Occurrence of Patterns: Martingale Technique and Applications to Scans

Vladimir Pozdnyakov

University of Connecticut

2006

Based on joint work with J. Glaz, M. Kulldorff, and M. Steele

Outline

- Occurrence of patterns
  - Examples
  - Problem Statement
  - Single sequence
  - Multiple sequences
  - Markov Chain
- Applications to scans
  - From scan to compound pattern
  - Approximations
  - Numerical results

Example 1

We flip a fair coin five times. What pattern is "more difficult" to get: $HHHHH$ or $HTHTH$? If we give this question to a street-wise guy, the most likely answer is: "the first one". Well, we know that answer is not correct. Both patterns have the same probability to occur – 1/32. However, there is a sense in which the street-wise guy *is*, in fact, correct. If we flip the coin without stopping, then the average waiting time for the first occurrence of the pattern $HHHHH$ is 62, while for the pattern $HTHTH$ it is 42. So, the pattern $HHHHH$ is indeed "more difficult" to get.

Example 2

Now, if we ask a person familiar with probability theory (but unfamiliar with this particular topic) to rank the average waiting times until patterns $HHHHH$, $HHHHT$, $HHHTH$ and $HTHTH$, then most likely the first pattern will get rank 1 (the longest average waiting time), the second – 2, the third – 3, and the last one – 4 (the shortest average waiting time). This ranking is based on an "intuitive" idea that the runs ($HHHH$ or $HHHHH$) typically require more time to occur. In fact, the average waiting times are 62 for $HHHHH$, 32 for $HHHHT$, 34 for $HHHTH$, and 42 for $HTHTH$.

Example 3

Suppose that Melanie flips a coin until she observes either $HHHTH$ or $HTHTH$ while Kyle flips another coin until he observes either $HHHHT$ or $HHHTH$. Since Kyle got the two patterns with the shortest waiting times, 32 and 34 versus 34 and 42, one would expect him to have a shorter average waiting time when in fact they are exactly the same – 22 for both Melanie and Kyle.

Example 4: Penney's Game

Consider three patterns:

$$A = HHTHH, \quad B = HTHHT, \quad C = THHTH$$

For these, in case of a fair coin :

$$P(A \text{ occurs before } B) = .583333...$$

$$P(B \text{ occurs before } C) = .590909...$$

$$P(C \text{ occurs before } A) = .625000...$$

Problem Statement

Let Z be an arbitrary discrete random variable with the set of possible values Σ, and let $\{Z, Z_k\}_{k \geq 1}$ be a sequence of independent, identically distributed random variables.

Consider a collection of finite patterns over Σ: $\{A_j\}_{1 \leq j \leq K}$. Assume that no pattern contains another as a subpattern. We will denote by τ_{A_j} the waiting time until A_j occurs as a run in the sequence $Z_1, Z_2, ...$

The objective is to find the expected time of

$$\tau = \min\{\tau_{A_1}, ..., \tau_{A_K}\}, \tag{1}$$

and probabilities $\pi_j = \mathbf{P}\Big(\tau = \tau_{A_j}\Big)$.

Single Pattern

We flip a fair coin and wait for the pattern $A = HTH$.

What is $\mathbf{E}\tau_A$?

Key Martingale

The standard martingale technique is as follows (Li (1981)). Assume that a new gambler arrives just before each time $n = 1, 2, ...$ He bets \$1 that

$$Z_n = H.$$

If he loses, he leaves the game. If he wins, he gets 2 dollars. Then he bets the whole amount, \$2, on the event that

$$Z_{n+1} = T.$$

Again if he loses, he leaves. If he wins his total capital is now \$4 dollars, and he bets his whole fortune on the next event

$$Z_{n+2} = H.$$

If the gambler is lucky and finishes the pattern, he leaves the game with his winnings.

Let X_n be the net amount of money collected by the casino from all the gamblers up until and including time n. Since the amount of the bets at round n depends only on history up to time $n - 1$, and the odds are fair for each gambler, X_n is a martingale.

What is the value of X_{τ_A}?

We flip a fair coin until the first time τ_A when the pattern $A = HTH$ will occur.

By this moment exactly τ_A gamblers entered the game, each of them paid a dollar, and almost all of them lost their money.

Only two gamblers won: the one that entered the game at time $\tau_A - 3$, and another one who started his betting at time $\tau_A - 1$. At time τ_A, the first gambler has got \$8 and the second \$2.

Thus, we get that $X_{\tau_A} = \tau_A - 8 - 2$.

Heavy Artillery

By the Optional Stopping Theorem (Williams, 1991, p. 100) we get that

$$0 = \mathbf{E}(X_0) = \mathbf{E}(X_{\tau_A}) = \mathbf{E}(\tau_A) - 10,$$

and, hence,

$$\mathbf{E}(\tau_A) = 10.$$

Multiple Patterns

We flip a fair coin again. But now we wait for one of two patterns: $A_1 = HTH$ and $A_2 = HH$.

Let $\tau = \min\{\tau_{A_1}, \tau_{A_2}\}$. What is $\mathbf{E}\tau_A$?

Methods:

- Martingale approach: Li (1980) and Gerber and Li (1981)
- Markov Chain embedding method: Fu (1996), Fu and Chang (2002), Antzoulacos (2001) and other
- Recurrent event theory, combinatorics etc: Feller (1968), Guibas and Odlyzko (1981) and other

First Attempt

Assume now that we have 2 teams of betters, and the first team bets on the pattern A_1, and the second team – on A_2.

Let X_n again be the net gain of the casino at time n. It is a martingale. What is X_τ now?

$$X_\tau = \begin{cases} 2 \times \tau - \quad (10 \quad + \quad 2), & \text{if } \tau = \tau_{A_1} \\ 2 \times \tau - \quad (2 \quad + \quad 6), & \text{if } \tau = \tau_{A_2} \end{cases}$$

After taking the expectation we get that

$$0 = \mathbf{E}(X_\tau) = 2\mathbf{E}(\tau) - 12\mathbf{P}(\tau = \tau_{A_1}) - 8\mathbf{P}(\tau = \tau_{A_2}).$$

Not good.

Free Parameters

Let y_j be the initial amount of money with which each of the gamblers from the j-th team start their betting.

Then

$$X_\tau = \begin{cases} (y_1 + y_2) \times \tau - & (10y_1 + 2y_2), \quad \text{if } \tau = \tau_{A_1} \\ (y_1 + y_2) \times \tau - & (2y_1 + 6y_2), \quad \text{if } \tau = \tau_{A_2} \end{cases}$$

Let us choose y_1 and y_2 in such way that

$$\begin{array}{lcll} 10y_1 & + & 2y_2 & = 1 \\ 2y_1 & + & 6y_2 & = 1 \end{array}$$

that is $y_1 = 1/14$ and $y_2 = 1/7$. As consequence, we get

$$0 = \mathbf{E}(X_\tau) = (y_1 + y_2)\mathbf{E}(\tau) - 1,$$

and

$$\mathbf{E}(\tau) = \frac{1}{y_1 + y_2} = 4\frac{2}{3}.$$

Wise Gamblers

Let us consider two other choices of the initial bets (y_1, y_2): $(0, 1)$ and $(1, 0)$. The first choice leads to the equation:

$$0 = \mathbf{E}(\tau) - 2\mathbf{P}(\tau = \tau_{A_1}) - 6\mathbf{P}(\tau = \tau_{A_2}).$$

The other one gives

$$0 = \mathbf{E}(\tau) - 10\mathbf{P}(\tau = \tau_{A_1}) - 2\mathbf{P}(\tau = \tau_{A_2}).$$

That allows us to find that

$$\mathbf{P}(\tau = \tau_{A_1}) = \frac{1}{3}, \quad \mathbf{P}(\tau = \tau_{A_2}) = \frac{2}{3}.$$

An Example

Suppose that we have three independent sequences of iid random variables: $\{Z^{(i)}, Z_k^{(i)}\}_{k\geq 1}$ with

$$\mathbf{P}(Z^{(i)} = A) = p_i, \quad \mathbf{P}(Z^{(i)} = B) = q_i, \quad p_i + q_i = 1, \quad i = 1, 2, 3.$$

Let τ be the waiting time for the 2-by-2 block:

$$\begin{array}{cc} \mathsf{A} & \mathsf{A} \\ \mathsf{A} & \mathsf{A} \end{array}$$

For instance, if the realization of $\{Z^{(i)}, Z_k^{(i)}\}_{k\geq 1}, i = 1, 2, 3$ produced the following three sequences:

$$\begin{array}{cccccccccc} \mathsf{A} & \mathsf{B} & \mathsf{A} & \mathsf{A} & \mathsf{B} & \mathsf{A} & \mathsf{B} & \mathsf{A} & \mathsf{B} & \ldots \\ \mathsf{A} & \mathsf{B} & \mathsf{A} & \mathsf{A} & \mathsf{A} & \mathsf{B} & \mathsf{A} & \mathsf{A} & \mathsf{B} & \ldots \\ \mathsf{A} & \mathsf{B} & \mathsf{B} & \mathsf{B} & \mathsf{B} & \mathsf{A} & \mathsf{A} & \mathsf{A} & \mathsf{A} & \ldots \end{array}$$

then $\tau = 4$.

What is $\mathbf{E}(\tau)$?

What can be done?

- IID Sequence
  - Generating function – initial bets are α^n
  - Moments – initial bets are n^k to get moment of order $k+1$
  - Expected number and generating function of occurrence of subpattern P till observing pattern PB (it works in Markov chain case as well)
- Markov Chain
  - Two-state chains of first (or higher) order
  - General markov chain?
  - Non-homogeneous trials?
  - “Conditional” situation?
  - Multi-dimensional Patterns?

Two-state Markov Chain

Now 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 has the initial distribution $\mathbf{P}(Z_1 = S) = p_S$, $\mathbf{P}(Z_1 = F) = p_F$ and the transition matrix

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

where p_{SF} is shorthand for $\mathbf{P}(Z_{n+1} = F | Z_n = S)$.

What is $\mathbf{E}[\tau_{FSF}]$?

Key Martingale – Watch Then Bet

Now, when gambler number $n + 1$ arrives he observes first the result of the n-th trial, Z_n.

So, he knows how to bet on the next letter in the fair way.

Too Many Ending Scenarios?

The problem is that now for one pattern FSF this time we need to consider three different *ending scenarios*:

1. FSF occurs at the beginning of the sequence $\{Z_n, n \geq 1\}$, or

2. the pattern $SFSF$ occurs, or

3. the pattern $FFSF$ occurs.

Two Teams for One Pattern

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 FSF. If he wins he then bets all of his capital on the next letter in the sequence FSF, and he continues in this way until he either loses his capital or he observes all of the letters of FSF. Such players are called *straightforward gamblers*.

2. The gamblers of the second team make use of the information that they observe. If gambler $n + 1$ observes $Z_n = S$ just before he begins his play, then he bets just 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 $Z_n = F$ when he first arrives, then wagers y_2 dollars on the first letter of the pattern SF. He then continues to wager on the successive letters of SF either until he loses or until he observes SF. Such players are called *smart gamblers*.

Stopped Martingale

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 of stopped martingale X_τ which represents the casino's net gain is given by

$$X_\tau = \begin{cases} (y_1 + y_2)(\tau - 1) - y_1 W_{11} - y_2 W_{12}, & \text{1-st scenario,} \\ (y_1 + y_2)(\tau - 1) - y_1 W_{21} - y_2 W_{22}, & \text{2-nd scenario,} \\ (y_1 + y_2)(\tau - 1) - y_1 W_{31} - y_2 W_{32}, & \text{3-rd scenario.} \end{cases}$$

Choosing Initial Bets

Now, if we take (y_1^*, y_2^*) to be a solution of 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:

$$X_\tau = \begin{cases} (y_1^* + y_2^*)(\tau - 1) - y_1^* W_{11} - y_2^* W_{12}, & \text{1-st scenario,} \\ (y_1^* + y_2^*)(\tau - 1) - 1, & \text{2-nd scenario,} \\ (y_1^* + y_2^*)(\tau - 1) - 1, & \text{3-rd scenario.} \end{cases}$$

Optional Stopping Theorem Routine

The optional stopping theorem then gives us

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

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

$$\mathbf{E}[\tau] = 1 + \frac{p_1(y_1^* W_{11} + y_2^* W_{12}) + (1 - p_1)}{y_1^* + y_2^*}. \tag{2}$$

Done!

$$\mathbf{E}[\tau_{FSF}] = 1 + \frac{p_S}{p_{SF}} + \frac{1}{p_{SF}^2} + \frac{1}{p_{FS}p_{SF}},$$

From scan to compound pattern

Scan. Assume that we observe a sequence of Bernoulli trials, and the probability of failure is known and relatively small – 5%. We have an alert if we observe too many failures during a short period of time. More specifically, we stop the process if we have at least three failures out of 5 sequential trials.

Compound pattern. We have an alert when the following runs occur first time:

1) 3 out of 3

$$FFF,$$

2) 3 out of 4

$$FFSF, \quad FSFF,$$

(note that the runs $SFFF$ and $FFFS$ were counted earlier)

3) 3 out of 5

$$FFSSF, \quad FSFSF, \quad FSSFF.$$

The expected time is 1608.4 and the standard deviation of the waiting time is 1604.8.

Approximations

- *exponential*

$$\mathbf{P}(\tau \leq n) \approx 1 - \exp(-(n-l)/\mu),$$

  where l is the length of the shortest sequence

- *gamma*

$$\mathbf{P}(\tau \leq n) \approx \frac{1}{\Gamma(a)} \int_0^{(n-l)/b} x^a e^{-x} dx,$$

  where l is again the length of the shortest sequence, $b = \sigma^2/\mu$, and $a = \mu/b$.

- *shifted exponential*

$$\mathbf{P}(\tau \leq n) \approx 1 - \exp(-(n + 0.5 + \sigma - \mu))/\sigma),$$

  where the 0.5 term is a continuity correction.

Numerical Results

n	exponential	shifted exponential	gamma	upper bound	lower bound
500	0.01600	0.01589	0.01597	0.01588	0.01589
1000	0.03183	0.03173	0.03179	0.03171	0.03174
1500	0.04741	0.04731	0.04736	0.04729	0.04733
2000	0.06274	0.06265	0.06267	0.06262	0.06267
2500	0.07782	0.07773	0.07775	0.07770	0.07776
3000	0.09266	0.09258	0.09258	0.09254	0.09261
4000	0.12162	0.12155	0.12154	0.12150	0.12169
5000	0.14966	0.14960	0.14957	0.14954	0.14965

Table 1. Fixed window scans: at least 3 out of 10, $\mathbf{P}(F) = .01$, $\mu = 30822$, $\sigma = 30815$

n	exponential	shifted exponential	gamma	upper bound	lower bound
50	0.09110	0.07827	0.08268	0.07713	0.07940
60	0.10977	0.09770	0.10059	0.09543	0.09989
70	0.12807	0.11672	0.11828	0.11337	0.11991
80	0.14599	0.13534	0.13573	0.13095	0.13949
90	0.16354	0.15357	0.15292	0.14819	0.15864
100	0.18073	0.17141	0.16985	0.16508	0.17736

Table 2. Fixed window scans: at least 4 out of 20, $\mathbf{P}(F) = .05$, $\mu = 481.59$, $\sigma = 469.35$

THANK YOU