

GAMBLING TEAMS AND WAITING TIMES FOR PATTERNS IN TWO-STATE MARKOV CHAINS

JOSEPH GLAZ,* *University of Connecticut*

MARTIN KULLDORFF,** *Harvard Medical School and Harvard Pilgrim Health Care*

VLADIMIR POZDNYAKOV,* *** *University of Connecticut*

J. MICHAEL STEELE,**** *University of Pennsylvania*

Abstract

Methods using gambling teams and martingales are developed and applied to find formulae for the expected value and the generating function of the waiting time to observation of an element of a finite collection of patterns in a sequence generated by a two-state Markov chain of first, or higher, order.

Keywords: Gambling; team; waiting time; pattern; success run; failure run; Markov chain; martingale; stopping time; generating function

2000 Mathematics Subject Classification: Primary 60J10
Secondary 60G42

1. Introduction

How long must we observe a stochastic process with values from a finite alphabet until we see a realization of a pattern that belongs to a specified collection, \mathcal{C} , of possible patterns? For independent processes this is an old question; in some special cases it was even considered by Feller (1968, pp. 326–328). Nevertheless, in the context of more general processes, or even for Markov chains, there are many natural problems that have not been fully addressed. The main goal here is to show how some progress can be made by further developing the martingale methods that were introduced by Li (1980) and Gerber and Li (1981) in their investigation of independent sequences. Their key observation was that information on the occurrence times of patterns can be obtained from the values assumed by a specially constructed auxiliary martingale at a certain well-chosen time.

In the case of Markov chains of first, or higher, order, this observation is still useful, but to utilize it requires a rather more elaborate plan for the construction of the auxiliary martingale. This construction depends in turn on several general devices that seem more broadly useful; these include ‘teams of gamblers,’ ‘watching then betting,’ ‘reward matching,’ and a couple of other devices that will be described shortly.

Before giving that description, we should note that pattern matching has been studied using many other techniques. The combinatorial methods of Guibas and Odlyzko (1981a), (1981b)

Received 4 April 2003; revision received 14 November 2005.

* Postal address: Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269-4120, USA.

** Postal address: Department of Ambulatory Care and Prevention, Harvard Medical School, 133 Brookline Avenue, Boston, MA 02215-3920, USA.

*** Email address: vladimir.pozdnyakov@uconn.edu

**** Postal address: Department of Statistics, Wharton School, Huntsman Hall 447, University of Pennsylvania, Philadelphia, PA 19104, USA.

are particularly effective, and there are numerous treatments of pattern matching problems using probabilistic techniques, such as Benevento (1984), Biggins and Cannings (1987a), (1987b), Blom and Thorburn (1982), Breen *et al.* (1985), Chryssaphinou and Papastavridis (1990), Han and Hirano (2003), Pozdnyakov *et al.* (2005), Pozdnyakov and Kulldorff (2006), Robin and Daudin (1999), Stefanov (2003), and Uchida (1998). One of the more general techniques is the Markov chain embedding method introduced by Fu (1986), which has been further developed by Antzoulakos (2001), Fu (2001), Fu and Chang (2002), and Fu and Koutras (1994). The approach of Stefanov (2000) and Stefanov and Pakes (1997) also uses Markov chain embedding, though their method differs substantially from that of Fu. Only a few investigations have considered waiting time problems for higher-order Markov chains, and these have all focused on specific waiting times such as the ‘sooner or later’ problem studied by Aki *et al.* (1996).

2. Expected waiting time until a pattern is observed

We take $\{Z_n, n \geq 1\}$ to be a Markov chain with two states, S and F , which may model ‘success’ and ‘failure.’ We suppose the chain to have the initial distribution with $P(Z_1 = S) = p_S$ and $P(Z_1 = F) = p_F$, and the transition matrix

$$\begin{pmatrix} p_{SS} & p_{FS} \\ p_{SF} & p_{FF} \end{pmatrix},$$

where p_{SF} is shorthand for $P(Z_{n+1} = F \mid Z_n = S)$. We then consider a collection, \mathcal{C} , of finite sequences A_i , $1 \leq i \leq K$, from the two-letter alphabet $\{S, F\}$. If τ_{A_i} denotes the time elapsed before the pattern A_i is first observed as a completed run in the series Z_1, Z_2, \dots , then the random variable of main interest here is $\tau_{\mathcal{C}} = \min\{\tau_{A_1}, \dots, \tau_{A_K}\}$, the time at which we first observe a pattern from \mathcal{C} . Throughout our discussion we assume that no pattern from \mathcal{C} contains another pattern from \mathcal{C} as an initial segment. Naturally, this assumption entails no loss of generality.

2.1. A run of ‘failures’ under a Markov model

To illustrate the construction, we first consider the rather easy case in which $K = 1$ and the pattern A_1 is a run of r consecutive F s. Thus, we will compute the expected value of $\tau = \tau_{A_1}$, the time of the first completion of a run of r ‘failures’ under our two-state Markov model. This example can be treated in several ways, and offers a useful benchmark for more challenging examples.

We consider a casino where gamblers may bet in successive rounds on the output of our given two-state Markov chain, and assume that the casino is *fair* in a sense that we will soon make precise. We then consider a sequence of gamblers, one of whom arrives just before each new round of betting. Thus, gambler $n + 1$ arrives in time to observe the result of the n th trial, Z_n , and we assume that he bets a dollar on the event that the next trial yields an F . If $Z_{n+1} = S$, he loses his dollar and leaves the game. If he is lucky, and $Z_{n+1} = F$, then he wins $1/p_{SF}$ when $Z_n = S$ and wins $1/p_{FF}$ when $Z_n = F$. This is the sense in which the casino is fair; the expected return on a one-dollar bet is one dollar.

After this gambler collects his money, he then bets his entire capital on the event that $Z_{n+2} = F$. Again, if $Z_{n+2} = S$ then the gambler leaves the game with nothing, while if $Z_{n+2} = F$ then the gambler wins this round and his capital is increased by a factor $1/p_{FF}$. Successive rounds proceed in the same way, with a new gambler arriving at each new round and with the gamblers from earlier periods either continuing to win or else going broke and leaving.

We need to be precise about how this process ends. If gambler $n + 1$ begins by observing $Z_n = S$, then he bets until he either goes broke or observes r successive F s, and if he begins by observing $Z_n = F$, then he bets until he either goes broke or observes $r - 1$ successive F s. Once some gambler stops without going broke, all of the gambling stops.

Finally, we let X_n denote the casino's net gain at the conclusion of round n . Since each bet is fair and since the bet sizes depend only on the previous observations, the sequence X_n is a martingale with respect to the σ -field generated by $\{Z_n, n \geq 1\}$. Now, to obtain a relation that will determine $E[\tau]$, we just need to consider the casino's net gain when the gambling stops, X_τ . By calculating $E[X_\tau]$ in two ways we will then obtain the expected value of the time τ .

At time τ , many gamblers are likely to have lost all their money; only those who entered the game after round number $\tau - r - 1$ have any money. We thus face two different ending scenarios. First, it could happen that we have a block (denoted by $F^{(r)}$) of r instances of F that occurs at the very beginning of the sequence $\{Z_n, n \geq 1\}$. Second, it could happen that we end with $SF^{(r)}$, an S followed by a block of r instances of F . Obviously we do not need to consider the possibility of ending with $FF^{(r)} = F^{(r+1)}$, since by definition $F^{(r)}$ cannot occur before time τ .

When we sum the wins and losses of all of the gamblers, we then find that the value of the stopped martingale X_τ is exactly given by

$$X_\tau = \begin{cases} \tau - 1 - \frac{1}{p_{FF}^{r-1}} - \frac{1}{p_{FF}^{r-2}} - \dots - \frac{1}{p_{FF}} & \text{in the first scenario,} \\ \tau - 1 - \frac{1}{p_{SF} p_{FF}^{r-1}} - \frac{1}{p_{FF}^{r-1}} - \frac{1}{p_{FF}^{r-2}} - \dots - \frac{1}{p_{FF}} & \text{in the second scenario,} \end{cases}$$

which can be written more succinctly as

$$X_\tau = \begin{cases} \tau - 1 - \frac{1 - p_{FF}^{r-1}}{p_{FF}^{r-1}(1 - p_{FF})} & \text{in the first scenario,} \\ \tau - 1 - \frac{1}{p_{SF} p_{FF}^{r-1}} - \frac{1 - p_{FF}^{r-1}}{p_{FF}^{r-1}(1 - p_{FF})} & \text{in the second scenario.} \end{cases}$$

Since $E[\tau] < \infty$ and the increments of X_n are bounded, the optional stopping theorem for martingales (see, e.g. Williams (1991, p. 100)) tells us that $0 = E[X_1] = E[X_\tau]$. From this identity and the formula for X_τ , algebraic simplification gives us

$$E[\tau] = 1 + p_F \frac{1 - p_{FF}^{r-1}}{1 - p_{FF}} + (1 - p_F p_{FF}^{r-1}) \left(\frac{1}{p_{SF} p_{FF}^{r-1}} + \frac{1 - p_{FF}^{r-1}}{p_{SF} p_{FF}^{r-1}} \right).$$

2.2. Second step: a single pattern

We now consider the more subtle case of a single (nonrun) pattern A of length r , and for specificity we assume that the pattern begins with an F ; hence, $A = FB$ with $B \in \{S, F\}^{r-1}$. As before, we consider a sequence of gamblers, but this time we need to consider three different ending scenarios:

1. A occurs at the beginning of the sequence $\{Z_n, n \geq 1\}$.
2. The pattern SA occurs.
3. The pattern FA occurs.

The probability, p_1 , of the first scenario is trivial to compute, but we then face a problem. We do not know the probability of the pattern SA appearing earlier than the pattern FA , so the probabilities of the second and third ending scenarios are not readily computable. To circumvent this problem, we introduce two teams of gamblers.

2.3. Rules for the gambling teams

1. A gambler from the first team who arrives before round n watches the result of the n th trial and then bets y_1 dollars on the first letter in the sequence A . If he wins he then bets all of his capital on the next letter in the sequence A , and continues in this way until he either loses his capital or observes all of the letters of A . Such players are called ‘straightforward gamblers.’

2. The gamblers of the second team make use of the information that they gather. On the one hand, if gambler $n + 1$ observes that $Z_n = S$ just before he begins his play, then he bets like a straightforward gambler, except that he begins by wagering y_2 dollars on the first letter of pattern A . On the other hand, if he observes that $Z_n = F$ when he first arrives, then he wagers y_2 dollars on the first letter of the pattern B . He then continues to wager on the successive letters of B until either he loses or he observes B . Such players are called ‘smart gamblers.’

The two gambling teams continue their betting until one team stops. At that time, all gambling stops and we consider the wins and losses. Only those gamblers who entered the game after the time $\tau - r - 1$ will have any money, and the amount they have will depend on the ending scenario. If we let $W_{ij}y_j$ denote the amount of money that team $j \in \{1, 2\}$ wins in scenario $i \in \{1, 2, 3\}$, then the values W_{ij} are easy to compute and, in terms of these values, the stopped martingale X_τ , which represents the casino’s net gain, is given by

$$X_\tau = \begin{cases} (y_1 + y_2)(\tau - 1) - y_1 W_{11} - y_2 W_{12} & \text{in the first scenario,} \\ (y_1 + y_2)(\tau - 1) - y_1 W_{21} - y_2 W_{22} & \text{in the second scenario,} \\ (y_1 + y_2)(\tau - 1) - y_1 W_{31} - y_2 W_{32} & \text{in the third scenario.} \end{cases}$$

Now, if we take (y_1^*, y_2^*) to be a solution to the system

$$y_1^* W_{21} + y_2^* W_{22} = 1, \quad y_1^* W_{31} + y_2^* W_{32} = 1,$$

we see that, with these bet sizes, we have a very simple formula for X_τ :

$$X_\tau = \begin{cases} (y_1^* + y_2^*)(\tau - 1) - y_1^* W_{11} - y_2^* W_{12} & \text{in the first scenario,} \\ (y_1^* + y_2^*)(\tau - 1) - 1 & \text{in the second scenario,} \\ (y_1^* + y_2^*)(\tau - 1) - 1 & \text{in the third scenario.} \end{cases}$$

The optional stopping theorem then gives us

$$0 = (y_1^* + y_2^*)(E[\tau] - 1) - p_1(y_1^* W_{11} + y_2^* W_{12}) - (1 - p_1),$$

where p_1 is the probability of scenario one occurring. We therefore find that

$$E[\tau] = 1 + \frac{p_1(y_1^* W_{11} + y_2^* W_{12}) + 1 - p_1}{y_1^* + y_2^*}. \tag{1}$$

Formula (1) is more explicit than it may at first appear. In the typical case, the calculation of $p_1, \{W_{ij}, 1 \leq i \leq 3, 1 \leq j \leq 2\}$, and $\{y_j^*, 1 \leq j \leq 2\}$ is genuinely routine, as can be seen in the following example.

2.4. Example: waiting time until FSF is observed

Here, our straightforward gamblers bet y_1 dollars on FSF without regard for the preceding observation, while the smart gamblers bet y_2 dollars on FSF if they observed S before placing their first bet, and bet y_2 dollars on SF if they observed F . The three ending scenarios are now the occurrence of FSF at the beginning (scenario one), the occurrence of $SFSF$ at the end (scenario two), or the occurrence of $FFSF$ at the end (scenario three). The 3×2 ‘profit matrix’ $\{W_{ij}\}$ is then given by

$$\begin{pmatrix} \frac{1}{PSF} & \frac{1}{PFSPSF} + \frac{1}{PSF} \\ \frac{1}{PSFPFSPSF} + \frac{1}{PSF} & \frac{1}{PSFPFSPSF} + \frac{1}{PFSPSF} + \frac{1}{PSF} \\ \frac{1}{PFFPFSPSF} + \frac{1}{PSF} & \frac{1}{PFSPSF} + \frac{1}{PSF} \end{pmatrix},$$

and the bet sizes y_1^* and y_2^* are determined by the relations

$$\begin{aligned} y_1^* \left(\frac{1}{PSFPFSPSF} + \frac{1}{PSF} \right) + y_2^* \left(\frac{1}{PSFPFSPSF} + \frac{1}{PFSPSF} + \frac{1}{PSF} \right) &= 1, \\ y_1^* \left(\frac{1}{PFFPFSPSF} + \frac{1}{PSF} \right) + y_2^* \left(\frac{1}{PFSPSF} + \frac{1}{PSF} \right) &= 1, \end{aligned}$$

which can be solved to obtain

$$y_1^* = \frac{PFFPFSPSF}{PFS + PSF + PFSPSF} \quad \text{and} \quad y_2^* = \frac{PFSPSF(PSF - PFF)}{PFS + PSF + PFSPSF}.$$

The probability, p_1 , of the first scenario is just $p_F p_{FS} p_{SF}$, so after substitution and simplification the general formula (1) yields

$$E[\tau_{FSF}] = 1 + \frac{PS}{PSF} + \frac{1}{p_{SF}^2} + \frac{1}{PFSPSF},$$

which is as explicit as could be desired.

3. Expected time until observing one of many patterns

We now consider a collection, $\mathcal{C} = \{A_i, 1 \leq i \leq K\}$, of K strings of possibly varying lengths from the two-letter alphabet $\{S, F\}$, and we take on the task of computing the expected value of $\tau_{\mathcal{C}} = \min\{\tau_{A_1}, \dots, \tau_{A_K}\}$, the time at which any one of the patterns in \mathcal{C} is first observed. The method we propose is analogous to the two-team method we used in the previous section, although many teams are now needed. The real challenge is the construction of the list of appropriate ending scenarios, which now requires some algorithmic techniques.

3.1. Listing the ending scenarios

Given the collection $\mathcal{C} = \{A_i\}_{1 \leq i \leq K}$, we first consider the set sequence transformation

$$\mathcal{C} = \{A_i\}_{1 \leq i \leq K} \rightarrow \{SA_i, FA_i\}_{1 \leq i \leq K} = \{B_i\}_{1 \leq i \leq 2K} = \mathcal{C}'$$

which doubles the cardinality of \mathcal{C} . We then delete from \mathcal{C}' each pattern B that can occur only after the stopping time $\tau_{\mathcal{C}}$. The resulting collection, \mathcal{C}'' , is called the ‘final list.’ We denote the elements of \mathcal{C}'' by $C_i, 1 \leq i \leq K'$, and note that $K \leq K' \leq 2K$.

To illustrate the construction, suppose that the initial collection is $\mathcal{C} = \{FSF, FF\}$. The ‘doubling step’ (from \mathcal{C} to \mathcal{C}') gives us $\mathcal{C}' = \{SFSF, FFSF, SFF, FFF\}$. Since FFS and FFF cannot occur before $\tau_{\mathcal{C}}$, these are eliminated from \mathcal{C}' and the final list is simply $\mathcal{C}'' = \{SFSF, SFF\}$. Similarly, if the initial collection is $\mathcal{C} = \{FS, SSS\}$, then the final list is $\mathcal{C}'' = \{SFS, FFS\}$.

Now, before we describe the ending scenarios, we need one further notion. If patterns C and C' in the final list \mathcal{C}'' satisfy $C = SA$ and $C' = FA$ for some pattern $A \in \mathcal{C}$, then we say that C and C' are ‘matched.’ Also, if C and C' are matched and $C = SA$ and $C' = FA$, then we say that C and C' are ‘generated’ by A . Finally, even though there are many ending scenarios, they may be classified into three basic kinds.

1. There are K scenarios in which we observe an element of \mathcal{C} to be an initial segment of the Markov sequence $\{Z_n, n \geq 1\}$.
2. There is a scenario for each unmatched pattern from \mathcal{C}'' . We denote the number of these by L .
3. There is pair of scenarios for each matched pattern from \mathcal{C}'' . We denote the number of these by $2M$.

3.2. From the listing to the teams

For each scenario associated with the unmatched pattern C_j , we introduce *one* team of straightforward gamblers who bet y_j dollars on the pattern A_i that generated C_j . For each pair of scenarios associated with the matched patterns C_p and C_m , which were generated by pattern A_k , say, we introduce *two* teams. One team bets y_p dollars on A_k in the straightforward way, and the other bets y_m dollars on A_k in the smart way. If $W_{ij}y_j, i = 1, 2, \dots, K + L + 2M, j = 1, 2, \dots, L + 2M$, denotes the amount of money that the j th team wins in the i th scenario, then the stopped martingale X_τ is given by the sum

$$X_\tau = \sum_{j=1}^{L+2M} y_j(\tau - 1) - S(y_1, \dots, y_{L+2M}),$$

where

$$S(y_1, \dots, y_{L+2M}) = \sum_{i=1}^{K+L+2M} \mathbf{1}_{E_i} \sum_{j=1}^{L+2M} y_j W_{ij}.$$

Here, $\mathbf{1}_{E_i}$ is the indicator function for the event, E_i , that the i th scenario occurs.

If $(y_1^*, \dots, y_{L+2M}^*)$ is a solution to the linear system

$$\begin{aligned} y_1^* W_{K+1,1} + \dots + y_{L+2M}^* W_{K+1,L+2M} &= 1, \\ &\vdots \\ y_1^* W_{K+L+2M,1} + \dots + y_{L+2M}^* W_{K+L+2M,L+2M} &= 1, \end{aligned} \tag{2}$$

then we have

$$S(y_1^*, \dots, y_{L+2M}^*) = \begin{cases} \sum_{j=1}^{L+2M} y_j^* W_{ij} & \text{in scenario } i \in \{1, 2, \dots, K\} \\ 1, & \text{in scenario } i > K. \end{cases}$$

By the optional stopping theorem, we have

$$0 = E[X_1] = E[X_{\tau_C}] = \sum_{j=1}^{L+2M} y_j^* (E[\tau_C] - 1) - \sum_{i=1}^K p_i \sum_{j=1}^{L+2M} y_j^* W_{ij} - \left(1 - \sum_{i=1}^K p_i\right),$$

where p_i is the probability that A_i is an initial segment of $\{Z_n, n \geq 1\}$. We can now solve this equation to obtain a formula for $E[\tau_C]$, which we summarize as a theorem as follows.

Theorem 1. *If $(y_1^*, y_2^*, \dots, y_{L+2M}^*)$ solves the linear system (2), then*

$$E[\tau_C] = 1 + \frac{\sum_{i=1}^K p_i \sum_{j=1}^{L+2M} y_j^* W_{ij} + (1 - \sum_{i=1}^K p_i)}{\sum_{j=1}^{L+2M} y_j^*}. \tag{3}$$

As before, this formula is more explicit than it may at first appear. For example, all of the required terms can be computed straightforwardly in problems of reasonable size.

3.3. Computation of the profit matrix

Formula (3) requires the computation of the profit matrix $\{W_{ij}\}$. We will first show how this can be done in general. The method will then be applied to a specific example to confirm that (3) may be rewritten in terms of the basic model parameters.

Consider a scenario that ends with pattern $C = c_1 c_2 \dots c_m \in \mathcal{C}''$. The team of straightforward gamblers who begin by betting one dollar and continue by betting on the successive terms of the pattern $A = a_1 a_2 \dots a_p$ will, by time τ_C , have won an amount

$$\sum_{i=1}^{\min\{m-1, p\}} \delta_i^{\text{st}}(A, C),$$

where

$$\delta_i^{\text{st}}(A, C) = \begin{cases} \frac{1}{p_{c_{m-i} a_1} p_{a_1 a_2} \dots p_{a_{i-1} a_i}} & \text{if } a_1 = c_{m-i+1}, a_2 = c_{m-i+2}, \dots, a_i = c_m, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, in the same time, the team of smart gamblers will have won an amount

$$\sum_{i=1}^{\min\{m-1, p\}} \delta_i^{\text{sm1}}(A, C) + \sum_{i=1}^{\min\{m-1, p-1\}} \delta_i^{\text{sm2}}(A, C),$$

where

$$\delta_i^{\text{sm}1}(A, C) = \begin{cases} \frac{1}{p_{c_{m-i}a_1} p_{a_1 a_2} \cdots p_{a_{i-1} a_i}} & \text{if } a_1 = c_{m-i+1}, a_2 = c_{m-i+2}, \dots, a_i = c_m, \\ & \text{and } c_{m-i} \neq a_1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\delta_i^{\text{sm}2}(A, C) = \begin{cases} \frac{1}{p_{a_1 a_2} p_{a_2 a_3} \cdots p_{a_i a_{i+1}}} & \text{if } a_1 = c_{m-i}, a_2 = c_{m-i+1}, \dots, a_{i+1} = c_m, \\ 0 & \text{otherwise.} \end{cases}$$

3.4. Explicit determination of $E[\tau_e]$

To illustrate the use of (3), we consider the collection $\mathcal{C} = \{SS, FFSF\}$. After doubling and elimination we obtain the final list $\{FSS, SF SF, FFSF\}$. We then need to work out the set of scenarios: we have two scenarios in which either $C_1 = SS$ or $C_2 = FFSF$ occurs as an initial segment of $\{Z_n, n \geq 1\}$; we have the unmatched scenario $C_3 = FSS$, associated with the pattern SS ; and we have a pair of matched scenarios, $C_4 = SF SF$ and $C_5 = FFSF$, which are associated with the pattern $FFSF$. The profit matrix $\{W_{ij}\}$ is then given by

$$\begin{pmatrix} \frac{1}{p_{SS}} & 0 & 0 \\ 0 & \frac{1}{p_{SF}} & \frac{1}{p_{FS} p_{SF}} + \frac{1}{p_{SF}} \\ \frac{1}{p_{FS} p_{SS}} + \frac{1}{p_{SS}} & 0 & 0 \\ 0 & \frac{1}{p_{SF} p_{FS} p_{SF}} + \frac{1}{p_{SF}} & \frac{1}{p_{SF} p_{FS} p_{SF}} + \frac{1}{p_{FS} p_{SF}} + \frac{1}{p_{SF}} \\ 0 & \frac{1}{p_{FF} p_{FS} p_{SF}} + \frac{1}{p_{SF}} & \frac{1}{p_{FS} p_{SF}} + \frac{1}{p_{SF}} \end{pmatrix},$$

and, after solving the corresponding linear system, we find that the appropriate initial team bets are given by

$$y_1^* = \frac{p_{FS} p_{SS}}{1 + p_{FS}}, \quad y_2^* = \frac{p_{FF} p_{FS} p_{SF}}{p_{FS} + p_{SF} + p_{FS} p_{SF}}, \quad y_3^* = \frac{p_{FS} p_{SF} (p_{SF} - p_{FF})}{p_{FS} + p_{SF} + p_{FS} p_{SF}}.$$

The probabilities, p_1 and p_2 , that SS and $FFSF$ are initial segments of the process $\{Z_n, n \geq 1\}$ are respectively given by $p_S p_{SS}$ and $p_F p_{FS} p_{SF}$. Thus, (3) leads to the pleasantly succinct result

$$E[\tau_e] = 2 + p_S p_{SF} + \frac{1 - p_S p_{SS}}{p_{FS}}.$$

4. Generating functions for pattern waiting times

To find the generating function of the waiting time τ_e , we need to introduce the same scenarios and the same betting teams, but we need to make some changes to the design of the initial bets. A gambler from the j th team who arrives at moment $k - 1$ and who begins his betting in round k will now begin with a bet of size $y_j \alpha^k$, where $0 < \alpha < 1$. If $\alpha^{\tau_e} W_{ij}(\alpha) y_j$ denotes the total winnings of the j th team when the game ends in the j th scenario, then we call

$W_{ij}(\alpha)$ the ‘ α -profit matrix.’ As before, the α -profit matrix does not depend on τ_c , and can be computed if we know the ending scenario.

If X_n again denotes the casino’s net gain at time n , then

$$X_{\tau_c} = \frac{\alpha^2 - \alpha\alpha^\tau}{1 - \alpha} \sum_{j=1}^{L+2M} y_j - S(\alpha, y_1, \dots, y_{L+2M}),$$

where

$$S(\alpha, y_1, \dots, y_{L+2M}) = \sum_{i=1}^{K+L+2M} \mathbf{1}_{E_i} \sum_{j=1}^{K+2M} \alpha^{\tau_c} y_j W_{ij}(\alpha).$$

As before, $\mathbf{1}_{E_i}$ is the indicator function for the event, E_i , that the j th scenario occurs.

If $(y_1^*, \dots, y_{L+2M}^*)$ is a solution to the linear system

$$\begin{aligned} y_1^* W_{K+1,1}(\alpha) + \dots + y_{L+2M}^* W_{K+1,L+2M}(\alpha) &= 1, \\ &\vdots \\ y_1^* W_{K+L+2M,1}(\alpha) + \dots + y_{L+2M}^* W_{K+L+2M,L+2M}(\alpha) &= 1, \end{aligned} \tag{4}$$

then we might hope to mimic our earlier calculation of $E[X_{\tau_c}]$, but unfortunately we run into trouble since $E[\alpha^{\tau_c} \mathbf{1}_{E_1}]$ may not equal $p_1 E[\alpha^{\tau_c}]$.

Nevertheless, if the i th, $i \leq K$, scenario occurs, then we know the value of τ_c exactly. It is equal to $|A_i|$, the length of the i th sequence. Therefore, we have a formula for the stopped martingale:

$$X_{\tau_c} = \frac{\alpha^2 - \alpha\alpha^\tau}{1 - \alpha} \sum_{j=1}^{L+2M} y_j^* - \alpha^\tau - I(\alpha, y_1^*, \dots, y_{L+2M}^*).$$

Here, $I(\alpha, y_1^*, \dots, y_{L+2M}^*)$ is defined by setting

$$I(\alpha, y_1^*, \dots, y_{L+2M}^*) = \begin{cases} \alpha^{|A_i|} \left[\sum_{j=1}^{L+2M} y_j^* W_{ij}(\alpha) - 1 \right] & \text{in scenario } i \in \{1, 2, \dots, K\} \\ 0, & \text{in scenario } i > K. \end{cases}$$

From this formula and the optional stopping theorem, we then find the anticipated formula for the moment generating function of τ_c .

Theorem 2. *If $(y_1^*, \dots, y_{L+2M}^*)$ is a solution to the linear system (4), then*

$$E[\alpha^{\tau_c}] = \frac{[\alpha^2/(1 - \alpha)] \sum_{j=1}^{L+2M} y_j^* - \sum_{i=1}^K p_i \alpha^{|A_i|} [\sum_{j=1}^{L+2M} y_j^* W_{ij} - 1]}{1 + [\alpha/(1 - \alpha)] \sum_{j=1}^{L+2M} y_j^*}. \tag{5}$$

4.1. Computation of the α -profit matrix

As before, we need to know how to compute the profit matrix before (5) may be properly regarded as an explicit formula. This is only a little more difficult than before. First, assume

that a scenario ends with the pattern $C = c_1c_2 \cdots c_m$. The team of straightforward gamblers who bet a dollar on pattern $A = a_1a_2 \cdots a_p$ will, by time τ , win an amount

$$\sum_{i=1}^{\min\{m-1,p\}} \frac{\delta_i^{\text{st}}(A, C)}{\alpha^{i-1}},$$

while, in the same time, the team of smart gamblers will win an amount

$$\sum_{i=1}^{\min\{m-1,p\}} \frac{\delta_i^{\text{sm1}}(A, C)}{\alpha^{i-1}} + \sum_{i=1}^{\min\{m-1,p-1\}} \frac{\delta_i^{\text{sm2}}(A, C)}{\alpha^{i-1}}.$$

These formulae provide almost everything we need, but before we can be completely explicit, we need to focus on a concrete example.

4.2. A generating function example

Consider the waiting time to observation of the three-letter pattern FSF in the random sequence $\{Z_n, n \geq 1\}$ produced by the Markov model. In this case, the α -profit matrix $\{W_{ij}(\alpha)\}$ is given by

$$\begin{pmatrix} \frac{1}{PSF} & \frac{\alpha^{-1}}{PFSPSF} + \frac{1}{PSF} \\ \frac{\alpha^{-2}}{PSFPFSPSF} + \frac{1}{PSF} & \frac{\alpha^{-2}}{PSFPFSPSF} + \frac{\alpha^{-1}}{PFSPSF} + \frac{1}{PSF} \\ \frac{\alpha^{-2}}{PFFPFSPSF} + \frac{1}{PSF} & \frac{\alpha^{-1}}{PFSPSF} + \frac{1}{PSF} \end{pmatrix},$$

and, by solving the associated linear system, we obtain

$$y_1^* = \frac{\alpha^2 PFFPFSPSF}{1 - \alpha PFF + \alpha PSF + \alpha^2 PFSPSF}, \quad y_2^* = \frac{\alpha^2 PFSPSF(PFS - PFF)}{1 - \alpha PFF + \alpha PSF + \alpha^2 PFSPSF}.$$

The general moment generating representation (5) then gives us the simple formula

$$E[\alpha^{\tau_e}] = \frac{\alpha^3 PFSPSF(PF + \alpha(PS - PSS))}{1 - \alpha(PSS + PFF - \alpha(PFF - PSF(1 - PFS(1 - \alpha PSS))))}.$$

Naturally, such a formula provides complete information on the distribution of τ_e , and to obtain an explicit formula for $P(\tau_e = k)$ we can rewrite the rational function (5) in its partial fraction expansion.

5. Higher-order Markov chains

In this work, we have applied the gambling team method only to two-state chains. For reasons that will be explained later, this limitation is not easily lifted. Nevertheless, there are more complex chains to which the team method applies, and it is instructive to consider one of these. Specifically, we briefly consider how the gambling team method may be applied to second-order two-state chains. Here we obviously need to avoid the naive representation of such chains as first-order chains with four states.

In the team approach to a second-order model, the gamblers need to observe *two* rounds of betting before they place their first bets, and, consequently, we need to consider a larger number

of final scenarios. Moreover, for each pattern $A = a_1a_2 \cdots a_p$ we will need to consider up to seven termination cases, including three ‘initial’ cases, which are associated with the patterns A , SA , and FA , and four ‘later’ cases, which are associated with the patterns SSA , SFA , FSA , and FFA .

As before, our main objective is to count accurately all the ending scenarios and create a matched number of gambling teams. However, for a second-order chain there is an additional difficulty that must be addressed. Specifically, two cases must be considered separately: (i) there are no runs in the initial list \mathcal{C} and (ii) there are runs in \mathcal{C} .

5.1. Case one: there are no runs in \mathcal{C}

First we need to replace the earlier doubling step with an analogous *quadrupling* step. Now, given the collection of patterns $\mathcal{C} = \{A_i\}_{1 \leq i \leq K}$, we consider the set sequence transformation

$$\mathcal{C} = \{A_i\}_{1 \leq i \leq K} \rightarrow \{SSA_i, SFA_i, FSA_i, FFA_i\}_{1 \leq i \leq K} = \{B_i\}_{1 \leq i \leq 4K} = \mathcal{C}'.$$

We then delete from \mathcal{C}' each scenario that can occur only after the stopping time $\tau_{\mathcal{C}}$, and we take the collection \mathcal{C}'' that remains to be our ‘final list’ of ending scenarios.

Each pattern from the collection \mathcal{C} leads us to four – or perhaps fewer – ending scenarios. For each pattern from \mathcal{C} we consider a sequence of gamblers who belong to teams of different types. As before, these gamblers arrive sequentially and observe the game before placing any bets, according to the following rules.

1. A new gambler from the ‘type-I’ team arrives two rounds before he begins to bet. He watches these rounds and then bets on the successive letters of pattern A , with complete indifference to what he may have seen in the first two rounds he observed.
2. A new gambler from the ‘type-II’ team also arrives two rounds before he begins to gamble, but is influenced by what he sees. If this gambler observes the pattern Sa_1 in these two rounds, then he bets on the sequence $a_2a_3 \cdots a_p$. If he observes anything other than Sa_1 , then he places his bets according to the sequence A .
3. Similarly, a new gambler from the ‘type-III’ team watches two rounds, and if he observes the pattern Fa_1 then he bets according to the sequence $a_2a_3 \cdots a_p$. If he observes anything other than Fa_1 , then he places his bets according to A .
4. Finally, a gambler from the ‘type-IV’ team watches two rounds, and if he observes a_1a_2 then he bets according to the sequence $a_3a_4 \cdots a_p$. Otherwise, he bets according to the sequence A .

Each pattern A from the initial list \mathcal{C} leads to zero, one, two, three, or four scenarios in the final list \mathcal{C}'' . Thus, instead of there being only matched and unmatched patterns, the patterns in final list \mathcal{C}'' are now of four kinds: ‘unmatched,’ ‘double-matched,’ ‘triple-matched,’ and ‘quadruple-matched.’

To see how this works, consider the initial collection $\mathcal{C} = \{FSFF, FFSF\}$. First, note that we have five initial cases, namely $FSFF$, $FFSF$, $SFSFF$, $SFFSF$, and $FFFSF$. Pattern $FFFSF$ cannot occur before $\tau_{\mathcal{C}}$; therefore, the scenario associated with this pattern is eliminated from the list of initial cases. Second, in the final list \mathcal{C}'' we have five patterns: the double-matched patterns $SSFSFF$ and $FSFSFF$ generated by $FSFF$, and the triple-matched patterns $SSFFSF$, $SFFFSF$, and $FFFFSF$ generated by $FFSF$. Thus, we need to introduce five teams: type-I and type-II teams that bet on $FSFF$, and type-I, type-II, and type-III teams that bet on $FFSF$.

5.2. Case two: special treatment of runs

If the initial list \mathcal{C} contains a run, then there may be a problem with the straightforward application of the method described above. The difficulty is that if we observe the game only until the moment $\tau_{\mathcal{C}}$, then there is no difference in behavior between teams of different types that place their bets on the run.

To illustrate the problem, let us consider the initial list $\mathcal{C} = \{F^{(r)}\}$. The straightforward usage of the above algorithm tells us that we must introduce two initial cases, $F^{(r)}$ and $SF^{(r)}$. The final list \mathcal{C}'' contains two double-matched patterns, $SSF^{(r)}$ and $FSF^{(r)}$. Therefore, according to the algorithm, we need two (type-I and type-II) teams that respectively bet y_1 and y_2 dollars on $F^{(r)}$. However, since before time $\tau_{\mathcal{C}}$ there is no difference in gambling between these two teams, we in effect have only one team, which bets $y_1 + y_2$ dollars on the run. Thus, the number of free parameters does not match the number of ending scenarios.

However, a simple modification of the gambling method easily solves the problem. Before time $\tau_{\mathcal{C}}$, the run $F^{(r)}$ can only occur as an initial segment of the sequence $\{Z_n, n \geq 1\}$, or, later, in the pattern $SF^{(r)}$. Thus, if the initial list \mathcal{C} contains runs (obviously we can have only one or two runs in \mathcal{C} , namely $F^{(r)}$ or $S^{(p)}$ or both), then we first need to substitute the runs $F^{(r)}$ and $S^{(p)}$ in \mathcal{C} by $SF^{(r)}$ and $FS^{(p)}$, respectively, to obtain a different collection $\tilde{\mathcal{C}}$. The collection $\tilde{\mathcal{C}}$ contains no runs; therefore, we can proceed as before. After application of the quadrupling and elimination processes to the list $\tilde{\mathcal{C}}$ we will obtain the final list of ending scenarios $\tilde{\mathcal{C}}''$, and for this list we will be able to create a matched number of gambling teams. Since we are interested in $\tau_{\mathcal{C}}$, not $\tau_{\tilde{\mathcal{C}}}$, the elimination process has to be based on $\tau_{\mathcal{C}}$, and the runs must be included in the list of initial cases.

For instance, if $\mathcal{C} = \{F^{(r)}\}$ then we must consider four initial cases, $F^{(r)}$, $SF^{(r)}$, $SSF^{(r)}$, and $FSF^{(r)}$, and four later cases, in which the game ends with $SSSF^{(r)}$, $SFSF^{(r)}$, $FSSSF^{(r)}$, or $FFSF^{(r)}$. For this collection, it can be shown that each of the four teams (type I, II, III, and IV) that bet on $SF^{(r)}$ bets in its own way.

5.3. Final step

After attending to this bookkeeping, we can now calculate the expected observation times in a way that parallels our earlier calculation. Since we have matched the number of (non-initial) ending scenarios and the number of teams, we can choose the size of the initial bet for each team in a way that makes all the expressions for the stopped martingale equal to 1, however the game may end.

Let us summarize this as a theorem. Assume that ultimately we have P initial cases and Q later cases. Let $W_{ij}y_j$, $i = 1, 2, \dots, P + Q$, $j = 1, 2, \dots, Q$, denote the amount of money won by the j th team that bets y_j dollars in the i th scenario. Finally, let p_i , $i = 1, 2, \dots, P$, be the probability that the i th initial scenario occurs.

Theorem 3. *If $(y_1^*, y_2^*, \dots, y_Q^*)$ solves the linear system*

$$\begin{aligned} y_1^* W_{P+1,1} + \dots + y_Q^* W_{P+1,Q} &= 1, \\ &\vdots \\ y_1^* W_{P+Q,1} + \dots + y_Q^* W_{P+Q,Q} &= 1, \end{aligned}$$

then

$$E[\tau_{\mathcal{C}}] = 2 + \frac{\sum_{i=1}^P p_i \sum_{j=1}^Q y_j^* W_{ij} + (1 - \sum_{i=1}^P p_i)}{\sum_{j=1}^Q y_j^*}.$$

6. Concluding remarks

The method of gambling teams deals quite effectively with the waiting time problems of two-state chains, but for N -state chains it is much less effective. The problem is that we typically find that the number of ending scenarios is larger than the number of teams, meaning that there are too few free parameters to achieve the required matching.

We might think to reduce the waiting time problems for an N -state chain by encoding the states $\{1, 2, \dots, N\}$ as sequences of 0s and 1s, but this idea typically fails, since the natural encodings do not lead to a waiting time problem for a *homogeneous* two-state Markov chain on $\{0, 1\}$. Ironically, for many of the pattern problems associated with N -state Markov chains, the method of gambling teams is ineffective when $N \geq 3$, even though it is typically the method of choice for the corresponding problems in a two-state chain.

A possible *computational* advantage of the martingale method over the Markov chain embedding method (see, e.g. Antzoulakos (2001), Fu (2001), and Fu and Chang (2002)) is the size of the matrices involved in the calculations. The size of the profit matrix depends only on the number of patterns, K , while the size of the transition matrix of the embedded Markov chain depends also on the length of patterns in \mathcal{C} . For example, if \mathcal{C} contains K patterns, each of which has a length that is approximately N , and K is much smaller than N , then the dimensions of the transition matrix in the Markov chain embedding method are approximately $(K \times N) \times (K \times N)$. For large N , this can cause technical problems. The size of profit matrix is at most $2K \times 2K$.

References

- AKI, S., BALAKRISHNAN, N. AND MOHANTY, S. G. (1996). Sooner and later waiting time problems and failure runs in higher order Markov dependent trials. *Ann. Inst. Statist. Math.* **48**, 773–787.
- ANTZOULAKOS, D. (2001). Waiting times for patterns in a sequence of multistate trials. *J. Appl. Prob.* **38**, 508–518.
- BENEVENTO, R. V. (1984). The occurrence of sequence patterns in ergodic Markov chains. *Stoch. Process. Appl.* **17**, 369–373.
- BIGGINS, J. D. AND CANNINGS, C. (1987a). Formulas for mean restriction-fragment lengths and related quantities. *Amer. J. Hum. Genet.* **41**, 258–265.
- BIGGINS, J. D. AND CANNINGS, C. (1987b). Markov renewal processes, counters and repeated sequences in Markov chains. *Adv. Appl. Prob.* **19**, 521–545.
- BLOM, G. AND THORBURN, D. (1982). How many random digits are required until given sequences are obtained? *J. Appl. Prob.* **19**, 518–531.
- BREEN, S., WATERMAN, M. AND ZHANG, N. (1985). Renewal theory for several patterns. *J. Appl. Prob.* **22**, 228–234.
- CHRYSSAPHINO, O. AND PASTAVRIDIS, S. (1990). The occurrence of a sequence of patterns in repeated dependent experiments. *Theory Prob. Appl.* **35**, 145–152.
- FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd edn. John Wiley, New York.
- FU, J. C. (1986). Reliability of consecutive- k -out-of- n : F systems with $(k - 1)$ -step Markov dependence. *IEEE Trans. Reliab.* **35**, 602–606.
- FU, J. C. (2001). Distribution of the scan statistics for a sequence of bistate trials. *J. Appl. Prob.* **38**, 908–916.
- FU, J. C. AND CHANG, Y. (2002). On probability generating functions for waiting time distribution of compound patterns in a sequence of multistate trials. *J. Appl. Prob.* **39**, 70–80.
- FU, J. C. AND KOUTRAS, M. V. (1994). Distribution theory of runs: a Markov chain approach. *J. Amer. Statist. Assoc.* **78**, 168–175.
- GERBER, H. AND LI, S. (1981). The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stoch. Process. Appl.* **11**, 101–108.
- GUIBAS, L. AND ODLYZKO, A. (1981a). Periods of strings. *J. Combinatorial Theory A* **30**, 19–42.
- GUIBAS, L. AND ODLYZKO, A. (1981b). String overlaps, pattern matching and nontransitive games. *J. Combinatorial Theory A* **30**, 183–208.
- HAN, Q. AND HIRANO, K. (2003). Sooner and later waiting time problems for patterns in Markov dependent trials. *J. Appl. Prob.* **40**, 73–86.
- LI, S. (1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann. Prob.* **8**, 1171–1176.

- POZDNYAKOV, V. AND KULLDORFF, M. (2006). Waiting times for patterns and a method of gambling teams. *Amer. Math. Monthly* **113**, 134–143.
- POZDNYAKOV, V., GLAZ, J., KULLDORFF, M. AND STEELE, J. M. (2005). A martingale approach to scan statistics. *Ann. Inst. Statist. Math.* **57**, 21–37.
- ROBIN, S. AND DAUDIN, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* **36**, 179–193.
- STEFANOV, V. T. (2000). On some waiting time problems. *J. Appl. Prob.* **37**, 756–764.
- STEFANOV, V. T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach. *J. Appl. Prob.* **40**, 881–892.
- STEFANOV, V. T. AND PAKES, A. G. (1997). Explicit distributional results in pattern formation. *Ann. Appl. Prob.* **7**, 666–678.
- UCHIDA, M. (1998). On generating functions of waiting time problems for sequence patterns of discrete random variables. *Ann. Inst. Statist. Math.* **50**, 655–671.
- WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge University Press.