A NOTE ON OCCURRENCE OF GAPPED PATTERNS IN I.I.D. SEQUENCES

VLADIMIR POZDNYAKOV

Abstract. A new martingale technique is developed to find formulas for the expected value and generating function of the waiting time until one observes a gapped pattern (or a structured motif) in an i.i.d. sequence of random letters from a finite alphabet.

Keywords: Gapped pattern, structured motif, waiting time, martingale, gambling.

1. Introduction

In this work the waiting time until the first occurrence of a gapped pattern (or a structured motif) in an i.i.d. sequence of random letters is studied. A gapped pattern is defined as a collection of patterns that are composed of two fixed patterns (we call them *prefix* and *suffix*) separated by a variable gap. For instance, promoters for *Bacillus subtilis* (Robin et al. (2002)) form the gapped pattern of the following structure: `ttgaca...tataat` with gaps of length 16, 17 or 18. In theory, the occurrence of gapped patterns can be treated by methods developed for the occurrence of compound patterns (for instance, Fu and Lou (2006)). But because of a huge cardinality of a compound pattern associated with a gapped pattern it is often computationally prohibitive. Therefore, algorithms that use the structure of gapped patterns are needed.

The occurrence of gapped patterns is a challenging probabilistic problem. Some useful approximations can be found in Robin et al. (2002). The first exact probability results on gapped patterns appeared in Stefanov et al. (2007). In particular, an expression for the generating functions of the waiting time in case of more general Markov dependent trials is obtained. But these results are derived for gapped patterns that are defined in a slightly different, nontraditional, way. According to the definition in Stefanov et al. (2007), a collection of patterns that compose a gapped pattern contains only such strings for which both prefix and suffix of the gapped pattern can appear only once in a string from the collection. In other words, they have some additional restrictions on symbols that can appear in the gap. Here we deal with a simpler i.i.d. sequences but we allow the multiple occurrences of prefix or suffix inside a gapped pattern, i.e. no restrictions on symbols in gaps.

The proposed approach gives a significant computational advantage in comparison to any method based on the occurrence of compound patterns. For instance, in one of the examples of section 4 instead of working with 64 single patterns that compose a gapped pattern we need to consider only 5 special patterns.

The paper is organized as follows. In section 2 we state the problem. In section 3 we recall an elegant martingale technique that Li (1980) introduced to treat the occurrence of patterns in i.i.d. sequences of random letters. In section 4 we derive a formula for the expected value of the waiting time till a gapped pattern, in section 5 the computational complicity of the method and further extensions are discussed, and in section 6 an expression of the generating function is given.

2. Problem statement

Let $\{Z_n, n \geq 1\}$ be a sequence of i.i.d. random letters from a finite alphabet $\Omega = \{A, B, C, ...\}$, with $|\Omega| = K$ and the following distribution

$$a = \mathbf{P}(Z_n = A), \quad b = \mathbf{P}(Z_n = B), \quad c = \mathbf{P}(Z_n = C), ...$$

The gapped pattern (or structured motif) $\mathbb{P}[d_1 * d_2]\mathbb{S}$ is a finite collection $\mathcal{C}$ of all finite ordered sequences (patterns) over the alphabet Ω that
(1) have the pattern $\mathbb{P} = P_1 P_2 \cdots P_p$ as a prefix,
(2) have the pattern $\mathbb{S} = S_1 S_2 \cdots S_s$ as a suffix,
(3) have d letters between $\mathbb{P}$ and $\mathbb{S}$, where $d_1 \leq d \leq d_2$.

Example 2.1. Let $\Omega = \{A, B, C\}$. Then $AA[1 * 2]BB$ is the following collection of 12 patterns

$$AAABB,\ AABBB,\ AACBB,$$
$$AAAABB,\ AAABBB,\ AAACBB,$$
$$AABABB,\ AABBBB,\ AABCBB,$$
$$AACABB,\ AACBBB,\ AACCBB.$$

Let τ be the first time when one observes $\mathbb{P}[d_1 * d_2]\mathbb{S}$ as a run of the sequence $\{Z_n, n \geq 1\}$ (i.e. τ is a position of the last letter of the first observed pattern from $\mathcal{C}$ in the stochastic sequence $\{Z_n, n \geq 1\}$). The goal is to find the expected value of τ and its generating function.

It is clear that τ can be viewed as the waiting time till the first occurrence a member of a smaller finite collection of *non-redundant* patterns $\tilde{\mathcal{C}}$ (with $|\tilde{\mathcal{C}}| \leq |\mathcal{C}|$) associated with a gapped pattern. To obtain $\tilde{\mathcal{C}}$ one needs to eliminate some patterns from $\mathcal{C}$ with help of the following rule: if one pattern from $\mathcal{C}$ is a subpattern of another pattern from $\mathcal{C}$ then the longer pattern must be deleted.

Example 2.2. Let $\Omega = \{A, B, C\}$. The finite collection of non-redundant patterns $\tilde{\mathcal{C}}$ (a compound pattern) associated with $AA[1 * 2]BB$ contains 7 patterns:

$$AAABB,\ AABBB,\ AACBB,$$
$$AABABB,\ AABCBB,$$
$$AACABB,\ AACCBB.$$

However, to apply our method we need to work with a different subcollection of patterns from $\mathcal{C}$. Let $\hat{\mathcal{C}}$ denotes the collection of patterns that *can be observed* at moment τ. To get this subcollection we use the following rules: (1) if a pattern from $\mathcal{C}$ is strictly inside of another one, then the longer pattern is excluded, (2) if a pattern is a suffix of another one then we keep both patterns.

Example 2.3. Let $\Omega = \{A, B, C\}$. The finite collection of *observable* patterns $\hat{\mathcal{C}}$ associated with $AA[1 * 2]BB$ contains 9 patterns:

$$AAABB,\ AABBB,\ AACBB,$$
$$AAAABB,\ AAACBB,$$
$$AABABB,\ AABCBB,$$

$$AACABB,\ AACCBB.$$

As we have mentioned in the introduction the distribution of τ can be found, at least in theory, by various methods developed for the waiting time till the compound pattern $\tilde{\mathcal{C}}$. But if the size of the gap d_1 is large, then any method based on the full count of patterns in $\tilde{\mathcal{C}}$ can be computationally infeasible. Here we develop a martingale technique that allows us significantly decrease the computational complexity.

3. Li's martingale technique

Before we deal with gapped patterns, let us recall the martingale technique introduced by Li (1980) and Gerber and Li (1981) for finding the expected waiting time till a single pattern $\mathbb{P}$.

Imagine that we have a flow of gamblers (or a gambling team) visiting a casino that generates the sequence $\{Z_n, n \geq 1\}$. The nth gambler arrives right before Z_n will be observed. This gambler places the \$1 bet that $Z_n = P_1$. If Z_n is not P_1 the gambler goes home empty handed. If Z_n indeed yields P_1, gambler wins $1/\mathbf{P}(Z_n = P_1)$. Then he bets his entire capital on $Z_{n+1} = P_2$. If it is not P_2, he goes home with nothing, otherwise he increases his capital by factor $1/\mathbf{P}(Z_{n+1} = P_2)$. Then he continues in the same fashion until the entire pattern $\mathbb{P}$ is exhausted. If the gambler is lucky he leaves the game with total winnings of

$$[\mathbf{P}(Z_n = P_1)\mathbf{P}(Z_{n+1} = P_2) \times \cdots \times \mathbf{P}(Z_{n+p-1} = P_p)]^{-1}$$

dollars. Otherwise, he loses his initial bet of \$1.

Now, let X_n denotes the total net gain of the casino. It is easy to see that $\{X_n, \sigma(Z_1, ..., Z_n)\})$ is a martingale with bounded increments, because the size of bets in the nth round depends only on the history of the process before the round, and odds are fair. The stopped martingale X_τ is given by

$$X_\tau = \tau - W,$$

where W is the total winning of gamblers by time τ. The trick is that W is not a random variable and it is fully determined by overlapping of the pattern $\mathbb{P}$ with itself. More specifically, most of the gamblers are losers. To win something one needs to see $\mathbb{P}$, and nobody who enters the game before time $\tau - p + 1$ sees it. Only the gambler who enters the game at time $\tau - p + 1$ and perhaps those who enters after him can have some amount in their pockets. The total amount of money that these few players have is represented by the following measure of overlapping of pattern $\mathbb{P}$ with itself. First, for $0 \leq i, j \leq p$ let

$$\delta_{ij} = \begin{cases} 1/\mathbf{P}(Z_1 = P_i), & \text{if } P_i = P_j, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$W = \delta_{11}\delta_{22} \cdots \delta_{pp} + \delta_{21}\delta_{32} \cdots \delta_{pp-1} + \cdots + \delta_{p1}.$$

(This explains why symmetric patterns take longer time to appear.)

Employing the optional stopping theorem (see, for instance, Williams (1991, p. 100)) we get

$$0 = \mathbf{E}X_\tau = \mathbf{E}\tau - W,$$

and, as a consequence,

$$\mathbf{E}\tau = W.$$

Note that to apply the optional stopping theorem one needs to make sure that $\mathbf{E}\tau < \infty$ which can be done by showing that τ is bounded by a random variable with a geometric distribution (see Li (1980)).

Example 3.1. Let $\Omega = \{A, B\}$. Consider pattern $AABA$.

$$\mathbf{E}\tau = W = (a \times a \times b \times a)^{-1} + (a)^{-1}$$

4. Method of gambling teams: expected time

Suppose that now one needs to find the expected waiting time till the first occurrence of a gapped pattern with a *fixed*—just for this instance—gap: $\mathbb{P}[d * d]\mathbb{S}$. The basic idea is quite simple: let gamblers bet on pattern $\mathbb{P}$ first, then (if they finished $\mathbb{P}$) pause for d rounds, and then continue their betting on $\mathbb{S}$. Again, the total casino net gain $\{X_n, \sigma(Z_1, ..., Z_n)\}$ forms a martingale. However, the total winnings of the team, W, now is a random variable. Indeed, the value of W depends on how we stopped the game. For instance, if the game is stopped by a pattern from $\hat{\mathcal{C}}$ which contains only one $\mathbb{P}$ as subpattern we have one value of W, if it contains two $\mathbb{P}$s then the value of W is different.

To address the difficulty we will employ the idea of gambling teams introduced by Pozdnyakov et al. (2005) (see also Pozdnyakov and Kulldorff (2006)). First, we compose a list of all possible ending scenarios depending on how many overlapping occurrences of $\mathbb{P}$ we observe right before stopping. Then we introduce a matching number of gambling teams each of which will bet on a particular scenario from the list. Finally, we choose the sizes of initial bets for each gambling team in such a way that the total winnings of all the teams is always equal to 1 regardless of the ending scenario.

Now, to simplify our exposition let us make the following assumption.

Assumption 1. *$\mathbb{S}$ is not a subpattern of $\mathbb{P}$.*

This assumption excludes the case when some pattern from $\hat{\mathcal{C}}$ has a suffix that coincides with a prefix of $\mathbb{P}$ which is longer than $\mathbb{S}$. This condition is a technical one, and later we will comment how one can treat occurrence of gapped patterns when this condition is not present.

We say that two patterns $\mathbb{G}_1$ and $\mathbb{G}_2$ from $\hat{\mathcal{C}}$ are *similar* if:
(1) patterns $\mathbb{G}_1$ and $\mathbb{G}_2$ have the same lengths,
(2) patterns $\mathbb{G}_1$ and $\mathbb{G}_2$ contain the same number of overlapping occurrences of $\mathbb{P}$,
(3) patterns $\mathbb{G}_1$ and $\mathbb{G}_2$ have $\mathbb{P}$s on the same positions.

It is clear that the introduced relation between the patterns from $\hat{\mathcal{C}}$ is an equivalence relation that partitions $\hat{\mathcal{C}}$ into disjoint equivalence classes.

Example 4.1. Let $\Omega = \{A, B, C\}$. Consider the gapped pattern $AA[1 * 2]BB$. Then $AABBB$ is similar to $AACBB$, but not similar to $AAABB$, because last one has two AAs. Patterns $AABABB$, $AABCBB$, $AACABB$ and $AACCBB$ are similar. The class that contains $AABBB$ will be denoted by $AA * BB$, $AABABB$ – by $AA * *BB$, and $AAACBB$ – by $AAA * BB$.

The list of the equivalence classes gives us the corresponding list of ending scenarios. Let us note here that when $\mathbb{G}_1$ is a suffix of $\mathbb{G}_2$ the ending scenario associated with the shorter pattern $\mathbb{G}_1$ includes only situations when the longer pattern $\mathbb{G}_2$ is not observed. With each ending scenario we associate a gambling team that bets on a compound pattern that corresponds the equivalence class, and stars mean pauses in betting.

Example 4.2. Let $\Omega = \{A, B, C\}$. Consider the gapped pattern $AA[6*6]BB$. For equivalence class $AAA*AA**BB$ gamblers from the associated team bet first on pattern AAA, then pause for one round, then bet the entire capital on pattern AA, then pause for two more rounds, and then finally bet on BB.

Assume that we have N ending scenarios and N matching gambling teams. The gamblers from the jth team bets y_j dollars on the compound pattern associated with the jth ending scenario. Let $y_j W_{ij}$ be the total winning of the jth team in case when the game is ended by the ith scenario. The key observation is that W_{ij} are *not* random variables, and a bit later an explicit expression for the W_{ij} will be provided. The net casino gain at the time τ is given by

$$X_\tau = (y_1 + y_2 + ... + y_N)\tau - \sum_{i=1}^{N} \sum_{j=1}^{N} y_j W_{ij} 1_{E_i},$$

where 1_{E_i} is an indicator that the game is ended by the ith ending scenario. Suppose that we can find such $(y_1, y_2, ..., y_N)$ that

$$\sum_{j=1}^{N} y_j W_{ij} = 1, \qquad \text{for all } i. \tag{1}$$

Then the stopped martingale is given by

$$X_\tau = (y_1 + y_2 + ... + y_N)\tau - 1.$$

Applying the optional stopping theorem we obtain the following result.

Theorem 4.1. *If $(y_1, y_2, ..., y_N)$ solves the linear system (1), then*

$$\mathbf{E}[\tau] = (y_1 + y_2 + ... + y_N)^{-1}. \tag{2}$$

Example 4.3. Let $\Omega = \{A, B, C\}$. Consider the gapped pattern $AA[1*2]AB$. In this case $|\mathcal{C}| = |\hat{\mathcal{C}}| = 12$, and $|\tilde{\mathcal{C}}| = 9$, but we have only 6 ending scenarios:

Ending scenario	Patterns from $\hat{\mathcal{C}}$
$AA**AB$	$AABBAB$, $AABCAB$, $AACBAB$, $AACCAB$
$AA*AB$	$AABAB$, $AACAB$
$AAA*AB$	$AAABAB$, $AAACAB$
$AA*AAB$	$AABAAB$, $AACAAB$
$AAAAB$	$AAAAB$
$AAAAAB$	$AAAAAB$

The matrix W_{ij} is given by

$$\begin{bmatrix} \frac{1}{a^3\,b} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{a^3\,b} & 0 & 0 & 0 & 0 \\ \frac{1}{a^3\,b} & \frac{1}{a^3\,b} & \frac{1}{a^4\,b} & 0 & 0 & 0 \\ \frac{1}{a^2} & \frac{1}{a^2} + \frac{1}{a^3\,b} & 0 & \frac{1}{a^2} + \frac{1}{a^4\,b} & 0 & 0 \\ \frac{1}{a^2} + \frac{1}{a^3\,b} & \frac{2}{a^2} & \frac{1}{a^3} & \frac{1}{a^2} & \frac{1}{a^4\,b} & 0 \\ \frac{1}{a^2} + \frac{1}{a^3\,b} & \frac{2}{a^2} + \frac{1}{a^3\,b} & \frac{1}{a^3} + \frac{1}{a^4\,b} & \frac{1}{a^2} + \frac{1}{a^4\,b} & \frac{1}{a^4\,b} & \frac{1}{a^5\,b} \end{bmatrix}$$

Solving linear system (1) and applying formula (2) we obtain

$$\mathbf{E}\tau = \frac{1 + a^2\,b}{(2-a)\;a^3\,b\;(1-a^2\,b)}.$$

Example 4.4. Let $\Omega = \{A, B, C, D\}$. Consider the gapped pattern $ABA[3*3]ACA$. Note that $|\mathcal{C}| = |\tilde{\mathcal{C}}| = |\hat{\mathcal{C}}| = 64$. Since here we deal with a longer gapped pattern and a larger alphabet our technique gives a significant computational simplification in comparison to methods based on the full count. The gambling teams approach leads us to only 5(!) ending scenarios:

$$ABA***ACA,$$

$$ABABA*ACA,$$

$$ABAABAACA,$$

$$ABA*ABACA,$$

$$ABABABACA.$$

Here some values of matrix W_{ij}. The gambling team that bets \$1 on $ABABA*ACA$ in the case when game is ended by $ABABA*ACA$ will win

$$(ababaaca)^{-1} + a^{-1.}$$

The same team in the case of ending scenario $ABABABACA$ will win

$$(ababaaca)^{-1} + (ababaa)^{-1} + a^{-1.}$$

After solving linear system (1) and applying formula (2) we find that

$$\mathbf{E}\tau = \frac{-1 + a^2\,b\,c\,\left(-1 - a + a^4\,(1+a)\,b\,c + a^4\,\left(-1 + a + a^3\right)\,b^2\,c^2 - a^6\,b^3\,c^3\right)}{a^4\,b\,c\,\left(-1 + a^3\,b\,c\,\left(1 + a + \left(-1 + a + a^3\right)\,b\,c - a^2\,b^2\,c^2\right)\right)}$$

In particular, when $a = 1/6$, $b = 1/5$ and $c = 1/4$, $\mathbf{E}\tau \approx 2.59688 \times 10^4$. Note that the expected value till the first occurrence of $ABAACA$ in this case is equal to 2.5926×10^4.

Remark 4.1. At the moment the existence of a solution of linear system (1) is an open question. It seems safe to conjecture that a solution always exists. Let us list two possible approaches that can be explored to prove it. First, note that in every row of matrix W_{ij} the largest element is on the main diagonal, because in the ith ending scenario the biggest profit is going to the gambling team that bets on the same scenario. Therefore, for a certain choice of probability distribution over Ω we can make this matrix almost diagonal. And then perhaps some kind of continuity arguments might be used to show the existence of a solution. Second approach might be based on an appropriate Markov chain imbedding as it was done by Gerber and Li (1981).

Remark 4.2. (*Calculation of W_{ij}*) Consider an *extended* alphabet

$$\bar{\Omega} = \{*, A, B, C, ...\}$$

with $|\bar{\Omega}| = K + 1$. Let $\mathbb{E} = E_1E_2\cdots E_m$ and $\mathbb{T} = T_1T_2\cdots T_n$ be two patterns over $\bar{\Omega}$.

First, we introduce the following measure of two letters coincidence:

$$\delta(E_i, T_j) = \begin{cases} 1, & \text{if } T_j = * \\ 1/\mathbf{P}(Z_1 = T_j), & \text{if } T_j \neq *, E_i = T_j, \\ 0, & \text{if } T_j \neq *, T_j \neq E_i. \end{cases}$$

Second, we define the following measure of overlapping of $\mathbb{E}$ and $\mathbb{T}$:

$$W(\mathbb{E},\mathbb{T}) = \sum_{i=1}^{\min(m,n)} \prod_{j=1}^{i} \delta(E_{m-i+j}, T_j).$$

Finally, if the ith ending scenario is associated with pattern $\mathbb{E}$, and the jth gambling team bets on $\mathbb{T}$, then

$$W_{ij} = W(\mathbb{E},\mathbb{T}).$$

5. Discussion: complexity of the method and extensions

To illustrate computational issues related to the method let us consider alphabet $\Omega = \{a, c, t, g\}$ and gapped pattern $ttgaca[16*18]tataat$ (promoters for *Bacillus subtilis* that were mentioned in the introduction).

To find the expected time $\mathbf{E}(\tau)$ via an appropriate Markov chain imbedding one needs to solve a linear system associated with the transition matrix of imbedded Markov chain—for instance, see Fu and Chang (2002, p. 73). Therefore, the size of transition matrix is important. If we take a straightforward approach and think about $ttgaca[16*18]tataat$ as a compound pattern $\tilde{\mathcal{C}}$, then the number of states of the imbedded Markov chain is too high because the cardinality of the compound pattern is high: $4^{18}+4^{17}+4^{16} \approx 9\times 10^{10}$. As a consequence, the use of the formula of Fu and Chang (2002) is not practical for this Markov chain.

However, it is possible to imbed a much smaller Markov chain. More specifically, in recent papers Nuel (2006a and 2006b) shows how with help of the deterministic finite state automata (see Hopcroft et al. (2001)) one can construct a significantly smaller Markov chain for $ttgaca[16*18]tataat$ with only 1059 transient states and 43 final states (in the case of i.i.d. letters). As a result, the formula of Fu and Chang (2002) can be employed now to obtain the expected waiting time.

Similarly, the size of linear system (1) (or the list of the ending scenarios) is the main computational issue of the presented martingale technique. So let us see how many ending scenarios we have in this case.

Note that subcollections $\mathcal{C}$, $\tilde{\mathcal{C}}$ and $\hat{\mathcal{C}}$ are not used in practical calculations. Those subcollections are needed to provide the exact meaning of the star-patterns associated with ending scenarios. A star in a star-pattern does not always mean "any letter", it is rather a pause in the betting.

In practice, we can directly start with listing of ending scenarios. First, let us count how many ending scenarios we have with gap 18. For this gap we can have exactly 1, 2, 3 or 4 prefixes $ttgaca$ in an ending scenario. Thus we get

- $\binom{18}{0} = 1$ ending scenario with 1 occurrence of prefix $ttgaca$ which is
$$ttgaca******************tataat;$$
- $\binom{13}{1} = 13$ ending scenarios with 2 occurrences of $ttgaca$, for example,
$$ttgaca******ttgaca******tataat;$$
- $\binom{8}{2} = 28$ ending scenarios with 3 occurrences of $ttgaca$, for example,
$$ttgaca**ttgaca**ttgaca**tataat;$$
- and $\binom{3}{3} = 1$ ending scenario with 4 occurrences of $ttgaca$ which is
$$ttgacattgacattgacattgacatataat.$$

Second, similar calculations give us $34 = \binom{17}{0} + \binom{12}{1} + \binom{7}{2}$ ending scenarios with gap 17 and $27 = \binom{16}{0} + \binom{11}{1} + \binom{6}{2}$ ending scenarios with gap 16. Thus the total number of ending scenarios is 104 (compare to 1059+43 states of imbedded Markov chain in Nuel (2006b)), and size of matrix W_{ij} is 104 by 104. Each entry of this matrix is computed with help of formulas provided in the previous section. This computation is not heavy at least in terms of needed memory because W_{ij} is fully determined by overlapping of two star-patterns.

Let us also mention here how this technique can be extended to the case of several gaps. It is a bit more difficult to provide a formal description of the algorithm in this case, so let us consider an example. Let $\Omega = \{A, B, C, D\}$ and consider pattern with two gaps:

$$AB[3*3]AC[3*3]CC.$$

We can treat this pattern as a gapped pattern with *one* gap between *gapped* prefix $\mathbb{P} = AB[3*3]AC$ and suffix $\mathbb{S} = CC$. As before, to apply the martingale method we need to track overlapping occurrences of prefix $\mathbb{P}$. The difference is that now we can have *fractional* occurrences of $\mathbb{P}$ in our ending scenarios. More specifically, we have 5 ending scenarios:

- 1 ending scenario with 1 occurrence of $\mathbb{P}$
$$AB***AC***CC;$$
- 2 ending scenarios with 1.5 occurrences of $\mathbb{P}$ ($AB***AC$ and AB later)
$$AB***ACAB*CC \text{ and } AB***AC*ABCC;$$
- 2 ending scenarios with 2 occurrences of $\mathbb{P}$
$$ABAB*ACAC*CC \text{ and } AB*ABAC*ACCC.$$

It is impossible to observe more than 2 (2.5 or higher) occurrences of $\mathbb{P}$ in a single pattern from the list of patterns associated with $AB[3*3]AC[3*3]CC$.

Finally, the last issue we would like to discuss is a possible extension of this technique to Markov dependent sequences of letters. The common perception is that the martingale approach does not work for Markov dependent trials. However, in our recent paper (Glaz et al. (2006)) we have shown how to treat occurrence of compound patterns with help of martingales in two-state Markov chains. At this moment we even now how one can extend martingale approach to the case of multi-state Markov chains.

The difficulty is obvious. To place a *fair* bet on a coming letter in the case of Markov dependent trials we need to know the result of previous round. This will increase a number of ending scenarios significantly. Because now, whenever we go from a sequence of stars to a real letter, we must know an exact value behind the last star. Most likely, it is doable, but the computational simplifications (if any) will not be as dramatic as in case of i.i.d. sequences. However, if one wants to obtain a direct formula for a higher moment the martingale technique still might be useful.

6. Method of gambling teams: generating function

To derive a formula for the generating function of τ, $\mathbf{E}\alpha^\tau, 0 \le \alpha \le 1$ we need to change the betting system just a bit. Now the gambler from jth team that arrives to place his bet in round n will bet $y_j\alpha^n$ dollars. Let $\alpha^\tau y_j W_{ij}(\alpha)$ denotes the total winning of the jth team in case when the game is ended by ith scenario.

Again $W_{ij}(\alpha)$ is not a random variable, it is fully determined by overlapping of patterns associated with the jth gambling team and the ith ending scenario. More specifically, if star-patterns $\mathbb{E}$ and $\mathbb{T}$ (over the extended alphabet $\bar{\Omega}$) are associated with the ith ending scenario and the jth gambling team, respectively, then

$$W_{ij}(\alpha) = W(\mathbb{E}, \mathbb{T}, \alpha) = \sum_{i=1}^{\min(m,n)} \prod_{j=1}^{i} \delta(E_{m-i+j}, T_j)\alpha^{1-i}.$$

The net gain of the casino at time τ is given by

$$X_\tau = (y_1 + y_2 + ... + y_N)\alpha\frac{\alpha^\tau - 1}{\alpha - 1} - \alpha^\tau \sum_{i=1}^{N}\sum_{j=1}^{N} y_j W_{ij}(\alpha) 1_{E_i},$$

where as before 1_{E_i} is an indicator that the game is ended by the ith ending scenario. Suppose that we can find such $y_j(\alpha)$ that

$$\sum_{j=1}^{N} y_j(\alpha) W_{ij}(\alpha) = 1, \quad \text{for all } i. \tag{3}$$

Then the stopped martingale is given by

$$X_\tau = (y_1(\alpha) + y_2(\alpha) + ... + y_N(\alpha))\alpha\frac{\alpha^\tau - 1}{\alpha - 1} - \alpha^\tau.$$

After a routine application of the optional stopping theorem we come to the following result.

Theorem 6.1. *If* $(y_1(\alpha), y_2(\alpha), ..., y_N(\alpha))$ *solves the linear system (3), then*

$$\mathbf{E}[\alpha^\tau] = 1 - \left(\frac{\alpha}{1-\alpha}[y_1(\alpha) + y_2(\alpha) + ... + y_N(\alpha)] + 1\right)^{-1}. \tag{4}$$

Example 6.1. Let $\Omega = \{A, B, C\}$. Consider again the gapped pattern $AA[1{*}2]AB$. In this case the matrix $W_{ij}(\alpha)$ is given by

$$\begin{bmatrix}
\frac{\alpha^{-4}}{a^3\,b} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\alpha^{-5}}{a^3\,b} & 0 & 0 & 0 & 0 \\
\frac{\alpha^{-4}}{a^3\,b} & \frac{\alpha^{-5}}{a^3\,b} & \frac{\alpha^{-5}}{a^4\,b} & 0 & 0 & 0 \\
\frac{\alpha^{-2}}{a^2} & \frac{\alpha^{-2}}{a^2} + \frac{\alpha^{-5}}{a^3\,b} & 0 & \frac{\alpha^{-2}}{a^2} + \frac{\alpha^{-5}}{a^4\,b} & 0 & 0 \\
\frac{\alpha^{-2}}{a^2} + \frac{\alpha^{-4}}{a^3\,b} & \frac{\alpha^{-2}+\alpha^{-3}}{a^2} & \frac{\alpha^{-3}}{a^3} & \frac{\alpha^{-2}}{a^2} & \frac{\alpha^{-4}}{a^4\,b} & 0 \\
\frac{\alpha^{-2}}{a^2} + \frac{\alpha^{-4}}{a^3\,b} & \frac{\alpha^{-2}+\alpha^{-3}}{a^2} + \frac{\alpha^{-5}}{a^3\,b} & \frac{\alpha^{-3}}{a^3} + \frac{\alpha^{-5}}{a^4\,b} & \frac{\alpha^{-2}}{a^2} + \frac{\alpha^{5}}{a^4\,b} & \frac{\alpha^{-4}}{a^4\,b} & \frac{\alpha^{-5}}{a^5\,b}
\end{bmatrix}$$

In particular, when $a = 1/3$ and $b = 1/4$ Theorem 6.1 gives us:

$$\mathbf{E}\alpha^\tau = \frac{\alpha^5}{108} + \frac{5\,\alpha^6}{324} + \frac{59\,\alpha^7}{3888} + \frac{43\,\alpha^8}{2916} + \frac{85\,\alpha^9}{5832} + \frac{2029\,\alpha^{10}}{139968} + O(\alpha^{11}).$$

Note that first two terms of the series can be easily computed directly.

7. Concluding remarks

As we said before, Assumption 1 can be omitted. All we need to do is to change the definition of similarity of patterns from $\hat{\mathcal{C}}$. Now we also have to look after the suffixes of $\mathbb{G}_1$ and $\mathbb{G}_2$. More specifically, in this case we say that two patterns $\mathbb{G}_1$ and $\mathbb{G}_2$ from $\hat{\mathcal{C}}$ are similar if they satisfy the three old rules and a new one: (4) patterns $\mathbb{G}_1$ and $\mathbb{G}_2$ have the same prefix of $\mathbb{P}$ as their suffix.

Another nice thing about the martingale technique is that it allows to find direct formulas for higher moments as well—see, for instance, Pozdnyakov et al. (2005). And in practical situations the mean and variance of the waiting time often provide a lot of information about its distribution.

8. Acknowledgment

I am pleased to thank a referee for very insightful comments, useful suggestions and references that allowed us to strengthen the presentation.

References

[1] J. Fu, Y. Chang, On probability generating functions for waiting time distribution of compound patterns in a sequence of multistate trials, Journal of Applied Probability, 39 (2002) 70-80.

[2] J. C. Fu, W. Y. W. Lou, Waiting time distributions of simple and compound patterns in a sequence of r-th order Markov dependent multi-state trials, Annals of the Institute of Statistical Mathematics, 58 (2006) 291-310.

[3] J. Glaz, M. Kulldorff, V. Pozdnyakov, J.M. Steele, Gambling teams and waiting times for patterns in two-state Markov chains. Journal of Applied Probability, 43 (2006) 127-140.

[4] H. Gerber, S. Li, The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain, Stochastic Processes and their Applications, 11 (1981) 101-108.

[5] J.E. Hopcroft, R. Motwani, J.D. Ullman, Introduction to Automata Theory, Languages, and Computation (second edition), Addison-Wesley, 2001.

[6] S. Li, A martingale approach to the study of occurrence of sequence patterns in repeated experiments, The Annals of Probability, 8 (1980) 1171-1176.

[7] G. Nuel, Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistic, Algorithms for Molecular Biology, 1:5 (2006a), doi:10.1186/1748-7188-1-5.

[8] G. Nuel, Distribution of patterns on Markov chains: a unified approach using deterministic finite state automata, (2006b), preprint.

[9] V. Pozdnyakov, J. Glaz, M. Kulldorff, J. M. Steele, A martingale approach to scan statistics, Annals of the Institute of Statistical Mathematics, 57 (2005) 21-37.

[10] V. Pozdnyakov, M. Kulldorff, Waiting Times for Patterns and a Method of Gambling Teams, The American Mathematical Monthly, 113 (2006) 134-143.

[11] S. Robin, J.-J. Daudin, H. Richard, M.-F. Sagot, S. Schbath, Occurrence probability of structured motifs in random sequences. Journal of Computational Biology, 9 (2002) 761-773.

[12] V. Stefanov, S. Robin, S. Schbath, Waiting times for clumps of patterns and for structured motifs in random sequences, Discrete Applied Mathematics, 155 (2007) 868-880.

[13] D. Williams, Probability with martingales, Cambridge University Press, Cambridge, 1991.

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269-4120

E-mail address: vladimir.pozdnyakov@uconn.edu