TRANSACTIONS

of the

FOURTH PRAGUE CONFERENCE

on

INFORMATION THEORY, STATISTICAL DECISION FUNCTIONS, RANDOM PROCESSES

held at

Prague, from August 31 to September 11, 1965

T. M. Cover

BEHAVIOUR OF SEQUENTIAL PREDICTORS OF BINARY SEQUENCES

PUBLISHING HOUSE
OF THE
CZECHOSLOVAK ACACEMY OF SCIENCES

PRAGUE 1967

BEHAVIOR OF SEQUENTIAL PREDICTORS OF BINARY SEQUENCES

THOMAS M. COVER

STANFORD

1. INTRODUCTION

This paper concerns the behavior of sequential predictors of finite sequences of 0's and 1's. Robbins [1], Hannan [2] and Blackwell [3] and [4] have suggested algorithms for the prediction of sequences which asymptotically achieve certain desirable scores depending upon the empirical distribution of the 0's and 1's in the observed sequence. Other researchers have proposed ad hoc schemes which detect certain types of recurring "patterns" in long sequences and use this information to predict the future. Fogel [5], [6], for example, has described a machine which changes its structure sequentially as a function of the observed sequence. He reports that his machine achieves a cumulative average accuracy tending to one on all periodic sequences and on asymptotically lopsided sequences like the primes. (In the case of the prime number series, the predictor eventually predicts all composites.)

In the examples just given, the predictor achieves high scores on sequences which, by usual standards, seem to have some degree of order or nonrandomness. Sequences upon which high scores are achieved can be considered by definition to be orderly (with respect to the given sequential decision algorithm). The question naturally arises as to whether a predictor can achieve high scores on every sequence in some large subset of sequences which are decided a priori to be nonrandom. What prescribed sets of scores may be achieved and how may they be achieved? What internal consistencies must be present in the set of sequences yielding high scores? This paper is addressed to these questions.

2. DETERMINISTIC PREDICTORS

Consider the set of 2^n sequences $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_n)$ in $\{0, 1\}^n$. At stage k, after the observation $\Theta_1, \Theta_2, \ldots, \Theta_{k-1}$, the prediction 1 or 0 (for Θ_k) will be made with probabilities p_k and $1 - p_k$ respectively. Thus a *sequential predictor* on $\{0, 1\}^n$ will be completely specified by the set of functions $p_1, p_2(\Theta_1), p_3(\Theta_1, \Theta_2), \ldots$ $\ldots, p_n(\Theta_1, \Theta_2, \ldots, \Theta_{n-1})$, taking values in the interval $[0, 1]$. For the purposes of our discussion, we are not interested in p_i which depend on the actual outcomes of our guesses. If the p_i are restricted to the values 1 and 0, we shall say that the predictor is *deterministic*. If the p_i are independent of Θ, we shall say that the predictor is *memoryless*, and if the p_i are also independent of i, we shall say that the predictor is *constant* or time invariant.

Let $\delta = (\delta_1, \delta_2, \ldots, \delta_n) \in \{0, 1\}^n$ be the sequence of random variables resulting from the predictor $p = (p_1, p_2, \ldots, p_n)$ and the sequence $\Theta \in \{0, 1\}^n$. Then the empirical average score (the fraction of guesses correct) will be given by

$$s = \frac{1}{n} \sum_{i=1}^{n} [\delta_i \Theta_i + (1 - \delta_i)(1 - \Theta_i)] \tag{2.1}$$

and the expected empirical average score will be given by

$$\bar{s} = \mathop{\mathbf{E}}_{p}\{s\} = \frac{1}{n} \sum_{i=1}^{n} [p_i \Theta_i + (1 - p_i)(1 - \Theta_i)]. \tag{2.2}$$

The only two memoryless constant deterministic predictors are $p = (1, 1, \ldots, 1)$ and $p = (0, 0, \ldots, 0)$ which yield guessing sequences $\delta = (1, 1, \ldots, 1)$ and $\delta =$ $= (0, 0, \ldots, 0)$ and scores $s = 1/n \sum_{i=1}^{n} \Theta_i$ and $s = 1 - 1/n \sum_{i=1}^{n} \Theta_i$ respectively. Thus, there will be precisely $\binom{n}{k}$ sequences for which a score of k/n is achieved.

Time varying memoryless deterministic predictors $\delta = (\delta_1, \delta_2, \ldots, \delta_n) \in \{0, 1\}^n$ behave like constant predictors after suitable relabeling – the score on Θ depending only on the number of places in which δ and Θ differ (the Hamming distance). So in order for a memoryless deterministic predictor to achieve high prediction scores on a given subset Θ^* of the binary n-cube, Θ^* must have small Hamming diameter. Thus in the case of memoryless predictors, the set upon which high scores is obtained is easily characterized.

Finally, even for the most general deterministic predictor one cannot avoid a certain binomial distribution on the attainable scores, although the simple Hamming distance characterization of the set of high scoring Θ's is no longer possible. The following statement summarizes this. (We shall assign Roman numerals to our observations as they develop.)

I. Any sequential deterministic predictor attains a score of k/n on precisely $\binom{n}{k}$ sequences in $\{0, 1\}^n$, where $k = 0, 1, 2, \ldots, n$.

Thus no algorithm can be all things to all sequences. In fact, for any deterministic predictor, there exists a sequence upon which a score of zero with be obtained.

However, in subsequent sections we shall see that if expected score rather than actual score is the criterion for success, or if we can place bets at each stage, we can seemingly violate the "law of conservation of fairness".

3. SEQUENTIAL BETTING SYSTEMS

In this section we shall be concerned solely with deterministic sequential betting schemes. In the next section we shall suitably restrict the betting schemes in order to develop many useful statements about the behavior of the expected score of various random prediction schemes.

A series of n bets $b = (b_1, b_2, \dots, b_n)$ is made by a gambler on the outcomes of a sequence $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_n) \in \{0, 1\}^n$. The gambler's net gain at bet k is b_k if $\Theta_k = 1$ and $-b_k$ if $\Theta_k = 0$. Hence, his net winnings $w(\Theta)$ using stategy b against sequence Θ is

$$w(\Theta) = \sum_{k=1}^{n} (b_k\Theta_k - b_k(1 - \Theta_k)) = \sum_{k=1}^{n} b_k(2\Theta_k - 1), \tag{3.1}$$

where, in general, b_k will be a real valued function of Θ.

Notice that a gambler may win any preassigned amount $w(\Theta)$ if Θ is known a priori. For example, any w could be achieved with the betting system

$$\begin{aligned} b_1 &= w(\Theta)\,\Theta_1 - w(\Theta)\,(1 - \Theta_1) \\ b_2 &= b_3 = \dots = b_n = 0\,. \end{aligned} \tag{3.2}$$

However, if he knows only $\Theta_1, \Theta_2, \dots, \Theta_{k-1}$ when he must place his bet b_k, his set of achievable winnings w on $\{0, 1\}^n$ is limited. For, if $\{b_1, b_2, \dots, b_n\}$ achieves w, then manipulation of (3.1), noting the functional independence of b_k and Θ_k, yields

$$w(\Theta_1, \dots, \Theta_{n-1}, 1) + w(\Theta_1, \dots, \Theta_{n-1}, 0) = 2\sum_{k=1}^{n-1} b_k(2\Theta_k - 1) \tag{3.3}$$

and

$$w(\Theta_1, \dots, \Theta_{n-1}, 1) - w(\Theta_1, \dots, \Theta_{n-1}, 0) = 2b_n\,. \tag{3.4}$$

So, b_n is determined and (3.1) is replaced by (3.3) for the determination of b_{n-1}. Proceeding, we find

$$\sum_{\Theta} w(\Theta) = 0 \tag{3.5}$$

and

$$(3.6)\quad b_k = (\tfrac{1}{2})^{n-k+1} \sum_{(\Theta_k, \Theta_{k+1}, \ldots, \Theta_n) \in \{0,1\}^{n-k+1}} w(\Theta)\,(2\Theta_k - 1)\,, \quad k = 1, 2, \ldots, n$$

Hence, for w to be achievable by a sequential betting scheme, it is necessary and sufficient that (3.5) be satisfied. The betting scheme achieving w is given in (3.6).

Summarizing, consider a betting strategy for a game against sequences $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_n) \in \{0, 1\}^n$ which allows the bet b_k at stage k to be some element in a subset B_k of the collection B of all functions from $\{0, 1\}^n$ to R. Let $w : \{0, 1\}^n \to R$ be a desired set of net winnings defined for each sequence Θ in $\{0, 1\}^n$. As before (3.1) expresses the net winnings $w(\Theta)$ as a function of $\{b_1, b_2, \ldots, b_n\}$. Then

II. Trivially, if $B_k = B$, any w is achievable.

III. If B_k is the set of all functions in B depending only on $\Theta_1, \Theta_2, \ldots, \Theta_{k-1}$ then w is achievable if and only if (3.5) is satisfied.

IV. If, for $k = 1, 2, \ldots, n$, $B_k \subseteq B$ is the set of functions bounded in absolute value by b, depending only on $\Theta_1, \Theta_2, \ldots, \Theta_{k-1}$, then w is achievable if and only if

$$(3.7)\qquad \sum_{\Theta} w(\Theta) = 0$$

and if, for $k = 1, 2, \ldots, n$,

$$(3.8)\qquad \left| (\tfrac{1}{2})^{n-k+1} \sum_{(\Theta_k, \Theta_{k+1}, \ldots, \Theta_n) \in \{0,1\}^{n-k+1}} w(\Theta)\,(2\Theta_k - 1) \right| < b$$

for every $(\Theta_1, \Theta_2, \ldots, \Theta_{k-1}) \in \{0, 1\}^{k-1}$. This is the sequential betting scheme with bounded bet size.

We may be interested in w's which are functions only of $\sum_{i=1}^{n} \Theta_i$, the number of 1's in Θ. In this case define, for every $\Theta \in \{0, 1\}^n$,

$$(3.9)\qquad \hat{w}\left(\sum_{i=1}^{n} \Theta_i\right) = w(\Theta)\,.$$

Then the conditions of (3.7) and (3.8) become respectively

$$(3.10)\qquad \sum_{k=0}^{n} \binom{n}{k} \hat{w}(k) = 0$$

and

$$(3.11)\qquad |b_k(i)| = \left| (\tfrac{1}{2})^{n-k+1} \sum_{j=0}^{n-k} (\hat{w}(i + j + 1) - \hat{w}(i + j)) \binom{n-k}{j} \right| < b$$

for $i = 0, 1, \ldots, k - 1$ and for $k = 1, 2, \ldots, n$. Letting $k = n$ in (3.11) we have the condition

$$\tag{3.12} \tfrac{1}{2}\left|\hat{w}(i+1) - \hat{w}(i)\right| < b$$

for $i = 0, 1, \ldots, n - 1$.

Fortunately, all other conditions of (3.11) are consequences of (3.12), since, assuming (3.12) true,

$$\tag{3.13} \left|b_k(i)\right| = \left|(\tfrac{1}{2})^{n-k+1}\sum_{j=0}^{n-k}(\hat{w}(i+j+1) - \hat{w}(i+j))\binom{n-k}{j}\right| \leqq$$

$$\leqq (\tfrac{1}{2})^{n-k+1}\sum_{j=0}^{n-k}\binom{n-k}{j}2b = b\,.$$

Hence, we have

V. A terminal score $\hat{w}$ depending only on $\sum_{i=1}^{n}\Theta_i$ is achievable by a sequential betting scheme with bounded bet size b if and only if

$$\tag{3.14} \sum_{k=0}^{n}\binom{n}{k}\hat{w}(k) = 0$$

and

$$\tag{3.15} \left|\hat{w}(k+1) - \hat{w}(k)\right| < 2b\,, \quad k = 1, 2, \ldots, n - 1\,.$$

The following example illustrates the possible strange statements which may result from III.

EXAMPLE: A gambler is confronted with a sequence of 1's and 0's of length n upon which he is allowed to bet, at each stage, an amount which may be allowed to depend upon the previously observed terms of the sequence. However, he has the foreknowledge that a given sequence Θ^* will *not* appear. Can he make a net profit? The answer, of course, is yes. He can win an infinite amount. In fact, for any preassigned goal function $w(\Theta)$, he has a betting strategy which will win the amount $w(\Theta)$ against the sequence $\Theta \in \{0, 1\}^n$, for every $\Theta \neq \Theta^*$. (Put $w(\Theta^*) = -\sum_{\Theta \neq \Theta^*} w(\Theta)$ and use the betting scheme in (3.6).)

Consider for a moment the case where there is a probability distribution on the Θ's and one wishes to gamble in such a manner as to achieve a given terminal distribution F on his accumulated capital. This is the general problem faced by the gambler – that of selecting a system which yields an interesting or desirable distribution on the outcomes. Presumbaly the gambler, by investigating his needs, can choose his favorite F. Our previous development demonstrates that almost any terminal distribution on the capital is achievable by a sequential betting scheme. An example best illustrates this point:

Example: Let $\Theta_1, \Theta_2, \ldots, \Theta_n$ be independent flips of a fair coin. For any distribution function F, there exists a sequential betting scheme (of length n) achieving a terminal distribution F_n such that

$$\sup_x |F(x) - F_n(x)| < \frac{1}{2^n}. \tag{3.15}$$

To achieve F_n, in the case of continuous F, choose w_i such that $F(w_i) = i/2^n$, $i = 1, 2, \ldots, 2^n - 1$, and let

$$w_0 = -\sum_{i=1}^{2^n-1} w_i. \tag{3.16}$$

Associate the w's with the Θ's in arbitrary fashion and use the betting scheme of (3.6).

Our development demonstrates that, for flips of a fair coin, almost any probability distribution (with zero mean) on the gambler's terminal capital is achievable by a sequential betting scheme. However, most terminal distributions will be rejected by the gambler as uninteresting. Why this is so is not always clear. But we may observe the empirical fact that most popular gambling systems such as the so-called Martingale and Progression Systems are characterized by a large probability of small gain balanced by a small probability of a large loss. Presumably most gamblers have high utilities for small gains. If we allow utility functions to induce, under expectation, a partial ordering on the terminal distributions (or equivalently, the betting systems B) we find most betting systems will be inherently nonoptimal, because the 2^n points of increase of the distribution of the terminal capital due to a general $b \in B$ far exceed the few points of increase called for in the optimal betting policy of "betting the extremes". (See Chernoff and Moses [7] for a simple description of the axiomatic definition of utility and the concept of "betting the extremes".)

4. ACHIEVABLE SCORES FOR RANDOM PREDICTORS

The random predictor $p = (p_1, \ldots, p_n)$ yields an average score

$$\bar{s}(\Theta) = \frac{1}{n}\sum_{k=1}^{n}\left(p_k\Theta_k + (1 - p_k)(1 - \Theta_k)\right). \tag{4.1}$$

The characterization of the set of all achievable scores for all $p \in P$ is a consequence of the previous section under the correspondence of bets with probabilities given by

$$b_k = \frac{1}{n}(p_k - \tfrac{1}{2}). \tag{4.2}$$

The probability constraints $0 \leqq p_k \leqq 1$, $k = 1, 2, \ldots, n$, imply bounded bets

(4.3) $$|b_k| \leqq \frac{1}{2n}.$$

Thus, if $b = (b_1, \ldots, b_n)$ yields $w(\Theta)$, the corresponding predictor $p = (p_1, p_2, \ldots, p_n)$ yields score

(4.4) $$\bar{s}(\Theta) = \frac{1}{n}\sum_{k=1}^{n}((\tfrac{1}{2} + nb_k)\,\Theta_k + (\tfrac{1}{2} - nb_k)(1 - \Theta_k)) = \tfrac{1}{2} + w(\Theta).$$

We readily have from III and IV:

VI. A score $\bar{s}(\Theta)$, $\Theta \in \{0, 1\}^n$, is achievable with a sequential predictor $p \in P$ if and only if

(4.5) $$(\tfrac{1}{2})^n \sum_{\Theta} \bar{s}(\Theta) = \tfrac{1}{2}$$

and

(4.6) $$\left|(\tfrac{1}{2})^{n-k+1} \sum_{(\Theta_k, \Theta_{k+1}, \ldots, \Theta_n) \in \{0,1\}^{n-k+1}} \bar{s}(\Theta)\,(2\Theta_k - 1)\right| \leqq \frac{1}{2n}$$

for all $\Theta \in \{0, 1\}^n$ and each $k = 1, 2, \ldots, n$.

VII. A score

(4.7) $$\bar{s}(\Theta) = \hat{s}\left(\frac{1}{n}\sum_{k=1}^{n}\Theta_k\right),$$

depending only on the weight of the sequence Θ, is achievable by a sequential predictor $\hat{p} \in P$ if and only if

(4.8) $$(\tfrac{1}{2})^n \sum_{k=0}^{n} \binom{n}{k} \hat{s}\left(\frac{k}{n}\right) = \frac{1}{2}$$

and

(4.9) $$\left|\hat{s}\left(\frac{k+1}{n}\right) - \hat{s}\left(\frac{k}{n}\right)\right| < \frac{1}{n},$$

for

$$k = 0, 1, 2, \ldots, n - 1.$$

In accordance with the previous notation let the predictor at stage k given history i be denoted

(4.10) $$\hat{p}_k(i) = \mathbf{P}_r\{\delta_k = 1 \mid \sum_{j=1}^{k-1}\Theta_j = i\},$$

for $i = 1, 2, \ldots, k - 1$ and $k = 1, 2, \ldots, n$. Then, utilizing (3.6), we exhibit explicitly the predictor achieving $\hat{s}$:

$$\hat{p}_k(i) = \tfrac{1}{2} + n(\tfrac{1}{2})^{n-k+1} \sum_{j=0}^{n-k} \left(\hat{s}\left(\frac{i+j+1}{n}\right) - \hat{s}\left(\frac{i+j}{n}\right)\right) \binom{n-k}{j} \tag{4.11}$$

EXAMPLE 1. $\bar{s}(\Theta) = \frac{1}{2}$ is achieved by $p = (\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2})$.

EXAMPLE 2. $\hat{s}(k/n) = k/n \,.\, \eta + (1 - k/n)(1 - \eta)$ for $0 \leqq \eta \leqq 1$, satisfies the achievability constraints and is achieved by the constant predictor $p = (\eta, \eta, \ldots, \eta)$.

NOTE: A score $\hat{s}(k/n)$ which satisfies (4.9) but not (4.8) can be achieved with error

$$\varepsilon_n = \tfrac{1}{2} - (\tfrac{1}{2})^n \sum_{k=0}^{n} \binom{n}{k} \hat{s}\left(\frac{k}{n}\right), \tag{4.12}$$

uniformly in Θ by the p which achieves $\hat{s} + \varepsilon_n$. If $\varepsilon_n \to 0$, we shall say that $\hat{s}$ is uniformly asymptotically approachable by $\hat{p}$.

Of particular interest is the Bayes envelope

$$\max \left\{ \frac{1}{n} \sum_{i=1}^{n} \Theta_i, 1 - \frac{1}{n} \sum_{i=1}^{n} \Theta_i \right\},$$

considered by Hannan and Robbins [1], corresponding to the score function

$$\hat{s}(\eta) = \max\{\eta, 1 - \eta\}, \quad 0 < \eta < 1 . \tag{4.13}$$

$\hat{s}$ satisfies the nondegeneracy requirements of (4.8) and hence can be achieved uniformly in Θ with error

$$\varepsilon_n = \tfrac{1}{2} - (\tfrac{1}{2})^n \sum_{k=0}^{n} \binom{n}{k} \max\left\{\frac{k}{n}, 1 - \frac{k}{n}\right\} \approx \frac{-1}{\sqrt{(2\pi n)}} .$$

Thus, $\max\{\eta, 1 - \eta\}$ is uniformly asymptotically approachable. See Hannan [2] for the proof of the asymptotic approachability of the Bayes envelope for $m \times n$ matrix games.

It is natural to ask what set of score functions $\hat{s}(\eta)$, $0 \leqq \eta \leqq 1$, satisfy (4.8) and (4.9) in the limit as $n \to \infty$. Clearly the set of all continuous differentiable such functions satisfies the constraints

$$\hat{s}(\tfrac{1}{2}) = \tfrac{1}{2} \tag{4.15}$$

$$|\hat{s}(\eta)| < 1 , \quad 0 \leqq \eta \leqq 1 . \tag{4.16}$$

Paralleling Blackwell's [3], [4], definitions of approachability of sets, let us define a score function $s(\eta)$, $0 \leqq \eta \leqq 1$ to be approachable in the sense of Blackwell if and only if there exists a sequential predictor $p = (p_1, p_2, \ldots)$ on $\{0, 1\}^\infty$ such that, for every $\Theta \in \{0, 1\}^\infty$,

$$\mathbf{P}_r \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} s_i(\Theta) - s\left(\frac{1}{n} \sum_{i=1}^{n} \Theta_i\right) \right| \to 0 \right\} = 1 . \tag{4.17}$$

(Recall that $s_i(\Theta) = \delta_i \Theta_i + (1 - \delta_i)(1 - \Theta_i)$ is a random variable governed by p_i).

Blackwell reports that necessary and sufficient conditions for approachability and excludability are known only for convex sets (in his problem) and convex or concave $s(\eta)$ (in our problem). In particular, $\eta\alpha + (1 - \eta)(1 - \alpha)$, for $0 \leqq \alpha \leqq 1$, and $\max \{\eta, 1 - \eta\}$ are approachable.

We offer the conjecture that $\hat{s}$ is approachable if and only if $\hat{s}(\frac{1}{2}) = \frac{1}{2}$ and

$$\left| \hat{s}\left(\frac{k+1}{n}\right) - \hat{s}\left(\frac{k}{n}\right) \right| < \frac{1}{n},$$

for all $k = 0, 1, 2, \ldots, n - 1$ and $n = 1, 2, \ldots$.

5. CONCLUSIONS

We have used constructive procedures to establish the set of achievable winnings using sequential betting schemes and the set of achievable expected scores using sequential prediction schemes. From this construction we can conclude that no predictor can score well on all sequences, although the asymptotic achievability of

$$\bar{s}(\Theta) = \max \left\{ \frac{1}{n} \sum_{i=1}^{n} \Theta_i, 1 - \frac{1}{n} \sum_{i=1}^{n} \Theta_i \right\}$$

seems almost to violate this conclusion. However, when one is restricted to deterministic predictors, the distribution of scores on the 2^n sequences of $\{0, 1\}^n$ is strictly fixed: A score of k will be achieved on precisely $\binom{n}{k}$ sequences, $k = 0, 1, \ldots, n$. In particular, every sequence upon which a high score is achieved will have its counterpart upon which a low score is achieved.

The set of sequentially achievable winnings has been simply characterized as the set of all functions $w(\Theta) : \{0, 1\}^n \to R$ such that

$$\sum_{\Theta} w(\Theta) = 0 .$$

Thus, the gambler's problem is essentially solved, and any zero mean terminal distribution can be approximated arbitrarily closely by a sequential betting system of n bets, for n sufficiently large.

A substantial reduction in complexity of the description of achievable scores occurs when the score function $s(\Theta)$ is made to depend only on the weight of the sequence $1/n \cdot \sum_{i=1}^{n} \Theta_i$. In this case, the set of achievable scores is characterized by a Lipschitz condition with one boundary condition.

REFERENCES

[1] Hannan, James F., Robbins, Herbert: *Asymptotic Solutions of the Compound Decision Problem for Two Completely Specified Distributions*, Annals of Mathematics and Statistics, 26, pp. 37—51, 1955.

[2] Hannan, James: *Approximation to Bayes Risk in Repeated Play*, Contributions to the Theory of Games, Vol. III, Annals of Mathematics Studies No. 39, Princeton 1957.

[3] Blackwell, David: *Controlled Random Walks*, Proceedings of International Congress of Mathematicians 1954, Vol. III, Amsterdam, North Holland, 1956, pp. 336—338.

[4] Blackwell, David: *An Analog of the Minimax Theorem for Vector Payoffs*, Pacific Journal of Mathematics 6, 1956, pp. 1—8.

[5] Fogel, L. J., A. Owens, M. Walsh: *Artificial Intelligence Through a Simulation of Evolution*, Internal Report of General Dynamics/Astronautics, San Diego, California.

[6] L. J. Fogel, A. J. Owens, M. J. Walsh: *An Evolutionary Prediction Technique*, Proceedings (Summaries) of ICMCI, Tokyo, Japan 1964.

[7] Chernoff, H., Moses, L. E.: *Elementary Decision Theory*, Wiley, 1959, pp. 99—104.

STANFORD UNIVERSITY