THE ROBBINS-ISBELL TWO-ARMED-BANDIT PROBLEM WITH FINITE MEMORY

By Carter Vincent Smith[1] and Ronald Pyke

The Boeing Company and University of Washington

1. Summary. This paper studies the sequential decision model known as the two-armed-bandit with finite memory. It was introduced by Robbins [8] in 1956 and studied further by Isbell [5] in 1959. In this paper, a set of rules is defined which are uniformly better than those given in [5] and [8]. A much larger class of rules is then defined, one member of which is conjectured to be a uniformly best rule.

2. Introduction. The so-called two-armed-bandit problem is described as follows. One is invited to perform a sequence of coin-tossing experiments and at each performance, the experimenter is allowed to choose one of two available coins. The coins have probabilities $p_1 = 1 - q_1$ and $p_2 = 1 - q_2$, respectively, of giving heads. These probabilities are assumed to be unknown. The problem for the experimenter is to find a sequential decision rule to determine his choice of coin at each play of the game in such a way as to maximize the limiting frequency of heads. That solutions to this problem exist is a consequence of the strong law of large numbers, as was shown by Robbins [5]. (In [5], the problem was not restricted to the Bernoulli case as is done here.)

In 1956, Robbins [8] posed a modification of this problem in which the memory of the experimenter is restricted at each toss of a coin to cover only the r preceding tosses, $(r \geqq 1)$. For this modified, finite memory problem, Robbins suggested the following decision rule: *at any play of the game, change coins if the last r tosses resulted in r tails.* Denote this rule by R_r^0. For R_r^0, it is shown in [8] that the limiting frequency of heads is

$$(p_1q_2^r + p_2q_1^r)/(q_2^r + q_1^r). \tag{2.1}$$

In 1959, Isbell [5] suggested a refinement of R_r^0 namely, R_r^1: *change coins only when the memory shows that either r tails have been thrown with the same coin, or when r − 1 tails with one coin have been followed by a single tail with the other.* It is shown in [5] that the limiting frequency of heads for this rule is

$$[p_1q_2^r(1 - q_1^rq_2) + p_2q_1^r(1 - q_1q_2^r)][q_2^r(1 - q_1^rq_2) + q_1^r(1 - q_1q_2^r)]^{-1}. \tag{2.2}$$

Isbell [5] showed that (2.2) is never less than (2.1), and is always greater except for the boundary values of p_1 and p_2. Therefore, R_r^1 is uniformly better than R_r^0.

The principal purpose of this paper is to show that the existing rules for the Bernoulli case may be considerably improved. Secondly, however, the purpose of

Received 26 March 1965.

[1] The work of this author is in partial fulfillment of the requirements for the degree of Master of Science in Mathematical Statistics at the University of Washington.

the paper is also to provide an insight into the possible construction of optimal rules, which insight should also be applicable to the more general formulations of the problem which have been considered in the literature. (See for example [1], [4], [10] and [11].) One specific generalization of this problem that might have application, is to the case where one is able to scan two Poisson processes for a total of r minutes with the proviso that only one process may be observed at each instant of time. Suppose one further postulates that each process is observed for multiples of a given unit of time. Then the decision problem of choosing which process to observe at each instant of time, in order to maximize the proportion of time during which the process with the greater parameter is observed, is closely related mathematically to the Bernoulli model studied in this paper.

As is done by Isbell, the finite memory two-armed-bandit problem may be described in terms of a finite state homogeneous Markov chain. The state space S for the Markov chain consists of the 4^r r-tuples, $s = (s_1, s_2, \cdots, s_r)$, in which each coordinate s_i takes on one of the 4 symbols H_1, T_1, H_2, T_2, where H_2, for example, denotes that coin 2 was used and a head resulted. If the last coordinate of a state s is either H_1 or H_2 (T_1 or T_2), call s a *head state* (*tail state*). Similarly, s is called a *coin* 1 *state* (*coin* 2 *state*) if the last coordinate of s is either H_1 or T_1 (H_2 or T_2). A *rule*, R, is a function with domain S and range $\{1, 2\}$, with the interpretation that $R(s)$ denotes the coin to be tossed next when s is recorded in the memory. Randomized rules could also be defined, but they will not be considered in this paper.

Let $\{J_n : n \geqq 1\}$ be a Markov chain determined by a rule R, probabilities p_1 and p_2 and some initial distribution. That is to say, the probability of a transition by this Markov chain between a state $s = (s_1, s_2, \cdots, s_r)$ and a state $s' = (s_2, s_3, \cdots, s_r, x)$ is p_1, p_2, q_1 or q_2 according as $x = H_1$ and $R(s) = 1$, $x = H_2$ and $R(s) = 2$, $x = T_1$ and $R(s) = 1$ or $x = T_2$ and $R(s) = 2$. (The transition probabilities are zero in all other cases.)

Let h be the function defined on S by $h(s) = 1$ or 0 according as s is a head state or not. Then the number of heads which occur during the first n tosses is $Y_n = \sum_{i=1}^{n} h(J_i)$, $(n \geqq 1)$. Since the Markov chain has a finite number of states it is known that the sample frequency of heads, Y_n/n, converges a.s. to a random variable (rv), defined on the sequence sample space of the process $\{J_n : n \geqq 0\}$. Moreover, Y is constant over the union of the sets $[J_0 = s]$ for all s within the same recurrent class.

The conditional limiting expected frequency of heads

$$E(Y \mid J_0) = \lim_{n\to\infty} E(Y_n/n \mid J_0) \tag{2.3}$$

is a function only of J_0, R, p_1 and p_2. We will therefore write $E(Y \mid J_0) = F(J_0, R, p_1, p_2)$. The *worth*, $W(R, p_1, p_2)$, of a rule is defined as in [5] by

$$W(R, p_1, p_2) = \min_{s\varepsilon S}\{\min [F(s, R, p_1, p_2), F(s, R, p_2, p_1)]\}. \tag{2.4}$$

By using this quantity to compare rules, we are taking the minimax approach to the problem. Whenever $W(R, p_1, p_2) \geqq W(R', p_1, p_2)$ for all $0 \leqq p_1, p_2 \leqq 1$

we shall say that R is uniformly as good as R' and write $R \succ R'$. The most desirable rule, for any fixed r, would be one whose worth was never less than that of any other rule for any values of p_1 and p_2. Such a rule would be called a uniformly best rule. Isbell [5] has shown that his rule R_r^1 is such that $R_r^1 \succ R_r^0$ and that for each $r \geqq 2$ its worth is as high as that of any other rule for the very restricted case of either p_1 or p_2 equal to 0 or 1. Moreover, for $r = 1$, Isbell shows that R_1^1 (which in this case is equivalent to Robbins' rule R_1^0) is uniformly best. Although these results are very special indeed, they do serve as useful test cases for other rules that might be suggested.

A rule, R, is said to be *symmetric* if for all $s \in S$, $R(s) \neq R(s')$ where memory state s' is obtained from s by changing the subscripts of each coordinate of s. It is conjectured that for any rule R which is not symmetric, there exists a symmetric rule R' such that $R' \succ R$. In view of this and the symmetry of the expression for worth given in (2.4), we only consider symmetric rules in this paper.

In the search for a uniformly best rule, one may restrict one's attention to rules which give rise to Markov chains which have only one recurrence class. That this can be done without loss of generality is a consequence of the definition of W in (2.4). Since W is constant over each recurrence class of the Markov chain and is some average of these values when evaluated at a transient state, it is possible to alter the rule so that the Markov chain of the new rule will always reach some specified one of the original recurrence classes, and hence will have only one recurrence class. The proof of this is left to the reader.

It is also worth pointing out that if a uniformly best rule exists for a given r, then it will be a rule which, like those of Robbins, Isbell and of this paper, dictates a change of coin whenever the memory state consists of all tails by the same coin. This may be seen by considering what would happen when $p_1 = 0$ if the r-tuple $(T_1, T_1, \cdots, T_1)$ did not dictate a change and if it were used as an initial state.

Suppose now that R is a symmetric rule whose associated Markov chain has exactly one recurrent class. Let K be any one of the recurrent states and consider it fixed throughout the remainder of this section. Let $\{T_n, n \geqq 1\}$ be the successive occurrence times of state K, set $T_0 = 0$ and define the nth block B_n by

$$B_n = (J_{T_{n+1}}, J_{T_{n+2}}, \cdots, J_{T_{n+1}}), \qquad (n \geqq 0).$$

It will sometimes be convenient to refer to the sequence of tosses defined by the last coordinate of each state in B_n as the block B_n. Define $N_{ni}, H_{ni}, (i = 1, 2)$, to be respectively the number of coin i states and head states by coin i in the nth block B_n. Set $N_n = N_{n1} + N_{n2} = T_{n+1} - T_n$ and $H_n = H_{n1} + H_{n2}$. For $n \geqq 1$, $\{N_n\}$ and $\{H_n\}$ are sequences of independent and identically distributed rv's. Set $E(N_{11}) = n_1$ and $E(N_{12}) = n_2$. Note that n_1 and n_2 are functions of p_1 and p_2 ! We are now prepared to prove the following:

LEMMA 2.1. *For any rule R with one recurrent class of memory states,*

$$W(R, p_1, p_2) = \min [\alpha p_1 + (1 - \alpha)p_2, \alpha p_2 + (1 - \alpha)p_1],$$

where $\alpha = n_1(n_1 + n_2)^{-1}$ *is a function of* p_1 *and* p_2 *but not of* K *and* K'. *If, moreover,* R *is symmetric then* $W(R, p_1, p_2) = \alpha p_1 + (1 - \alpha)p_2$.

PROOF. Since there is only one recurrent class, it is known (cf. [1], p. 85) that $Y = \lim_{n\to\infty} Y_n/n$ is a constant, a.s. It is also well known that the proportion of coin 1 states among $\{J_1, J_2, \cdots, J_n\}$ has the same limit as $n^{-1}\sum_{m=1}^{n} N_{m1}$, namely $\alpha = n_1(n_1 + n_2)^{-1}$. Moreover, for $n > 0$,

$$E[h(J_n)] = p_1 P[R(J_{n-1}) = 1] + p_2 P[R(J_{n-1}) = 2],$$

which, when summed, relates the expected number of head states to the expected number of coin 1 and coin 2 states. The proof may then be completed straightforwardly.

For symmetric rules with one recurrence class, it is more convenient to work with the ratio $V(R, p_1, p_2) = n_1/n_2$ which, by Lemma 2.1, is equivalent to working with W, since W is a strictly increasing function of V for $0 < p_2 < p_1 < 1$.

3. The class of rules R_r^s. We define a class of new rules which are proved to be improvements on R_r^1.

DEFINITION. For memory length r, and integer s $(1 \leqq s \leqq r)$, the rule R_r^s is defined on the state space by:

R_r^s maps a memory state x into the subscript of the last coordinate (that is, it says do not switch coins) *except when the coordinates of x are* (1): *all tails by the same coin, or* (2): $(r - t - 1)$ *tails by one coin followed by t heads and one tail of the other* $(0 \leqq t \leqq s - 1)$.

The notation introduced here is consistent with that used earlier since the rule R_r^1 is Isbell's rule as defined in the previous section. For the boundary cases of either p_1 or p_2 equal to 0 or 1, the worth of any rule R_r^s for $s \leqq r - 2$ $(r \geqq 3)$ is equal to that of R_r^1.

For a rule R_r^s, define the state K to be the one whose first $(r - s)$ coordinates are T_2 and the remainder are H_1. Also define the state K' to be the one whose first $(r - s)$ coordinates are T_1 and the remainder H_2. That the rule R_r^s is symmetric is apparent from its definition. Consider a rule R_r^s with $s \leqq r - 2$ $(r \geqq 3)$. Since $0 < p_2 < p_1 < 1$, it is clear that with probability 1, the Markov chain determined by this rule must sometime pass through a sequence of $2r$ head, r tail, s head, r tail and s head states consecutively. A careful observance of rule R_r^s will show that both states K and K' must occur in this sequence. Consequently K and K' can be reached from all other states, and since this is a finite Markov chain, K and K' must be recurrent and there must be only one recurrent class (cf. Chung [2]). Thus we have proven

LEMMA 3.1. *For* $0 < p_2 < p_1 < 1$, $r \geqq 3$, *and* $s \leqq r - 2$, *the rule R_r^s is symmetric and has only one recurrent class which includes the states K and K'.*

In finding the worth of a rule R_r^s it is convenient to define several kinds of blocks, or sequences, of states within the Markov chain generated by it.

A *type B block* is a sequence of states from (but not including) an occurrence

of state K to (and including) the next occurrence. This type of block was discussed in Section 2.

A *type B_1 block* is a sequence of states from an occurrence of state K to the next occurrence of state K'.

A *type B_2 block* is a sequence of states from an occurrence of state K' to the next occurrence of state K.

A *type L_1 block* is a sequence of coin 1 states which ends with the first occurrence of the state of all tails by coin 1.

A *type S_1 block* is a sequence of coin 1 states which ends with the first occurrence of a tail state or with s successive head states.

Type L_2 and S_2 blocks are symmetrically defined. Type L_1 and L_2 blocks will be called *long blocks*. Type S_1 and S_2 blocks will be called *testing blocks* (since they test to determine if an L_1 or L_2 type block will follow). Testing blocks of all heads will be called *successful testing blocks* (since there is a change to a new kind of long block). All other testing blocks will be called *failed testing blocks*.

It can be shown that the number of states in any particular block is independent of previous history as given by the first $(r-1)$ coordinates of its first state. Thus the lengths of, or number of states in, blocks of type L_1, L_2, S_1, S_2 are independent random variables whose respective expected values are straightforwardly computed to be

$$\lambda_1 = (1 - q_1^r)/p_1q_1^r, \qquad \lambda_2 = (1 - q_2^r)/p_2q_2^r,$$
$$\sigma_1 = (1 - p_1^s)/q_1, \qquad \sigma_2 = (1 - p_2^s)/q_2.$$

For $s \leqq r - 2$, it can be seen that the occurrences of states K and K' always alternate. (The restriction $s \leqq r - 2$ is necessary for alternation since, for example, R_3^2 can generate the sequence $T_2, T_2, T_2, H_1, H_1, T_1, T_1, T_1, T_2, H_1, H_1$ with successive occurrences of state $K = (T_2, H_1, H_1)$.) Thus for $s \leqq r - 2$ a type B block always decomposes into a type B_1 and type B_2 block. The type B_1 block further decomposes into a sequence of pairs of long blocks and testing blocks which ends with a successful testing block. The B_2 block similarly decomposes. The number of pairs in a B_1 block is an independent random variable whose expected value is simply the inverse of the probability of a S_2 block being a successful testing block. Thus for a rule $R_r^s (s \leqq r - 2)$, the number of type L_1 or S_2 blocks in a type B block is a random variable with an expected value of $m_1 = 1/p_2^s$. Similarly the expected number of L_2 or S_1 blocks in a type B block is $m_2 = 1/p_1^s$. By Wald's fundamental identity the expected number of coin 1 and coin 2 states in a type B block are given respectively by $n_1 = m_1\lambda_1 + m_2\sigma_1$ and $n_2 = m_2\lambda_2 + m_1\sigma_2$.

We have thus proved the first part of the following theorem. The motivation for the second part stems from the fact that if σ_1 and σ_2 are neglected,

$$V \doteq m_1\lambda_1/m_2\lambda_2 = (p_1/p_2)^s\lambda_1/\lambda_2$$

which increases as s increases. In looser words, by increasing s, we make a

"successful test" more difficult to achieve and therefore make it more difficult to go from block B_1 to block B_2.

THEOREM 3.1. *Given memory length* $r \geqq 3$, $1 \leqq s \leqq r-2$, $1 > p_1 > p_2 > 0$, *and a rule* R_r^s, *then*

$$V(R_r^s, p_1, p_2) = (q_2^r/q_1^r) \cdot \{[p_1^{s-1}(1-q_1^r) + q_1^{r-1}p_2^s(1-p_1^s)]/[p_2^{s-1}(1-q_2^r) + q_2^{r-1}p_1^s(1-p_2^s)]\} \tag{3.1}$$

and

$$W(R_r^{s+1}, p_1, p_2) > W(R_r^s, p_1, p_2), \text{ for } r-s \geqq 3. \tag{3.2}$$

PROOF. (3.1) was proved above. By Lemma 3.1 and the relationship between V and W, (3.2) is equivalent to proving that

$$[p_1^s(1-q_1^r) + q_1^{r-1}p_2^{s+1}(1-p_1^{s+1})]/[p_2^s(1-q_2^r) + q_2^{r-1}p_1^{s+1}(1-p_2^{s+1})] > [p_1^{s-1}(1-q_1^r) + q_1^{r-1}p_2^s(1-p_1^s)]/[p_2^{s-1}(1-q_2^r) + q_2^{r-1}p_1^s(1-p_2^s)].$$

Abbreviate the terms of this inequality by introducing the notation as indicated in the obvious way in

$$(a'+b')/(c'+d') > (a+b)/(c+d). \tag{3.3}$$

Since all terms are positive, (3.3) is equivalent to $Z \equiv a'c + a'd + b'c + b'd - ac' - ad' - bc' - bd' > 0$. Grouping these terms and simplifying, one obtains

$$\begin{aligned} a'c - ac' &= (p_1 - p_2)p_1^{s-1}p_2^{s-1}(1-q_1^r)(1-q_2^r), \\ a'd + b'c - ad' - bc' &= p_1^s p_2^s[q_1^r q_2^r(p_1^s - p_2^s) - (p_1^s q_2^r - p_2^s q_1^r)], \\ b'd - bd' &= p_1^s p_2^s[q_1^r q_2^r(p_2^s - p_1^s) - q_1^{r-1}q_2^{r-1}\{q_2(1-p_1^s) \\ &\qquad - q_1(1-p_2^s)\}]. \end{aligned}$$

Adding up these terms, we obtain

$$Z = (p_1 - p_2)p_1^{s-1}p_2^{s-1}(1-q_1^r)(1-q_2^r) - p_1^s p_2^s[(p_1^s q_2^r - p_2^s q_1^r) + q_1^{r-1}q_2^{r-1}\{q_2(1-p_1^s) - q_1(1-p_2^s)\}].$$

Now $Z > 0$ if $Z' = Z/p_1^s p_2^s > 0$. Also, since $(1-q_1^r)(1-q_2^r) > p_1 p_2$, Z' will be positive if

$$N(s) \equiv (p_1 - p_2) - (p_1^s q_2^r - p_2^s q_1^r) - q_1^{r-1}q_2^{r-1}[q_2(1-p_1^s) - q_1(1-p_2^s)] > 0.$$

We will in fact show by induction that $N(s)$ is positive and increasing. First of all, one obtains

$$\begin{aligned} N(1) &= (p_1 - p_2) - (p_1 q_2^r - p_2 q_1^r) = p_1(1-q_2^r) - p_2(1-q_1^r) \\ &= p_1 p_2(1 + q_2 + \cdots + q_2^{r-1} - 1 - q_1 \cdots - q_1^{r-1}) > 0. \end{aligned}$$

Suppose $N(s)$ is greater than zero, then

$$N(s+1) - N(s) = p_1^s q_1 q_2^r - p_2^s q_2 q_1^r - q_1^{r-1} q_2^{r-1}[q_1 q_2 p_1^s - q_1 q_2 p_2^s]$$
$$= q_1 q_2[p_1^s q_2^{r-1}(1 - q_1^{r-1}) - p_2^s q_1^{r-1}(1 - q_2^{r-1})],$$

and a term by term comparison shows this to be positive.

4. The class of rules R_r^δ. The rules R_r^s of the preceding section, for $r \geqq 3$ and $s \leqq r - 2$, are nice rules in the sense that the type B_1 and B_2 blocks generated by them alternate and are separable into sequences of pairs of long blocks and testing blocks which end with a successful (all heads) testing block. These rules can be modified so that they still generate sequences of pairs of long and testing blocks, but in which the testing blocks contain tosses by both coins. For example, rule R_4^2 can be modified so that the successful S_2 testing block is (H_2, T_1, H_2) instead of (H_2, H_2). The other (or failed) S_2 testing blocks would then be (T_2), (H_2, H_1), and (H_2, T_1, T_2) instead of (T_2) and (H_2, T_2). The successful S_1 testing block would be (H_1, T_2, H_1) by symmetry.

The motivation behind this type of modification is as follows. In the last section we saw that as s increased the rules R_r^s improved. But increasing s can be interpreted as making a successful testing block harder to obtain. This suggests that when working with a finite-memory, sequential decision problem, one should be guided by the philosophy that the coin is assumed to be the best one (innocent) until "proved" to be the worst one (guilty). Moreover, the more reliable the "proof" (that is, the longer the test) the better the rule.

We shall now introduce a class of rules which, like the previous example, can be defined completely in terms of their successful S_2 testing blocks, but which blocks are much longer than those studied in Section 3. By incorporating tosses by both coins into a successful testing block one can code the possible memory states in such a way as to increase the length of a successful testing block to an order of magnitude of 2^r instead of r. It is certainly natural to require that a successful S_2 testing block demand heads by coin 2 and tails by coin 1. Such a block would then be a sequence whose coordinates are either H_2 or T_1 and so therefore could be described by a vector whose coordinates are 1 (for H_2) and 0 (for T_1). Failed S_2 testing blocks would then consist of a portion of the successful testing block and a failure, either a T_1 or H_2. We now define the type of sequence of 0's and 1's which may be used to represent a successful testing block, and which in turn may be used to determine a rule.

DEFINITION. For a given memory length r, a *δ-vector* is any vector whose coordinates are either 0 or 1 which satisfies Conditions (1) and (2) below. An *extended δ-vector* is constructed from a δ-vector by adding r 0's to the front of the δ-vector and sufficient 1's to the end to make a total of r in succession. [For example let (101) be a δ-vector for $r = 4$, then (0000101111) is the associated extended δ-vector.]

CONDITION 1. In the extended δ-vector, each successive r-tuple is unique.

[In the above example, the r-tuples (0000), (0001), (0010), (0101), (1011), (0111), and (1111) are unique.]

CONDITION 2. There are no more than $(r - 2)$ 0's or $(r - 2)$ 1's in succession in the δ-vector.

The problem of constructing vectors whose coordinates are 0 or 1 such that each successive r-tuple is unique has been widely dealt with. The reader is referred to Chapter 9 of the recent book [9] by S. K. Stein for a general review of the subject. In 1934, M. H. Martin [6] proved that it was possible to construct a vector of the maximum length, $2^r + (r - 1)$. In 1946, N. G. de Bruijn [3] proved that the number of maximum length vectors is 2^N where $N = 2^{r-1} - r$. The vector that Martin used in his proof was constructed by starting with r 1's and adding a 0 if the newly formed r-tuple has not occurred before and a 1 otherwise. [For $r = 4$ this construction leads to (1111000010011010111).] Denote by δ_r^* (or sometimes δ^*) the vector formed by dropping the initial r 1's and r 0's and the final 1 from the vector formed according to this construction [giving $\delta^* = (1001101011)$ for $r = 4$]. The proof that δ_r^* is a δ-vector follows from the construction of the vector which determined it. The length of δ_r^* is $2^r - r - 2$ and it always starts with one 1 followed by $(r - 2)$ 0's.

In order to define the rule which is determined by a given δ-vector, it is necessary to make several definitions. The *head-tail vector* of a state is constructed by replacing a H_1 or H_2 coordinate with a 1 and a T_1 or T_2 coordinate with a 0. The *coin vector* of a state is constructed by replacing a H_2 or T_2 coordinate with a 1 and a H_1 or T_1 coordinate with a 0. A *complement coin vector* of a state is constructed by replacing 1's with 0's and 0's with 1's in the coin vector of the state. [For (T_1, T_2, H_1, H_2) these vectors are respectively (0011), (0101), and (1010).] For a given r and a given δ-vector define *δ-r-tuples* to be all sequential r-subtuples of the vector constructed by adding r 0's to the front of the δ-vector. [Thus for $r = 3$, the δ-vector (101) has δ-r-tuples (000), (001), (010), (101).]

DEFINITION. For memory length r, and δ-vector δ, the rule R_r^δ is defined on the state space by:

(i) R_r^δ maps a memory state x into 1 (2) if the coin vector (complement coin vector) of x is a δ-r-tuple of δ and if the head-tail vector of x is equal to this δ-r-vector except for the last coordinate. In other words, at the end of a failed testing block use the coin used in the last long block.

(ii) R_r^δ switches coins (i.e. it maps a memory state x into 1 (2) if x is a coin 2 (1) state) if the head-tail vector of x is (a) a δ-r-tuple whose last coordinate is not equal to the one following it in the extended δ-vector, and (b) either a coin vector or a complement coin vector. In other words, within testing blocks which have not failed continue to follow the pattern of the δ-vector.

(iii) In all other cases R_r^δ maps x into the subscript of its last coordinate (do not switch coins).

Blocks can be defined for a rule R_r^δ as in Section 3 with the key state K being defined as a state whose complement coin vector and head-tail vector are equal

to the last r coordinates of δ (preceded by at most r 0's if δ has fewer than r coordinates). In other words, K is the last state of a successful S_1 testing block. Blocks of type B, B_1, B_2, L_1, and L_2 are defined as before, and successful and failed testing blocks are defined in a straightforward manner from the δ-vector. It is easily seen that a rule R_r^{δ} will initiate switches (if necessary) at the end of long and testing blocks and also within the testing blocks. That the rule will not initiate a switch for some state that contains part of a testing block and part of a long block follows from the requirements of the δ-vector.

Thus the lengths of the S_1, S_2, L_1, and L_2 type blocks are independent random variables as are the numbers of pairs of long blocks and testing blocks in B_1 or B_2 blocks.

The analysis of Section 3 can be used to find a value for V in this general case. The only change is that testing blocks include tosses by both coins. One proceeds as follows.

For memory length r, let δ be a δ-vector of length s. Let δ_k be the kth coordinate of δ and set $\delta_0 = 0$ and $\rho_k = \sum_{i=0}^{k} \delta_i$ for $k \geqq 0$. Then the probability of completing a successful S_2 testing block is given by $t_2 = p_2^{\rho_s} q_1^{s-\rho_s}$. The expected number of tosses in a S_2 block is $\sum_{k=0}^{s-1} p_2^{\rho_k} q_1^{k-\rho_k}$. This can be separated into the expected number of coin 2 tosses,

$$\sigma_2 = \sum_{k=0}^{s-1} \delta_{k+1} p_2^{\rho_k} q_1^{k-\rho_k},$$

and the expected number of coin 1 tosses,

$$\sigma_1' = \sum_{k=0}^{s-1} (1 - \delta_{k+1}) p_2^{\rho_k} q_1^{k-\rho_k}$$

The expected number of occurrences of a L_1 or S_2 type block in a B_1 type block is $m_1 = 1/t_2$. The expected number of tosses in a L_1 type block is still given by $\lambda_1 = (1 - q_1^r)/p_1 q_1^r$. If t_1, σ_1, σ_2', and m_2 are defined in a symmetrical manner then by the same argument as used in Section 3, the following theorem is true:

THEOREM 4.1. *For $r \geqq 3$, a given δ-vector δ, and $0 < p_2 < p_1 < 1$, R_r is symmetric, has one recurrence class and*

$$V(R_r, p_1, p_2) = [t_1(\lambda_1 + \sigma_1') + t_2\sigma_1]/[t_2(\lambda_2 + \sigma_2') + t_1\sigma_2].$$

For the boundary cases of either p_1 or p_2 equal to 0 or 1, it is verifiable that any R_r^{δ} is equivalent to R_r^1.

For the special rule δ_r^* described earlier,

$$m_1/m_2 = (p_1/p_2)^M (q_2/q_1)^N$$

where $M = 2^{r-1} - 2$ and $N = 2^{r-1} - r$. Thus if the expected lengths of the testing blocks are ignored,

$$V(R_r^{\delta^*}, p_1, p_2) \doteq (p_1/p_2)^M (q_2/q_1)^N (\lambda_1/\lambda_2), \tag{4.1}$$

as compared with

$$V(R_r^{r-2}, p_1, p_2) \doteq (p_1/p_2)^{r-2} (\lambda_1/\lambda_2) \tag{4.2}$$

TABLE I

Comparison of V ratios

r	p_1	p_2	$V(R_r^1)$	$V(R_r^{r-2})$	$V(R_r^{\delta^*})$
4	0.500	0.400	2.13	2.55	9.11
5	0.500	0.400	2.54	3.76	50.02
6	0.500	0.400	3.03	5.62	122.33
7	0.500	0.400	3.62	8.46	250.36
8	0.500	0.400	4.33	12.76	506.34
9	0.500	0.400	5.18	19.23	1018.33
10	0.500	0.400	6.21	28.98	2042.33
11	0.500	0.400	7.44	43.64	4090.32
12	0.500	0.400	8.92	65.67	8186.32
13	0.500	0.400	10.71	98.76	16378.32
14	0.500	0.400	12.84	148.45	32762.32
15	0.500	0.400	15.41	223.03	65530.32
16	0.500	0.400	18.49	334.96	131066.32
17	0.500	0.400	22.19	502.92	262138.32
18	0.500	0.400	26.62	754.94	524282.32
19	0.500	0.400	31.95	1133.04	1048570.33
20	0.500	0.400	38.34	1700.29	2097146.33
5	0.500	0.499	1.01	1.01	1.06
5	0.500	0.490	1.11	1.15	1.75
5	0.500	0.400	2.54	3.76	50.02
5	0.500	0.100	26.06	53.21	61.33
10	0.500	0.499	1.02	1.03	7.68
10	0.500	0.490	1.22	1.40	2041.14
10	0.500	0.400	6.21	28.98	2042.33
10	0.500	0.100	432.06	1841.32	2045.32

and

$$V(R_r^1, p_1, p_2) \doteq (p_1/p_2)(\lambda_1/\lambda_2). \tag{4.3}$$

This indicates a considerable superiority of R_r^{δ} over the rules considered in Section 3. A numerical comparison of the exact V's for these rules is given in Table I.

Although any maximum length δ-vector has the same ratio m_1/m_2 as δ_r^*, it is felt that δ_r^* is best because it appears to have the minimum expected number of coin 2 tosses in an S_2 testing block. *It is conjectured that $R_r^{\delta^*}$ is the uniformly best rule for memory length r.* Theorem 4.2 which follows lends a little support to this conjecture. Moreover, Theorem 4.3 proves that the worth of $R_3^{\delta^*}$ is greater than that of any other rule based on a δ-vector for $r = 3$.

THEOREM 4.2. *Let R_r^{s0} denote the rule with δ-vector consisting of s 1's followed by one 0 for $1 \leqq s \leqq r - 2$. Then for $0 < p_2 < p_1 < 1$,*

$$W(R_r^{s0}, p_1, p_2) > W(R_r^s, p_1, p_2). \tag{4.4}$$

PROOF. From the results preceding Theorem 3.1 one obtains for rule R_r^s that

$$t_1 = p_1^s, \qquad t_2 = p_2^s,$$
$$\lambda_1 = (1 - q_1^r)/p_1 q_1^r, \qquad \lambda_2 = (1 - q_2^r)/p_2 q_2^r,$$
$$\sigma_1 = (1 - p_1^s)/q_1, \qquad \sigma_2 = (1 - p_2^s)/q_2,$$
$$\sigma_1' = 0 = \sigma_2'.$$

For rule R_r^{s0}, the superscript 0 will be added to the above symbols to stand for the appropriate expected values. Then by straightforward calculation one obtains

$$t_1^0 = t_1 q_2, \quad t_2^0 = t_2 q_1, \quad \lambda_1^0 = \lambda_1, \quad \lambda_2^0 = \lambda_2,$$
$$\sigma_1^0 = \sigma_1, \quad \sigma_2^0 = \sigma_2, \quad \sigma_1'^0 = t_2, \quad \sigma_2'^0 = t_1.$$

Upon substituting these values into Theorem 4.1, one obtains that (4.4) will hold if and only if

$$(4.5) \qquad (q_2 - q_1)(\lambda_1\lambda_2 - \sigma_1\sigma_2) + t_1 q_2 \sigma_2 + q_2 t_2 \lambda_2 - t_2 q_1 \sigma_1 - q_1 t_1 \lambda_1 > 0.$$

Now set the last four terms of (4.5) equal to $Q(s)$. Then by substitution one obtains

$$Q(s) = p_2^{s-1}(q_2^{1-r} - 1) - p_1^{s-1}(q_1^{1-r} - 1)$$

which is straightforwardly checked to be an increasing function of s. Therefore $Q(s) > Q(1) = q_2^{1-r} - q_1^{1-r}$. Consequently, (4.5) is true if

$$(4.6) \qquad (q_2 - q_1)\lambda_1\lambda_2 > (q_2 - q_1)\sigma_1\sigma_2 - Q(1).$$

Now

$$\sigma_1\sigma_2 = (1 - p_1^s)(1 - p_2^s)/q_1 q_2 < 1/q_1 q_2,$$

so that (4.6) holds if $(q_2 - q_1)\lambda_1\lambda_2 > (q_2 - q_1)/q_1 q_2 - Q(1)$, or if

$$(4.7) \qquad (q_2 - q_1)(1 - q_1^r)(1 - q_2^r)/p_1 p_2 > q_1 q_2[(q_2 - q_1)q_2^{r-2} q_1^{r-2} + q_2^{r-1} - q_1^{r-1}].$$

But (4.7) is easily checked to be true, thereby completing the proof.

THEOREM 4.3. *For $r = 3$, and any δ-vector rule R_3^δ, $W(R_3^{\delta^*}, p_1, p_2) \geqq W(R_3, p_1, p_1)$ for all p_1, p_2. Specifically,*

$$R_3^{(101)} > R_3^{(10)} > R_3^{(1)}.$$

PROOF. Since the only δ-vectors for $r = 3$ are (101), (10), and (1), the first inequality is compatible with the stated ordering of the corresponding rules. The second part of this inequality was proved in Theorem 4.2 and therefore, it remains to be proved that

$$(4.8) \qquad V(R_3^{(101)}, p_1, p_2) - V(R_3^{(10)}, p_1, p_2) > 0.$$

It is easily verified, using Theorem 4.1, that

$$V(R_3^{(10)}, p_1, p_2) = [p_1 q_2(\lambda_1 + p_2) + p_2 q_1]/[p_2 q_1(\lambda_2 + p_1) + p_1 q_2],$$

and

$$V(R_3^{(101)}, p_1, p_2) = [p_1^2 q_2(\lambda_1 + p_2) + p_2^2 q_1(1 + p_1 q_2)]/[p_2^2 q_1(\lambda_2 + p_1) + p_1^2 q_2(1 + p_2 q_1)],$$

where for $r = 3$, $\lambda_i = (1 - q_i^3)/p_i q_i^3$, for $i = 1$ and 2. Upon substitution of these identities into (4.8), the proof can be completed straightforwardly.

The authors are indebted to Professor N. D. Ylvisaker for supplying references [3], [6], and [9].

REFERENCES

[1] BRADT, R. N., JOHNSON, S. M. and KARLIN, S. (1956). On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.* **27** 1060–1074.

[2] CHUNG, KAI LAI (1960). *Markov Chains with Stationary Transition Probabilities*. Springer-Verlag, Berlin.

[3] DE BRUIJN, N. G. (1946). A combinatorial problem. *Nederl. Akad. Wetensch. Indag. Math.* **8** 461–467.

[4] FELDMAN, DORIAN (1962). Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33** 847–856.

[5] ISBELL, J. R. (1959). On a problem of Robbins. *Ann. Math. Statist.* **30** 606–610.

[6] MARTIN, M. H. (1934). A problem in arrangements. *Bull. Amer. Math. Soc.* **40** 859–864.

[7] ROBBINS, HERBERT (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 529–532.

[8] ROBBINS, HERBERT (1956). A sequential decision problem with a finite memory. *Proc. Nat. Acad. Sci.* **42** 920–933.

[9] STEIN, SHERMAN K. (1963). *Mathematics, The Man Made Universe*. Freeman, San Francisco, 110–121.

[10] VOGEL, WALTER (1960). A sequential design for the two armed bandit. *Ann. Math. Statist.* **31** 430–443.

[11] VOGEL, WALTER (1960). An asymptotic minimax theorem for the two armed bandit problem. *Ann. Math. Statist.* **31** 444–451.