The Annals of Probability
1980, Vol. 8, No. 6, 1171–1176

A MARTINGALE APPROACH TO THE STUDY OF OCCURRENCE OF SEQUENCE PATTERNS IN REPEATED EXPERIMENTS[1]

BY SHUO-YEN ROBERT LI

University of Illinois at Chicago

We apply the concept of stopping times of martingales to problems in classical probability theory regarding the occurrence of sequence patterns in repeated experiments. For every finite collection of sequences of possible outcomes, we compute the expected waiting time till one of them is observed in a run of experiments. Also we compute the probability for each sequence to be the first to appear. The main result, with a transparent proof, is a generalization of some well-known facts on Bernoulli process including formulas of Feller and the "leading number" algorithm of Conway.

1. Outline and background. This paper introduces a martingale approach for studying the occurrence of sequence patterns in repeated experiments. The purpose is to unify some well-known results by a simple general theorem with a transparent proof.

Consider an experiment with only countably many outcomes. Let the experiment be performed repeatedly. Given a collection of n finite sequences of possible outcomes, we compute the expected waiting time till one of the sequences is observed in a run of experiments. Also we compute the probability for each sequence to be the first to appear. Theorem 3.1, the main result, gives a system of $n + 1$ linear equations on these $n + 1$ quantities (the expected waiting time plus one probability for each competing sequence). Each coefficient in this system describes the overlapping between two sequences and can easily be calculated. Theorem 3.3 computes the probability of tie in a certain case.

Our main machinery of computation is Doob's fundamental theorem on stopping times of martingales. Theorem 3.1 generalizes some formulas in [6], an algorithm in [4], and some theorems in [2], which are discussed below.

It is a classical problem in probability theory to study occurrence of patterns in the Bernoulli process. In the book of Feller [6], quite a few sections are devoted to the discussion on randomness and recurrent patterns connected with repeated trials. Some of the results are as follows. In the Bernoulli process with the probability p $(= 1 - q)$ of success on each trial, the probability that a run of α consecutive successes occurs before a run of β failures is $p^{\alpha-1}(1 - q^{\beta})/(p^{\alpha-1} + q^{\beta-1})$. The expected waiting time until either run occurs is $(1 - p^{\alpha})(1 - q^{\beta})/(p^{\alpha}q + pq^{\beta} - p^{\alpha}q^{\beta})$. Moreover, the expected waiting time for the pattern SSFFSS is $p^{-4}q^{-2} + q^{-2} + p^{-1}$.

In the symmetric Bernoulli process the expected waiting time for SSFFSS is 70, whereas the expected waiting time for a success run of length six is 126. This shows that, contrary to intuition, there is an essential difference in coin tossing between head runs and other patterns of the same length. Another interesting phenomenon occurs in the following coin-flip game that was first introduced in literature by Penney [8]. Given two sequence patterns of heads and tails, a coin is tossed repeatedly until one of the patterns appears as a run; that pattern wins. If the patterns are HTHT and THTT, then the odds are 9 to 5 in favor of the former, though it has expected waiting time 20 whereas the latter has only 18.

An algorithm for computing the odds of winning for the competing patterns is discovered by Conway and described by Gardner [4]. Conway's procedure is to calculate some binary

---

Received October 3, 1977; revised August 25, 1979.
[1] Research supported by NSF Grant MCS 77-03533.
AMS 1970 *subject classification.* Primary 60C05.
Key words and phrases. Leading number, martingale, stopping time, waiting time.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Probability.* www.jstor.org

numbers that he called *leading numbers* (see Section 4). One of the facts that can be deduced from this algorithm is: given a sequence of length at least three, there always exists another sequence of the same length that beats the given sequence more than half of the time. Chen and Zame [1] has a proof of this fact.

The first proof of the Conway algorithm is by Collings [3], who extends the algorithm from the coin-flip game to the game played with a balanced k-sided die instead of a coin. Wendel [10] has a generalization of the Conway algorithm to the game with n equal-lengthed competing patterns of possible outcomes of an arbitrary die.

Actually the concept of leading numbers is implicitly contained in the formula (8.7) on page 328 of [6]. Sections 13.7 and 13.8 of [6] introduce an alternative way of computing the expected waiting time and probabilities of sequence patterns, i.e., by computing their generating functions. Results from this approach, though, are usually in relatively complicated forms, but reveal more information than just the expected waiting time and probabilities of winning. In the unpublished manuscript [9] Roberts solves the game with n competing patterns of outcomes of a balanced k-sided die in terms of generating functions. Recently Guibas and Odlyzko [7] rediscover this solution from a combinatorial method and generalize it to an arbitrary die.

2. Motivation and the martingale approach. Throughout the article we employ the following notation. Let Z be an arbitrary but fixed discrete random variable. Let Σ be the set of possible values of Z. Let $Z_1, Z_2, \cdots$ be a sequence of independent, identically distributed random variables having the same distribution as Z.

If B is a finite sequence over Σ, we shall denote by N_B the waiting time until B occurs as a run in the process $Z_1, Z_2, \cdots$. Clearly, N_B is a stopping time of the process. We also consider a more general situation when a sequence is already given at the beginning of the process: Let $A = (a_1, a_2, \cdots, a_m)$ be a sequence over Σ and assume that B is not a connected subsequence of $(a_1, a_2, \cdots, a_{m-1})$. Define

$$N_{AB} = \min\{k \mid B \text{ is a connected subsequence of } (a_1, \cdots, a_m, Z_1, \cdots, Z_K)\}. \tag{2.1}$$

Again N_{AB} is a stopping time since it means the waiting time for B, given A to start with.

EXAMPLE 2.1. Let a die, which shows x, y, and z with respective probabilities ½, ⅓, and ⅙, be rolled repeatedly. Let B be the sequence (x, z, x). We want to compute EN_B.

Imagine that a gambler bets 1 dollar on the sequence B according to the following rules of fair odds. At the first roll, if x appears, he receives 2 dollars (including his bet) and must parlay the 2 dollars on the occurrence of z at the second roll. In case he wins again, he receives 12 dollars and must parlay the whole amount of 12 dollars on the occurrence of x at the third roll. If he wins three times in a row, he receives the total amount of 24 dollars and the game is over.

Now suppose that, before each roll, a new gambler joins the game and starts betting 1 dollar on the same B sequence. Say the rolls turn out to be $(y, x, x, z, y, x, x, z, x)$. Gamblers 1, 2, 3, 4, 5, 6, and 8 lose at the first, third, fifth, fourth, fifth, seventh and eighth rolls, respectively. At the ninth roll the occurrence of the B sequence ends the game and Gambler 7 receives 24 dollars. The only other winner is Gambler 9, who receives 2 dollars. In general, at the end of the game, the last gambler should receive 2 dollars and the third last gambler should receive 24 dollars. Thus, however the rolls turn out, the receipts of all the gamblers in the game total up to 26 dollars. So their net gain is $26 - N_B$. Since the odds are fair, the expected net gain should be 0. Therefore $EN_B = 26$.

Now let $A = (y, x, x, z)$ be another sequence. We want to compute EN_{AB}. Suppose that the first four rolls yield the sequence A. Right before the fifth gambler joins the game, the total fortune of the gamblers amounts to $0 + 0 + 12 + 0$ dollars. The total net gain of all the gamblers at the *subsequent* rolls will be $26 - N_{AB} - 12$. Again, since the odds are fair, we find that $EN_{AB} = 14$.

DEFINITION 2.2. Let $A = (a_1, a_2, \cdots, a_m)$ and $B = (b_1, b_2, \cdots, b_n)$ be two sequences over Σ. For every pair (i, j) of integers, write

$$\delta_{ij} = P(Z = b_j)^{-1} \quad \text{if} \quad 1 \le i \le m,\ 1 \le j \le n, \quad \text{and} \quad a_i = b_j \tag{2.2}$$
$$= 0 \qquad \text{otherwise.}$$

Then define

$$A*B = \delta_{11}\delta_{22} \cdots \delta_{mm} + \delta_{21}\delta_{32} \cdots \delta_{m,m-1} + \cdots + \delta_{m1}. \tag{2.3}$$

EXAMPLE 2.3. Let Z, A, B, be as in Example 2.1. Then $B*B = 2.6.2 + 0.0 + 2 = 26$ and $A*B = 0.0.2.0 + 2.0.0 + 2.6 + 0 = 14$.

In view of Example 2.1, the following lemma is no surprise.

LEMMA 2.4. *Given a starting sequence A, the expected waiting time for a sequence B is $EN_{AB} = B*B - A*B$, provided that B is not a connected subsequence of A. In particular the expected waiting time of the sequence B (without a starting sequence) is $B*B$.*

In order to obtain a rigorous proof of this lemma, we first quote the usual definition of martingales.

DEFINITION 2.5. A process $X_1, X_2, \cdots$ is called a martingale if, for all k, $E|X_k| < \infty$ and

$$E(X_{k+1} | X_k, \cdots, X_1) = X_k. \tag{2.4}$$

The following theorem is well known (see, for instance, Theorem 2.2 on page 302 of [3] or page 214 of [5]).

THEOREM 2.6. (Doob). *Let $X_1, X_2, \cdots$ be a martingale and N a stopping time. If $E|X_N| < \infty$ and*

$$\liminf_{k\to\infty} \int_{\{N>k\}} |X_k|\, dP = 0, \tag{2.5}$$

then $\{X_1, X_N\}$ is a martingale and hence $EX_N = EX_1$.

PROOF OF LEMMA 2.4. For every nonnegative integer k, let ω_k denote the sequence $(Z_1, \cdots, Z_k)$. Thus ω_k is a random sequence over Σ. Define the random variable.

$$X_k = \overline{A\omega_k}*B - \overline{k}, \tag{2.6}$$

where $\overline{A\omega_k}$ is the sequence A followed by the sequence ω_k.

Claim that the process $\{X_{k\wedge N_{AB}}\}_{k=0,1,2,\cdots}$ is a martingale (here '$\wedge$' stands for minimum). As before we write $A = a_1a_2 \cdots a_m$ and $B = b_1b_2 \cdots b_n$. For integers $k \ge 0$ and $j \ge 1 - m$, define

$$\begin{aligned} M_k^{(j)} &= 0, && \text{if} \quad k < j \\ &= \frac{1}{P(Z = b_1) \cdots P(Z = b_{k-j+1})} - 1, && \text{if the final } k - j + 1 \text{ terms in the sequence } \overline{A\omega_k} \text{ are identical with } b_1, \cdots, b_{k-j+1} \\ &= -1, && \text{otherwise.} \end{aligned} \tag{2.7}$$

When $k \le N_{AB}$, we may interprete the quantity $M_k^{(j)}$ as the net gain of the jth gambler at the time k in the game of fair odds described at the end of Example 2.1. This shows that, for every fixed j, the process $\{M_{k\wedge N_{AB}}^{(j)}\}_{k=0,1,2,\cdots}$ is a martingale. Since

$$\sum_{j=1-m}^{\infty} M_k^{(j)} = \sum_{j=1-m}^{k} M_k^{(j)} = \overline{A\omega_k}*B - (k + m), \tag{2.8}$$

for $k \ge N_{AB}$, we see that $\{X_{k\wedge N_{AB}}\}$ is also a martingale. Clearly

$$X_{N_{AB}} = B*B - N_{AB}. \tag{2.9}$$

Since the random variable Z assumes every value in Σ with a positive probability, N_{AB} is dominated by a geometric random variable. Therefore EN_{AB} is finite, and

$$E|X_{N_{AB}}| \leq B*B + EN_{AB} < \infty. \tag{2.10}$$

On the set $\{N_{AB} > k\}$, we have

$$\begin{aligned} |X_k| &\leq |\overline{A\omega_k}*B| + k \\ &\leq B*B + N_{AB}. \end{aligned} \tag{2.11}$$

This implies that

$$\lim_{k\to\infty} \int_{\{N_{AB}>k\}} |X_k|\, dP \leq \lim_{k\to\infty} \int_{\{N_{AB}>k\}} (B*B + N_{AB})\, dP = 0. \tag{2.12}$$

So, we can apply Theorem 2.6 and obtain that

$$EX_{N_{AB}} = EX_0 = A*B. \tag{2.13}$$

From (2.9), it follows that

$$EN_{AB} = B*B - A*B. \tag{2.14}$$

This proves the theorem. □

3. Main result. Let $A_1, A_2, \cdots, A_n$ be sequences over Σ. For each i, we want to calculate the probability that A_i precedes all the remaining $n-1$ sequences in a realization of the process $Z_1, Z_2, \cdots$. Naturally we assume that none of the sequences contain any other as a connected subsequence except possibly as a tail subsequence. As before we consider the situation when a sequence A is given at the beginning of the process. Write N_i for N_{AA_i}. Let N be the minimum among $N_1, \cdots, N_n$. We want to compute $P(N = N_j)$ for each j. In case that A_j contains A_i as a tail subsequence, removing A_j from the collection of sequences does not affect $P(N = N_k)$ for all $k \neq j$. So, to avoid ties, we assume in the next theorem that no sequence contains another, the probability of a tie in a special case being considered in a later theorem (3.3).

THEOREM 3.1. *Let $Z, Z_1, Z_2, \cdots$ be discrete* i.i.d. *random variables and $A_1, \cdots, A_n$ be finite sequences of possible values of Z not containing one another. Let A be another such sequence not containing any A_i. Given the starting sequence A, let p_i be the probability that A_i precedes the remaining $n-1$ sequences in a realization of the process $Z_1, Z_2, \cdots$. Then for every i,*

$$\sum_{j=1}^{n} p_j A_j*A_i = EN + A*A_i, \tag{3.1}$$

where N is the stopping time when any A_j occurs and A_j*A_i is defined by (2.3). *In particular when A is void, we have, for every i,*

$$\sum_{j=1}^{n} p_j A_j*A_i = EN. \tag{3.2}$$

PROOF. Write N_i for N_{AA_i}. Then N is the miminum among $N_1, \cdots, N_n$. We have

$$\begin{aligned} EN_i &= EN + E(N_i - N) \\ &= EN + \sum_{j=1}^{n} p_j E(N_i - N \mid N = N_j). \end{aligned} \tag{3.3}$$

From Lemma 2.4 we also know that

$$EN_i = A_i*A_i - A*A_i \tag{3.4}$$

and

$$E(N_i - N \mid N = N_j) = A_i{*}A_i - A_j{*}A_i. \tag{3.5}$$

Substituting (3.4) and (3.5) in (3.3), we have

$$\begin{aligned} A_i{*}A_i - A{*}A_i &= EN + \textstyle\sum_{j=1}^{n} p_j(A_i{*}A_i - A_j{*}A_i). \\ &= EN + A_i{*}A_i - \textstyle\sum_{j=1}^{n} p_j A_j{*}A_i. \end{aligned} \tag{3.6}$$

This proves the theorem. □

In the matrix form, the equalities (3.1) become

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ -1 & A_1{*}A_1 & A_2{*}A_1 & \cdots & A_n{*}A_1 \\ -1 & A_1{*}A_2 & A_2{*}A_2 & \cdots & A_n{*}A_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -1 & A_1{*}A_n & A_2{*}A_n & \cdots & A_n{*}A_n \end{bmatrix} \begin{bmatrix} EN \\ p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} 1 \\ A{*}A_1 \\ A{*}A_2 \\ \vdots \\ A{*}A_n \end{bmatrix}. \tag{3.7}$$

Let M denote the coefficient matrix in this system. We can solve for the values of EN and p_j provided that M is nonsingular. An equivalent condition is the nonsingularity of the n by n submatrix $(A_j{*}A_i - A{*}A_i)$. This equivalence is because of the identity

$$EN \cdot \det M = \det(A_j{*}A_i - A{*}A_i), \tag{3.8}$$

which is obtained by elementary row operations followed by the application of Cramér's rule. When the matrix m is singular, we need a limiting process to compute p_j. (Since this paper was written, Gerber has proved the nonsingularity of the matrix $(A_j{*}A_i - A{*}A_i)$. Thus the limiting process in the next paragraph becomes superfluous.)

Imagine that the distribution of Z is unknown and $P(Z = k)$ is represented by z_k^{-1} for every $k \in \Sigma$. Then $A_j{*}A_i$ is a polynomial in the formal variables z_k. If $i \neq j$, then $A_i{*}A_i$ is a polynomial of higher degree than $A{*}A_i$ and $A_j{*}A_i$. So the determinant of the n by n matrix $(A_j{*}A_i - A{*}A_i)$ is a nonzero polynomial in z_k, $k \in \Sigma$. Again, by Cramér's rule, $p_1, \cdots, p_n$ are rational functions in z_k. For a random variable Z with known distribution, the probabilities p_j can be obtained from L'Hospital's rule by differentiating numerator and denominator. This limiting process is similar to the treatment on page 328 of [6].

COROLLARY 3.2. *If A, B are not connected subsequences of each other, then the odds that the sequence B precedes the sequence A in a realization of the process $Z_1, Z_2, \cdots$ are*

$$(A{*}A - A{*}B) : (B{*}B - B{*}A).$$

THEOREM 3.3. *Let $A = (a_1 \cdots a_m)$ be a sequence over Σ and B be the subsequence $(a_2 \cdots a_m)$. Then the probability that A and B occur at the same time in a realization of the process $Z_1, Z_2, \cdots$ is*

$$P(N_A = N_B) = \frac{B{*}B - B{*}A}{A{*}A - A{*}B - B{*}A + B{*}B}. \tag{3.9}$$

PROOF. Write $\omega_k = (Z_1, \cdots, Z_k)$ for all $k \geq 0$. As in the proof of Lemma 2.4, we know

$$E[\omega_{N_B}{*}B - N_B] = 0. \tag{3.10}$$

Similarly,

$$E[\omega_{N_B}{*}A - N_B] = 0. \tag{3.11}$$

The value of $\omega_{N_B}{*}A - \omega_{N_B}{*}B$ is either $A{*}A - A{*}B$ or $B{*}A - B{*}B$ according as $N_B = N_A$ or $N_B \neq N_A$. Therefore,

$$\begin{aligned} 0 &= E[\omega_{N_B}{*}A - \omega_{N_B}{*}B] \\ &= (B{*}A - B{*}B) \cdot P(N_A = N_B) + (A{*}A - A{*}B) \cdot P(N_A \neq N_B). \end{aligned} \tag{3.12}$$

This implies (3.9). □

4. Conway's leading-number algorithm. Given two sequences $A = (a_1, \cdots, a_m)$ and $B = (b_1, \cdots, b_n)$, define the *leading number* of A over B as a binary integer $\epsilon_n\epsilon_{n-1}\cdots\epsilon_1$ via

$$\begin{aligned} \epsilon_i &= 1 \quad \text{if } 1 \le i \le \min\{m, n\} \text{ and the two sequences} \\ &\quad\ (a_{m-i+1}, \cdots, a_m) \text{ and } (b_1, \cdots, b_i) \text{ are} \\ &\quad\ \text{identical,} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{4.1}$$

The following theorem was once considered as a somewhat magic algorithm.

THEOREM 4.1. (Conway). *Let AB denote the leading number of A over B. Then the odds for B to precede A in a symmetric Bernoulli process are*

$$(AA - AB) : (BB - BA).$$

It is easy to see that if the random variable Z assumes just two values with equal probabilities, then $A*B$ reduces to 2 times the leading number of A over B. Thus Theorem 4.1 is a special case of Corollary 3.2. Incidentally, the quantity $A*B$ can be expressed in terms of ϵ_i as follows.

$$\begin{aligned} A*B = [[\cdots[[\epsilon_n \cdot P(Z = b_n)^{-1} + \epsilon_{n-1}] \cdot P(Z = b_{n-1})^{-1} + \epsilon_{n-2}] \\ \cdot P(Z = b_{n-2})^{-1} + \cdots] + \epsilon_1] \cdot P(Z = b_1)^{-1}. \end{aligned} \tag{4.2}$$

Acknowledgment. The author is grateful to R. Chen, S. T. Huang, A. Odlyzko, J. G. Wendel and the referee for their helpful comments.

REFERENCES

[1] CHEN, R. and ZAME, A. (1977). A remark on fair coin-tossing process. *Bull. Inst. Math. Statist.* **6** 278.
[2] COLLINGS, S. (1977). Improbable probabilities. Unpublished manuscript.
[3] DOOB, J. L. (1953). *Stochastic Processes.* Wiley, New York.
[4] GARDNER, M. (1974). Mathematical games. *Sci. Amer.* **10** 120–125.
[5] FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications, Vol.* II. Wiley, New York.
[6] FELLER, W. (1967). *An Introduction to Probability Theory and Its Applications, Vol.* I, *3rd ed.* Wiley, New York.
[7] GUIBAS, L. and ODLYZKO, A. (1978). String overlaps, pattern matching, and nontransitive games. Unpublished manuscript.
[8] PENNEY, W. (1969). Problem: Penney-ante. *J. Recreational Math.* **2** 241.
[9] ROBERTS, S. W. (1963). Unpublished manuscript.
[10] WENDEL, J. G. (1977). Private communication.

BELL LABORATORIES, ROOM 6N-340
WARRENVILLE-NAPERVILLE RD.
NAPERVILLE, IL 60540