The Blackwell Prediction Algorithm for Infinite 0-1 Sequences, and a Generalization*

H. R. Lerche[1] and J. Sarkar[2]

[1] Universität Freiburg i. Br.
[2] Indiana University-Purdue University, Indianapolis

Abstract. Let $x_1, x_2, \ldots$ be a (not necessarily random) infinite 0-1 sequence. We wish to sequentially predict the sequence. This means that, for each $n \geq 1$, we will guess the value of x_{n+1}, basing our guess on knowledge of $x_1, x_2, \ldots, x_n$. Of interest are algorithms which predict well for all 0-1 sequences. An example is the Blackwell algorithm discussed in Sect. 1. In Sect. 2 we introduce a generalization of Blackwell's algorithm to the case of three categories. This three-category algorithm will be explained using a geometric model (the so-called prediction prism), and it will be shown to be a natural generalization of Blackwell's two-category algorithm.
The Blackwell algorithm has interesting properties. It predicts arbitrary 0-1 sequences as well or better than independent, identically distributed Bernoulli variables, for which it is optimal. Such Bernoulli variables are consequently the hardest to predict. Similar results hold for the three-category generalization of Blackwell's algorithm.

1 The Blackwell Prediction Algorithm

Let $x_1, x_2, \ldots$ be an infinite 0-1 sequence. A prediction algorithm $p_1, p_2, \ldots$ is a random infinite 0-1 sequence, with p_{n+1} being the predicted value of x_{n+1}. The value of p_{n+1} may depend on $x_1, \ldots, x_n$ and also on other random variables (so-called randomizers) which are independent of the x's. Let $e_i = 1\{p_i = x_i\}$ be the indicator function of the event that the i^{th} observation x_i is correctly predicted. Let $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the relative frequency of "1" in the sequence $x_1, x_2, \ldots$ up to n, and let $\bar{e}_n = \frac{1}{n} \sum_{i=1}^{n} e_i$ be the relative frequency of correct prediction.

We next consider a plausible deterministic prediction scheme. Let

$$p_n^o = \begin{cases} 1 & \text{if } \bar{x}_n > \frac{1}{2} \\ 0 & \text{if } \bar{x}_n \leq \frac{1}{2} \end{cases} \tag{1}$$

This algorithm has both strengths and weaknesses.

If $x_1, x_2, \ldots$ is a sequence of independent Bernoulli (p) random variables, then the law of large numbers implies for $(p_n^o, n \geq 1)$ that

$$\bar{e}_n \longrightarrow \max(p, 1-p) \tag{2}$$

* We thank T. Sellke for translating and helping to revise this paper.

as $n \to \infty$ for every p, $0 \le p \le 1$. The $(p_n^o, n \ge 1)$ algorithm is asymptotically optimal for independent Bernoulli (p) variables. If the value of p is known, for example with $p > \frac{1}{2}$, then the best strategy always predicts "1" and attains $\bar{e}_n \to p$. If $p \le \frac{1}{2}$ is known, then $\bar{e}_n \to 1-p$, providing one always predicts "0". However, the deterministic algorithm (1) fails for the cyclic sequence 1, 0, 1, 0, 1, 0, ..., since there $\bar{e}_n = 0$ for $n \ge 1$.

The Blackwell algorithm does not have such weaknesses. We explain it using Fig. 1. Let

$$\mu_n = (\bar{x}_n, \bar{e}_n) \in [0,1]^2 \text{ and } S = \{(x,y) \in [0,1]^2 | y \ge \max(x, 1-x)\}.$$

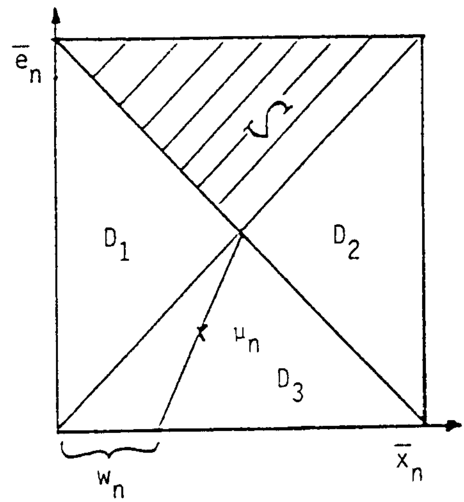

Fig. 1.

In Fig. 1, let D_1, D_2, and D_3 be the left, right, and bottom triangles, respectively, in the unit square, so that $D_1 = \{(x,y) \in [0,1]^2 | x \le y \le 1-x\}$, etc. When $\mu_n \in D_3$, draw the line through the points μ_n and $(\frac{1}{2}, \frac{1}{2})$, and let $(w_n, 0)$ be the point where this line crosses the horizontal axis. The Blackwell algorithm chooses its prediction $\bar{p}_{n+1}$ on the basis of μ_n according to the (conditional) probabilities

$$P(\bar{p}_{n+1} = 1) = \begin{cases} 0 & \text{if } \mu_n \in D_1 \\ 1 & \text{if } \mu_n \in D_2 \\ w_n & \text{if } \mu_n \in D_3. \end{cases}$$

When μ_n is in the interior of S, $\bar{p}_{n+1}$ can be chosen arbitrarily. Let $\bar{p}_1 \equiv 0$.

In what follows, d denotes Euclidean distance in $\mathbb{R}^2$, and $d(x, A)$ is the distance from the point x to the set A.

Theorem 1 *For the Blackwell algorithm applied to any infinite 0-1 sequence x_1, $x_2, \ldots$, the sequence $(\mu_n; n \geq 1)$ converges almost surely to S, i.e.*

$$d(\mu_n, S) \to 0 \text{ as } n \to \infty, \text{ almost surely.} \tag{3}$$

The conclusion of the theorem has a minimax character. The remarks following (2) show that one cannot do better than (3) for iid Bernoulli variables. For every other 0-1 sequence the Blackwell algorithm is (asymptotically) at least as successful. Consequently, iid Bernoulli variables are the hardest to predict.

The convergence behavior of the Blackwell algorithm can be explained geometrically. We view the convergence of $(\mu_n; n \geq 1)$ to S as the approach of a point sequence toward a convex set.

Case 1. μ_n is in the interior of D_1.

Here $\bar{p}_{n+1} \equiv 0$. In general we have

$$\mu_{n+1} = \mu_n + \frac{1}{n+1}(x_{n+1} - \bar{x}_n, e_{n+1} - \bar{e}_n). \tag{4}$$

Since $\bar{p}_{n+1} \equiv 0$, $(x_{n+1} - \bar{x}_n, e_{n+1} - \bar{e}_n)$ equals $(-\bar{x}_n, 1 - \bar{e}_n)$ when $x_{n+1} = 0$ and equals $(1 - \bar{x}_n, -\bar{e}_n)$ when $x_{n+1} = 1$. These two vectors are shown in Fig. 2 emanating from the point μ_n. Let $d_n = d(\mu_n, S)$. An argument using similar triangles shows that $d_{n+1} = \frac{n}{n+1} d_n$. (Note that $\mu_{n+1} \in D_1$ whenever μ_n is in the interior of D_1.)

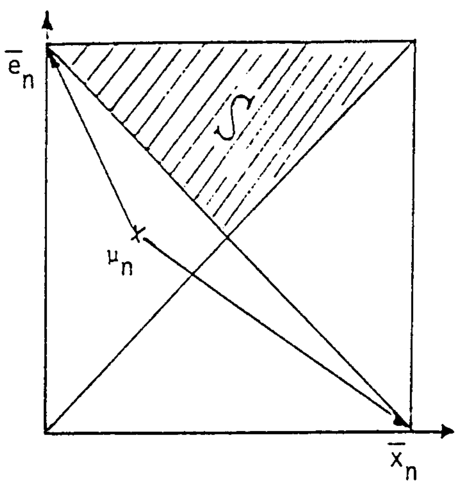

Fig. 2.

Case 2. μ_n is in the interior of D_2.

The arguments for Case 1 apply.

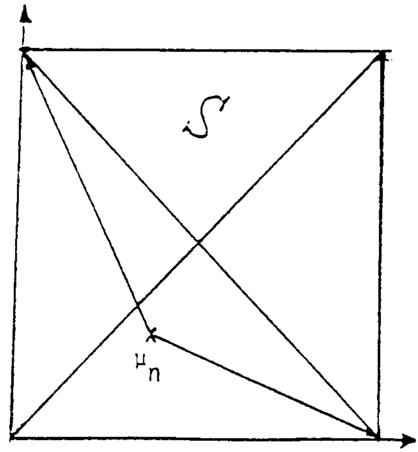

Fig. 3a.

Case 3. $\mu_n \in D_3$.

We discuss several possibilities for randomization using the following figures.
a) If we set $p_{n+1} \equiv 0$ and observe $x_{n+1} = 1$, then μ_{n+1} will be farther from S than μ_n was. See Fig. 3a.
b) Suppose the prediction of x_{n+1} is based upon tossing a fair coin.
The vectors emanating from μ_n in Fig. 3b are the conditional expectations of $(x_{n+1} - \bar{x}_n, e_{n+1} - \bar{e}_n)$ when x_{n+1} is given and e_{n+1} is random with

$$P(e_{n+1} = 1) = P(\bar{p}_{n+1} = x_{n+1}) = \frac{1}{2}.$$

One sees that when x_{n+1} takes on the "wrong" value, the conditional expected value of μ_{n+1}, given x_{n+1}, can be farther from S than μ_n is. (In Fig. 3b, the "wrong" value for x_{n+1} is 0.) Since distance from S is a convex function of $\mathbb{R}^2$, Jensen's inequality implies that the conditional expected distance (given the past and x_{n+1}) can increase when x_{n+1} takes on the "wrong" value and we predict x_{n+1} using a fair coin toss.
c) The situation is different for the Blackwell algorithm. Here we have

$$E(e_{n+1} | x_{n+1}, \text{ and past until } n) = \begin{cases} 1 - w_n & \text{if } x_{n+1} = 0 \\ w_n & \text{if } x_{n+1} = 1, \end{cases} \tag{5}$$

and the conditional expected change from μ_n to μ_{n+1} is a move toward S, provided the change is small enough. If we denote by T the line through $(\frac{1}{2}, \frac{1}{2})$ which is perpendicular to the line through $(\frac{1}{2}, \frac{1}{2})$ and μ_n, then the conditional expectation of μ_{n+1} is closer to T than μ_n was. See Fig. 3c.

Because of this orthogonality property of the Blackwell algorithm, we have

$$E(d_{n+1}^2 | \text{ past until } n) \leq \left(\frac{n}{n+1}\right)^2 d_n^2 + \frac{1}{2(n+1)^2}. \tag{6}$$

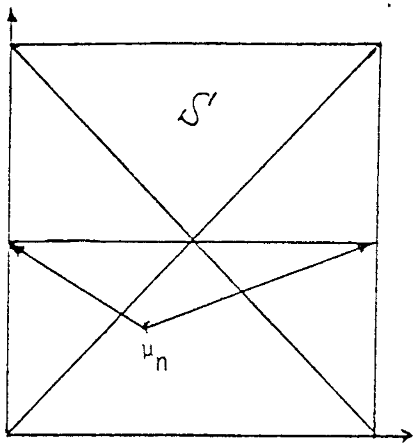

Fig. 3b.

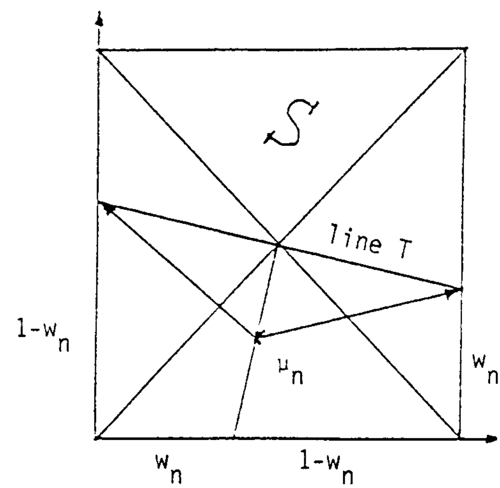

Fig. 3c.

One sees this as follows. By (4),

$$\mu_{n+1} = \frac{n}{n+1}\mu_n + \frac{1}{n+1}(x_{n+1}, e_{n+1}),$$

so that

$$d_{n+1}^2 \le d(\mu_{n+1}, (\frac{1}{2}, \frac{1}{2})) = (\bar{x}_{n+1} - \frac{1}{2})^2 + (\bar{e}_{n+1} - \frac{1}{2})^2 \tag{7}$$

$$= \left(\frac{n}{n+1}\right)^2 d_n^2 + \frac{1}{2(n+1)^2} + \frac{2n}{(n+1)^2}\{(\bar{x} - \frac{1}{2})(x_{n+1} - \frac{1}{2}) + (\bar{e}_n - \frac{1}{2})(\bar{e}_{n+1} - \frac{1}{2})\}.$$

Now take the conditional expectation in (7), with x_{n+1} given but e_{n+1} having the conditional probabilities given by (5). For either $x_{n+1} = 0$ or $x_{n+1} = 1$, the conditional expectation of the last (cross product) term in (7) vanishes, leaving us with (6).

If μ_n is in S, so that $d_n = 0$, then it is easily shown that $d_{n+1}^2 \leq \frac{1}{2(n+1)^2}$. Thus, (6) holds when $\mu_n \in S$. Since $d_{n+1} = \frac{n}{n+1} d_n$ when μ_n is in the interior of D_1 or D_2, (6) holds regardless of where μ_n is.

By (6), d_n^2 is an almost supermartingale. Theorem 1 follows from the convergence theorem for almost supermartingales in Robbins and Siegmund (1971).

2 A Three-Category Generalization of the Blackwell Algorithm

Let $x_1, x_2, \ldots$ be a sequence with values in $\{0, 1, 2\}$, and let $p_1, p_2, \ldots$ be a prediction algorithm. Let e_i and $\bar{e}_n$ be as in Sect. 1, and let $\bar{x}_n = (\bar{x}_{o,n}, \bar{x}_{1,n}, \bar{x}_{2,n})$ be the relative frequencies of the categories "0", "1", and "2". Of course $\bar{x}_{i,n} \geq 0$, $i = 0, 1, 2$, and $\bar{x}_{o,n} + \bar{x}_{1,n} + \bar{x}_{2,n} = 1$. Let $\mu_n = (\bar{x}_n, \bar{e}_n)$, and let

$$\Sigma_2 = \{(u, v, w) \in [0, 1]^3 | u + v + w = 1\}$$

denote the two-dimensional unit simplex. Also define

$$S' = \{(u, v, w, y) \in \Sigma_2 \times [0, 1] | y \geq \max(u, v, w)\}.$$

Theorem 2 *There is a generalized Blackwell algorithm for which $d(\mu_n, S') \to 0$ as $n \to \infty$, almost surely, for every infinite sequence with values in $\{0, 1, 2\}$.*

In the following, we will try to explain Theorem 2 and the generalized Blackwell algorithm. Let it first be said that the "triangle" Σ_2 is the natural domain of $\bar{x}_n$. On this triangle erect the perpendicular "success coordinate axis", so that the "prism" $\Sigma_2 \times [0, 1]$ with base Σ_2 results as the frequency-success space of μ_n. See Fig. 4.

The counterpart of the two-dimensional triangle $S = \{(x, y) | y \geq x \text{ and } y \geq 1 - x\}$ is S'. For each upper corner, cut the prism with the plane through that upper corner and the opposite bottom edge. These three cuts separate the prism into eight pieces. The top piece of the prism is S'. Figure 5 shows the prism with S' removed.

The generalized Blackwell algorithm can now be explained using the cuts just described. As before, $\bar{p}_{n+1}$ can be chosen arbitrarily when $\mu_n \in S'$. If S' is removed, seven pieces of the prism remain.

In the three pieces containing the vertical edges of the prism, one deterministically predicts the category corresponding to the vertical edge. In the pyramid at the base of the prism, one randomizes between all three categories by projecting the point $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ onto the base $(= \Sigma_2)$ through the point μ_n. The coordinates of the projection point in Σ_2 are the randomization probabilities for the corresponding categories. The three remaining pieces each touch two of the bottom corners of the prism. In these pieces, one randomizes between the two categories corresponding to these two bottom corners.

The theorem has a character similar to the theorem of Sect. 1. The hardest sequences to predict are iid trinomial variables, in the sense that for such sequences

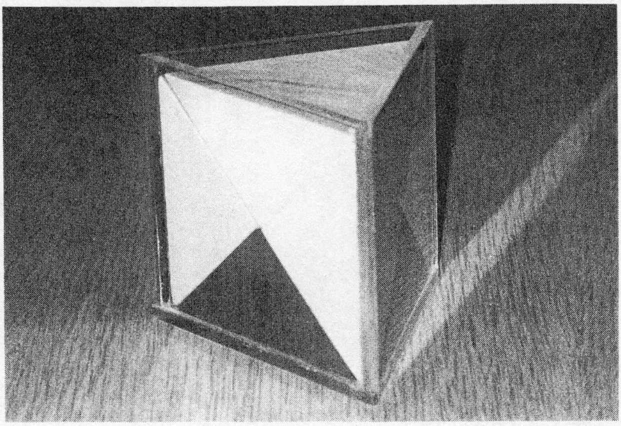

Fig. 4. $\Sigma_2 \times [0,1]$.

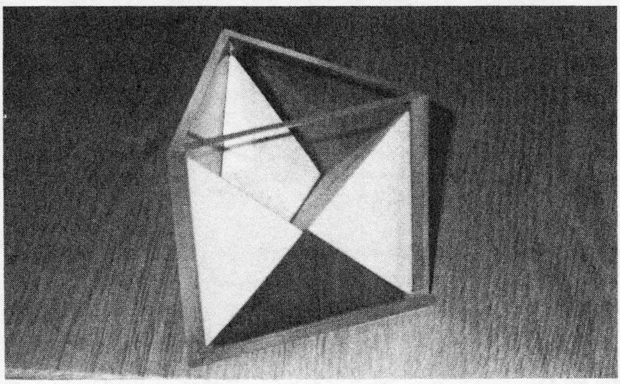

Fig. 5.

one cannot do better than to have μ_n converge to the bottom of S'. Among these sequences, the uniform distribution is the least pleasant, in that the limiting relative frequency of correct prediction is minimized at $\frac{1}{3}$.

Here is another connection with Sect. 1. If one projects the prism onto its vertical sides in the proper way, one gets the Blackwell two-category algorithm on each side.

As for the proof, one can give a geometric argument similar to that of Sect. 1. One should note, however, that the prism must be stretched by a factor of $\sqrt{3}$ in the

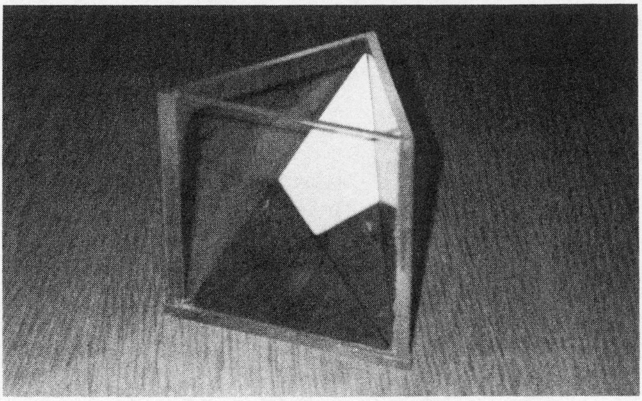

Fig. 6.

direction of the success axis to make the corresponding orthogonality relations hold. After this rescaling, $d^2(\mu_n, S')$ is an almost super martingale, so Theorem 2 follows from the same theorem of Robbins and Siegmund (1971) used in Sect. 1.

It seems to be possible to extend the procedure to more than three categories. Finally, note that one gets a puzzle with 18 pieces if the prediction prism is also cut from the bottom corners to the opposite top edges. See Fig. 7.

3 Connection with Blackwell's Theorem

Here we assume that the reader is familiar with Theorem 1 of Blackwell (1956). Instead of the prism $\Sigma_2 \times [0,1]$ and S', we consider the stretched prism

$$P'' = \{(u+y, v+y, w+y) | (u, v, w, y) \in \Sigma_2 \times [0,1]\}$$

and its subset

$$S'' = \{(u+y, v+y, w+y) | (u+y, v+y, w+y) \in P'' \text{ and } y \geq \max(u, v, w)\}.$$

The edges of the prism P'' are the possible outcomes of the game with vector payoff matrix

$$M = \begin{vmatrix} (2,1,1) & (0,1,0) & (0,0,1) \\ (1,0,0) & (1,2,1) & (0,0,1) \\ (1,0,0) & (0,1,0) & (1,1,2) \end{vmatrix}.$$

Some elementary calculations show that the "sides" of S'' which contain the point $(\frac{2}{3}, \frac{2}{3}, \frac{2}{3})$ are perpendicular to each other. This enables one to show that the assumptions of Blackwell's Theorem 1 are satisfied with S'' in place of S, which implies our Theorem 2.

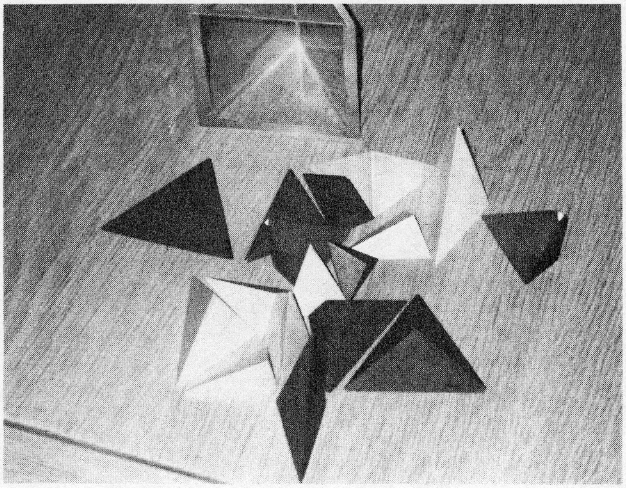

Fig. 7.

References

Blackwell, David. (1956). An analogy of the minimax theorem for vector payoffs, *Pac. Journal of Mathematics* **6**, 1–8.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*. (J. S. Rustagi, ed.) 233–257. Academic Press, New York.

This article was processed using the LaTeX macro package with LLNCS style