Reprinted from INFORMATION AND CONTROL, Volume 12, Nos. 5/6, May-June 1968

INFORMATION AND CONTROL **12**, 371–377 (1968)

A Note on the Two-Armed Bandit Problem with Finite Memory

THOMAS M. COVER[1]

Stanford University, Electronic Research Laboratories, Stanford, California 94305

Robbins has proposed a finite memory constraint on the two-armed bandit problem in which the coin to be tossed at each stage may depend on the history of the previous tosses only through the outcomes of the last r tosses. Letting the choice of coin depend on the time at which the toss is made, we exhibit a deterministic rule with memory $r = 2$, the description of which is independent of the coin biases p_1 and p_2, which achieves, with probability one, a limiting proportion of heads equal to $\max\{p_1, p_2\}$. Thus this rule is asymptotically uniformly best among the class of time-varying finite memory rules.

1. INTRODUCTION

In the two-armed bandit problem with finite memory, we are given two coins with unknown probabilities, p_1 and p_2, of heads. At each stage, based only on the results of the previous r tosses, we must decide which coin to toss next. (Following Robbins 1956) we define the result of a toss to mean both which coin was used and which face came up.) Our goal is to find a rule which maximizes the limiting proportion of heads.

If the best coin were known initially, this coin could be used for each toss, and the limiting proportion of heads $\max\{p_1, p_2\}$ would be obtained with probability one. It is clear that this limiting proportion may be obtained only if one resolves, with probability of error zero, the hypothesis $p_1 > p_2$ vs. the hypothesis $p_1 < p_2$. Hence, it is necessary, during the infinite sequence of trials, that an infinite number of observations be made and "remembered" on *each* coin. This appears to be impossible under a finite memory constraint.

Robbins (1952) investigated the two-armed bandit problem in a

[1] This work was supported at Stanford University under contract AF 49(638) 1517 and was completed at IBM Research Center, Yorktown Heights, N. Y.

case where the memory is unrestricted and demonstrated a procedure which sampled the "inferior" coin infinitely often, but with an ever-decreasing frequency, in such a manner that max $\{p_1, p_2\}$ was achieved. Robbins (1956) posed the memory r constraint and suggested a rule which begins to change coins when sufficient negative information (r consecutive tails) is obtained. Isbell (1959) improved Robbins' rule uniformly in $\{p_1, p_2\}$; and Smith and Pyke (1965) considered a family of rules, containing Isbell's as a special case, which obtained an even further improvement. Finally, Samuels (1966) investigated randomized versions of the rules of the previous three papers to obtain improvements in each case. None of the above time-invariant rules with finite memory achieves max $\{p_1, p_2\}$.

The subtleties and pitfalls in making a good definition of memory may be seen when we examine the work of Varshavskii and Vorontsova (1963) in which they present a time-invariant stochastic automaton with apparent memory $r = 1$ which achieves max $\{p_1, p_2\}$ for certain special values of p_1 and p_2. This rule involves incrementing the probability of selecting a coin when that coin is tossed and yields a head. However, this procedure violates the finite memory requirement, because the amount of memory required to store the description of the coin transition probabilities is infinite. (When the memory of possible coin transitions is constrained to be finite, the procedure fails.) With such a memory, one is essentially remembering the infinite past. Given Varshavskii's freedom, we may greatly improve on his technique. For example, let a sequence of n coin flips be represented by a real number r_n whose binary expansion has n 0's to the right of the decimal place followed by a sequence of 1's and 0's encoding the outcomes of the first n tosses. Clearly r_n may be updated to r_{n+1} on the basis solely of the knowledge of r_n and the current coin toss outcome.

Let $\tilde{r}_n$ equal r_n or $(1 - r_n)$ accordingly as Robbins' 1952 procedure prescribes the selection of coin 1 or coin 2. Then $\tilde{r}_n$ describes a random automaton, with the interpretation that coin 1 is to be flipped with probability $\tilde{r}_n$. Thus we are following Robbins' deterministic infinite memory procedure "most" of the time, and max $\{p_1, p_2\}$ is achieved. Hence Varshavskii's memory constraint is not really finite and results in only a minor inconvenience in achieving the desired goals.

In this note, we shall assume that we have access to a clock which provides a suitable notion of the number of tosses observed at any given stage. However, no attempt will be made to cheat by storing data in the

clock. Note that our knowledge of the precise outcomes of the past experiments is really finite, because the clock time is independent of the data in the sense that it furnishes no information about the hypothesis. Under the assumption of a time-varying decision rule, we shall describe a procedure achieving $\max \{p_1, p_2\}$, with a memory of $r = 2$.

The solution of the problem involves three main ideas. The first is the observation that information may be "passed on ahead," even with the finite memory limitation, by cleverly selecting future coins to be tossed.[2] The second is the idea, in the finite memory case, of compounding experiments with arbitrarily high information out of experiments of limited information. Such a technique is a special case of a general theory of hypothesis testing with finite memory in the sequential data case Cover (1967). We are thus able to test the hypothesis $p_1 > p_2$ with arbitrarily small probability of error. The third idea is that of interleaving trials and tests, with trial lengths increasing in such a manner as to swamp out preceding trials and tests. This last idea has the flavor of Robbins' 1952 solution of the unrestricted 2-armed bandit problem.

2. THE PROCEDURE

We shall follow the procedure of interleaving test blocks $T_1, T_2, \cdots$ with trial blocks $U_1, U_2, \cdots$. Each test block will test the hypothesis $p_1 > p_2$ vs. $p_1 < p_2$. The "favorite" coin resulting from this test will then be used exclusively for the ensuing trial block.

Each test block T will be begun arbitrarily with coin 1 as the favorite. (This precaution yields independence of the test blocks.) A test block will be broken into m subblocks each consisting of $2s$ tosses. Let $\theta \in \{1, 2,\}$ be the label of the favorite coin at the beginning of a subblock, and let $\bar{\theta}$ be the other coin. A subblock test will be said to be a success if 2s tosses yield an unbroken sequence of TH's.

During a test subblock, on every odd-numbered trial the favorite coin (the one with which the subblock was begun) will be thrown. This enables us to remember the favorite coin at any time. On even-numbered tosses, the alternate coin will be thrown as long as the desired $THTH \cdots TH$ sequence is being observed. As soon as a break in this sequence is observed, the alternate coin will not be tossed again during that subblock. Instead, the favorite coin will be thrown from then on, and we will know henceforth, from observing that the last two coins

[2] I am indebted to Professor M. Arbib for this remark.

tossed were the same, that a break in the sequence has occurred. In any case, $2s$ tosses will be made in the subblock. Thus, when the test subblock is terminated, we shall know the identity of the favorite coin, and either (i) that the desired TH sequence was unbroken, in which case the alternate coin to the favorite becomes the new favorite, or (ii) that the desired sequence was not obtained, in which case the old favorite is retained as the new favorite. In the next section there appears an explicit description of this rule in which the details of the orderly transition from the unbroken-sequence mode to the broken-sequence mode are made clear.

At the termination of each subblock, the new favorite coin is used to begin the next subblock until m subblock tests have been performed. This collection of subblocks comprises a test block. Thus $4ms$ tosses of the coin are made in the test block T.

3. DETAILS OF THE TEST BLOCK

Let the sequence of coins tossed $(\theta_1, \theta_2, \cdots, \theta_{2s})$, $\theta_i \in \{1, 2\}$, and outcomes observed $(x_1, x_2, \cdots, x_{2s})$, $x_i \in \{H, T\}$, be divided into pairs

$$\binom{\theta_1, \theta_2}{x_1, x_2}\binom{\theta_3, \theta_4}{x_3, x_4}\cdots\binom{\theta_{2s-1}, \theta_{2s}}{x_{2s-1}, x_{2s}}$$

The memory of the past at time n is either the state

$$\binom{\theta_{n-1}, \theta_n}{x_{n-1}, x_n} \quad \text{or} \quad \begin{matrix}\theta_{n-1} \\ x_{n-1}\end{matrix}\binom{\theta_n}{x_n}$$

accordingly as n is even or odd. Thus the memory is of length $r = 2$. Figure 1 exhibits a strategy for coin selection as a function of the state.

The subblock test begins in the starting state with the tossing of coin Θ, the current favorite coin. If coin Θ results in T, then coin $\bar{\Theta}$ (the other coin) is tossed, resulting in the new state $\binom{\Theta\bar{\Theta}}{T}$. If now coin $\bar{\Theta}$ results in H, coin Θ is tossed and the new state is $\binom{\bar{\Theta}}{H}\binom{\Theta}{}$. As long as heads and tails continue to alternate, these 2 states and the coin choices alternate. As soon as a break in this string occurs, an orderly transition is begun into the "copying mode" in which a terminal sequence of Θ's is eventually produced. Inspection will verify that this block test will result in $\Theta_{2s-1} = \Theta_1$, thus keeping track of the initial state Θ_1, and will indicate a string of s consecutive TH's only if the final state is of the

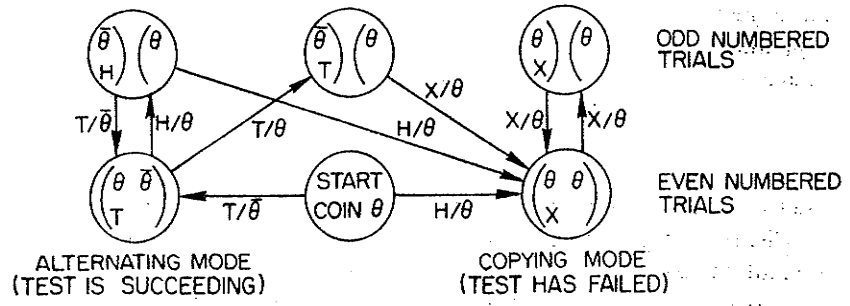

FIG. 1. State transition diagram (memory $r = 2$).

form $\binom{\theta\ \bar{\theta}}{T\ H}$. The following are examples of this process:

$$\binom{1}{T}\binom{2}{H}\binom{1}{T}\binom{2}{H}\binom{1}{H}\binom{1}{X}\binom{1}{X}\binom{1}{X}\cdots\binom{1}{X}\binom{1}{X}$$ failure; coin 1 is the initial coin.

$$\binom{2}{T}\binom{1}{H}\binom{2}{T}\binom{1}{T}\binom{2}{X}\binom{2}{X}\binom{2}{X}\binom{2}{X}\cdots\binom{2}{X}\binom{2}{X}$$ failure; coin 2 is the initial coin.

$$\binom{2}{T}\binom{1}{H}\binom{2}{T}\binom{1}{H}\binom{2}{T}\binom{1}{H}\binom{2}{T}\binom{1}{H}\cdots\binom{2}{T}\binom{1}{H}$$ success; coin 2 is the initial coin.

4. ANALYSIS

A subblock of $2s$ tosses results in transition probabilities from favorite coin 1 to 2 with probability $(p_2q_1)^s$ and from 2 to 1 with probability $(p_1q_2)^s$. The stationary probabilities of coin 1 and 2 being the favorite are then $1/(1 + \alpha^s)$ and $\alpha^s/(1 + \alpha^s)$ respectively, where

$$\alpha = q_1p_2/p_1q_2 .$$

Observe that $\alpha < 1$, if and only if, $p_1 > p_2$. These probabilities are approached exponentially by the finite trial block probabilities as m, the number of subblocks, tends to infinity.

We see that the stationary probability t of deciding on the inferior

coin is just $t = \min \{1/(1 + \alpha^s), \alpha^s/(1 + \alpha^s)\}$. By the symmetry of t in p_1 and p_2, we may, without loss of generality, assume $p_1 > p_2$. Thus $\alpha < 1$ and

$$t = \frac{\alpha^s}{1 + \alpha^s} < \alpha^s < 1.$$

Let t_i be the probability of selecting the inferior coin with test block T_i. Clearly t_i depends on α, m_i, s_i and approaches t as $m_i \to \infty$. First we shall choose $\{s_i\}$ so that $\Sigma\alpha^{s_i} < \infty$, e.g., $s_i = i$. We then choose $\{m_i\}$ large enough so that, for each $\alpha < 1$, $t_i < \alpha^{s_i}$ for sufficiently large i. This may easily be done. Thus it is assured that $\Sigma t_i < \infty$.

If we take care to make the test blocks $i = 1, 2, \cdots$, independent, we may conclude, from the Borel zero-one law and the finiteness of Σt_i, that with probability one only a finite number of test blocks T_i will result in an incorrect choice of coin. (The block tests have been made independent, for the sake of the argument, by letting coin 1 be the favorite at the beginning of each test block, thus ignoring all the previous information.)

Let u_i be the number of trials in the trial block U_i. We shall select numbers $u_1, u_2, \cdots$ so that

$$\frac{u_i}{\sum_{j=1}^{i+1} 4m_j s_j + \sum_{j=1}^{i} u_j} \to 1.$$

Note that u_i has been chosen large enough to dominate the number of trials $4m_{i+1}s_{i+1}$ in the *next* test block.

If we now use, for the next u_i trials, the coin chosen by block T_i, we may conclude that the proportion of heads obtained in the first n trials tends to $\max \{p_1, p_2\}$, with probability one, as $n \to \infty$. This, of course, is the maximum achievable limit.

5. CONCLUSIONS

Good attempted solutions to Robbins' two-armed bandit problem with finite memory, both for the time-varying rules discussed here and the time-invariant rules discussed in previous literature, tend to violate the spirit of the memory constraint in the sense that the effective memory into the past is much greater than r. For example, the δ^*-rule of Smith and Pyke uses the memory as a counter to achieve direct influence as far back as 2^r tosses as well as indirect influence indefinitely far back.

The time-varying finite memory rule discussed here has an effective memory which is a slowly-growing unbounded function of the sample number n, thus enabling us to achieve the ultimate goal. In light of this we should like to suggest a family of problems in which the memory constraint is defined differently.

Specifically, let the decision rule for the transition to the next state in memory depend on the previous state and the outcome of the current coin toss, where the previous state is one of a finite number m of states in the memory. The coin to be tossed may depend only on the current memory state. As before, we have the distinction between rules which allow the state transition to depend on n and those that do not. Thus the constraint is on how much is remembered rather than on how long ago. It may be seen from the considerations in this paper that a time-varying rule with finite memory ($m = 4$ is sufficient) will achieve $\max \{p_1, p_2\}$. The time-invariant case will require new techniques.

ACKNOWLEDGMENTS

I would like to thank M. Arbib, D. Sagalowicz, S. Samuels and M. Hellman for their helpful remarks.

RECEIVED: June 21, 1967

REFERENCES

ROBBINS, H. (1956), A sequential decision problem with finite memory. *Proc. Natl. Acad. Sci.* **42**, 920–923.

ROBBINS, H. (1952), Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**, 527–535.

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

SMITH, C. V., AND PYKE, R. (1965), The Robbins–Isbell two-armed bandit problem with finite memory. *Ann. Math. Stat.* **36**, 1375–1386.

SAMUELS, S. M. (1966), "Randomized Rules for the Two-Armed Bandit with Finite Memory." Purdue Univ. Tech. Rept. Mim. Series No. 71, Dept. of Statistics.

VARSHAVSKII, V. I., AND VORONTSOVA, I. P. (1963), On the behavior of stochastic automata with a variable structure. *Automation, Remote Control*, **24**, 327–333.

COVER, T. M. (1967), "Learning with Finite Memory," paper delivered at IEEE International Symposium on Information Theory, Athens, Greece, September 1967. (Submitted to *Ann. Math. Stat.*)