states to utilizing this information to maximize the probability of success.

6) Transitions up and down are made only on maximal information events, and then only one step at a time. This is the conservatism that the finite-memory constraint demands. The price is the long waiting times between transitions. The TABP with finite memory given a finite sequence of observations has not been solved.

References

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Am. Math. Soc.*, vol. 58, pp. 529–532, 1952.

[2] R. N. Bradt, S. M. Johnson, and S. Karlin, "On sequential designs for maximizing the sum of n observations," *Ann. Math. Stat.*, vol. 27, pp. 1060–1074, 1956.

[3] R. N. Bradt and S. Karlin, "On the design and comparison of certain dichotomous experiments," *Ann. Math. Stat.*, vol. 27, pp. 390–409, 1956.

[4] D. Feldman, "Contributions to the 'two-armed bandit' problem," *Ann. Math. Stat.*, vol. 33, pp. 817–856, 1962.

[5] H. Robbins, "A sequential decision problem with a finite memory," *Proc. 1956 Natl. Acad. Sci.*, vol. 42, pp. 920–933.

[6] J. R. Isbell, "On a problem of Robbins," *Ann. Math. Stat.*, vol. 30, pp. 606–610, 1959.

[7] C. V. Smith and R. Pyke, "The Robbins-Isbell two-armed bandit problem with finite memory," *Ann. Math. Stat.*, vol. 36, pp. 1375–1386, 1965.

[8] S. M. Samuels, "Randomized rules for the two-armed bandit with finite memory," *Ann. Math. Stat.*, vol. 39, no. 6, pp. 2103–2107, 1968.

[9] T. M. Cover, "A note on the two-armed bandit problem with finite memory," *Inform. and Control*, vol. 12, pp. 371–377, May–June 1968.

[10] ——, "Hypothesis testing with finite statistics," *Ann. Math. Stat.*, vol. 40, pp. 828–835, June 1969.

[11] M. L. Tsetlin, "Certain problems in the behavior of finite automata," *Soviet Phys. Doklady*, vol. 6, no. 8, pp. 670–673, 1961.

[12] ——, "On the behavior of finite automata in random media," *Avtomatica i Telemekhanika*, vol. 22, pp. 1345–1354, October 1961 (available in transl.).

[13] ——, "A game between a finite automaton and an opponent using a mixed strategy," *Soviet Phys. Doklady*, vol. 8, no. 3, pp. 230–231, 1963.

[14] V. Y. Krylov, "On one automaton that is asymptotically optimal in a random medium," *Avtomatica i Telemekhanika*, vol. 24, pp. 1226–1228, September 1963 (available in transl.).

[15] V. Yu Krylov and M. L. Tsetlin, "Examples of games with automata," *Soviet Phys. Doklady*, vol. 8, no. 3, pp. 232–234, 1963.

[16] V. I. Varshavskii and I. P. Vorontsova, "On the behavior of stochastic automata with a variable structure," *Avtomatica and Telemekhanika*, vol. 24, March 1963 (available in transl.).

[17] V. L. Stefanyuk, "Example of a problem in the joint behavior of two automata," *Avtomatika i Telemekhanika*, vol. 24, no. 6, pp. 716–719, 1963 (transl. in *Automation and Remote Control*).

[18] I. M. Gel'fand, I. I. Pyatetskii-Shapiro, and M. L. Tsetlin, "Certain classes of games and automata games," *Soviet Phys. Doklady*, vol. 8, no. 10, pp. 964–966, 1963.

[19] S. L. Ginsburg, V. Yu Krylov, and M. L. Tsetlin, "One example of a game for many identical automata," *Avtomatika i Telemekhanika*, vol. 25, no. 5, 1964 (transl. in *Automation and Remote Control*, vol. 25, no. 5, pp. 608–611, 1964).

[20] D. I. Kalinin and I. M. Rotvain, "Some asymptotic estimates for games of automata in distribution," *Avtomatika i Tele-*

mekhanika, vol. 27, no. 4, 1966 (transl. in *Automation and Remote Control*, vol. 27, no. 4, pp. 642–644, 1966).

[21] N. P. Kandelaki and G. N. Tsertsvadze, "Behavior of certain classes of stochastic automata in random media," *Avtomatika i Telemekhanika*, vol. 27, June 1966 (available in transl.).

[22] K. S. Fu and T. J. Li, "On the behavior of learning automata and its applications," Purdue University, Lafayette, Ind., Tech. Rept. TR-EE 68-20, 1968.

[23] ——, "Formulation of learning automata and automata games," *Information Sci.*, vol. 1, pp. 237–256, 1969.

[24] B. Chandrasekaran, "Contributions to the theory of learning automata," Ph.D. dissertation, University of Pennsylvania, Philadelphia, May 1967.

[25] B. Chandrasekaran and D. W. C. Shen, "Adaption of stochastic automata in nonstationary environments," *Proc. 1967 Natl. Electronics Conf., 6th Symp. on Discrete Adaptive Processes*, (Chicago, Ill.), October 1967.

[26] ——, "On expediency and convergence in variable structure automata," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-4, pp. 52–60, March 1968.

[27] ——, "Stochastic automata games," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-5, pp. 145–149, April 1969.

[28] M. E. Hellman and T. M. Cover, "Learning with finite memory" (to be published in *Ann. Math. Stat.*, 1970).

[29] M. E. Hellman, "A faster method of solution for random walk problems," IBM Research, RC 2389, no. 11615.

Signals That Can Be Calculated from Their Ambiguity Function

RUDOLF DE BUDA

***Abstract*—A new lemma relates the analytic extensions of two time functions $u(t)$ and $v(t)$ to the Laplace transform of their ambiguity function ψ_{uv}. This lemma is used to derive necessary conditions for u and v from two bounds on the behavior of ψ_{uv} at infinity. In particular, if the first bound is fulfilled, then $u(z)$ and $v(z)$ must be integral analytic functions. If both bounds are fulfilled, then u and v are each equal to $\exp\{-\pi t^2\}$ times a polynomial in t, and the two polynomials can be found from ψ_{uv} by comparing coefficients.**

Introduction

Although one can easily calculate an ambiguity function from two given time functions, the converse problem of finding the time functions from the given ambiguity function remains as yet unsolved. This paper will study a class of ambiguity functions that fulfill certain bounds. From these bounds, some important properties of the corresponding time functions will be derived.

Definition of the Ambiguity Function $\psi_{uv}(\tau, \phi)$

For the Fourier pair and for the time-frequency correlation function $\chi(\tau, \phi)$, we follow Woodward's notation [1]. We define the ambiguity function as the squared magnitude of χ, and introduce for it the symbol $\psi_{uv}(\tau, \phi)$. τ and ϕ are the time- and frequency-shift variables, and the subscripts u and v indicate the complex signal and filter functions $u(t)$ and $v(t)$, which we assume to be square integrable and normalized to unit energy:

$$\int_{-\infty}^{\infty} |u(t)|^2\, dt = \int_{-\infty}^{\infty} |v(t)|^2\, dt = \int_{-\infty}^{\infty} |U(f)|^2\, df = 1. \tag{1}$$

u and v combine to form the ambiguity function $\psi_{uv}(\tau, \phi)$ by

$$\psi_{uv}(\tau, \phi) = \left| \int_{-\infty}^{\infty} u(t) v^*(t + \tau) \exp\{-2\pi j \phi t\}\, dt \right|^2 \cdot \tag{2}$$

When we want to distinguish specifically between ψ_{uu} and ψ_{uv}, $u \neq v$ [2], we shall use "auto" or "cross" as a prefix to the ambiguity function.

Results

If we bound the ambiguity function $\psi_{uv}(\tau, \phi)$ by

$$\psi_{uv}(\tau, \phi) \le \exp\{-\pi k(\tau^2 + \phi^2)\}, \qquad 0 < k \le 1, \tag{3}$$

then the time functions $u(t)$ and $v(t)$

1) are bounded by

$$|u(t)|^2 < A \exp(-\pi k t^2),$$
$$|v(t)|^2 < A \exp(-\pi k t^2),$$

2) are differentiable infinitely often,
3) have analytic extensions that are integral analytic functions.

If we further bound the ambiguity function by

$$\psi_{uv}(\tau, \phi) \le P_{2n}(\tau, \phi) \exp\{-\pi(\tau^2 + \phi^2)\} \tag{4}$$

where P_{2n} is a polynomial of order $2n$, then u and v have the form $\exp(-\pi t^2)$ times a polynomial of t. The polynomials for u and v have combined degree n; they can

Manuscript received February 19, 1968; revised September 10, 1969. This work was performed for the Defense Research Board, under Contract DIR E-18.

The author is with the Canadian General Electric Co., Ltd., Toronto, Ont., Canada.