ON THE NUMBER OF EVENT APPEARANCES
IN A MARKOV CHAIN

Natalia Mezhennaya
Applied Mathematics Department
Bauman Moscow State Technical University
ul. Baumanskaya 2-ya, 5/1
Moscow - 105005, RUSSIA

Abstract: The paper presents the estimate for the total variation distance between the distribution of the number of appearances of homogeneous disjoint events in a segment of strongly ergodic Markov chain on the finite state space and accompanying Poisson distribution (i.e., Poisson distribution with a parameter equal to the expectation of the random variable under consideration). For this purpose the Chen–Stein method was used. As a result Poisson and normal limit theorems for the number of events appearances are derived. The considered scheme describes the well-known number of runs on consecutive letters, the number of \( f \)-recurrent runs, etc., and can be used for describing the properties of distribution of the special form scan statistic.

AMS Subject Classification: 60B10, 60J10, 60F05, 62E20
Key Words: number of events, Markov chain, Chen–Stein method, limit theorem

1. Introduction

The problem of the number of runs of consecutive identical letters in a sequence of independent random variables [7] and its generalization to the case of a sequence forming a Markov chain (see, e.g., [4]) are well known. We consider the following more general scheme.
Let \( \{X_j, j = 1, \ldots, n\} \) be a Markov chain, \( s \geq 1 \). We suppose (informally) that the random event \( A_j \) depends only on the random variables \( X_j, \ldots, X_{j+s} \), and that the set of events \( \{A_j, j = 1, 2, \ldots, n-s\} \) is homogeneous and has the property \( \Pr \{A_i A_j\} = 0 \) for \( |i-j| \leq s \).

In this paper, we study the distribution of the number of appearances of events \( A_j \) in a segment of a Markov chain. For example, such a scheme describes the well-known run of consecutive letters. Thus, according to [7, p. 62], the letters \( X_j, \ldots, X_{j+s} \) form a run of consecutive \( a \)'s of at least \( t \) length if \( X_{j+1} \neq a, X_{j+2} = a, \ldots, X_{j+t} = a \).

This definition could be expanded. Let \( f : A_N^l \to A_N^t \) be a numerical function, \( l \geq 1 \). We define the event as follows:

\[
A_j = \{X_{j+t} \neq f(X_j, \ldots, X_{j+t-1}), X_{j+t+1} = f(X_{j+1}, \ldots, X_{j+t}), \ldots, X_{j+2l+t} = f(X_{j+t}, \ldots, X_{j+t+l-1})\}.
\]

In this case, the event \( A_j \) means that \( X_j, \ldots, X_{j+t+l} \) form an \( f \)-recurrent run of length at least \( t \) (see. [17]). The events \( A_i \) and \( A_j \) are incompatible for \( |i-j| \leq s = t + l \).

It is easy to see that the concept of the \( f \)-recurrent run includes a run of consecutive \( a \)'s ([7, p. 62]). Indeed, if \( l = 1 \) and the function \( f \equiv a, a \in A_N^t \), then

\[
A_j = \{X_{j+1} \neq a, X_{j+2} = a, \ldots, X_{j+t+1} = a\}
\]

and the \( f \)-recurrent run coincides with the run of consecutive \( a \)'s of at least \( t \) length.

The exact distributions of the numbers of runs in binary Markov chains were studied by Savelyev and Balakin [23, 24], Antzoulakos [1], Inoue [11], and their limit distributions in Markov chains with any number of states were obtained by Tikhomirova [25], Chryssaphinou et al. [5], and Fu et al. [9]. The distribution of the length of the longest run was considered by Erdos and Revesz [6], Fu [8], and Lou [13] for a sequence of independent random variables and by Chryssaphinou and Vaggelatou [4, 26] and Zhang [27] for a Markov chain.

The distribution of the number of \( f \)-recurrent runs in a sequence of independent random variables was studied by Mikhailov [16, 17]. Similar results for the number of usual and \( f \)-recurrent runs with possible omissions of letters were obtained by Mezhennaya in [14] and [15], respectively. There are other generalizations of the problem about the distribution of the number of runs. For example, Minakov [22] discussed the distribution of the number of monotone tuples and runs in a finite ergodic Markov chain.
The problem under consideration is closely related to the problem of the number of repetitions of tuples in a random sequence. The limiting distribution of the number of pairs of identical tuples in a Markov chain was studied by Mikhailov and Shoitov [18, 20, 21], and the number of tuples with the same structure by Mikhailov [19].

The considered problem is also related to the description of properties of widely known scan statistics. According to [10, p. 58], scan statistic is defined as follows:

\[ T_{n,s}(u) = \max_{1 \leq i \leq n-k+1} \sum_{j=i}^{i+s-1} I\{X_j > u\} \]

(in what follows \( I_A \) is an indicator of a random event \( A \)). The value of \( T_{n,s}(u) \) is equal to the maximum number of exceedance occurrences by the random values \( \{X_j, j = 1, \ldots, n\} \) of the threshold value \( u \) among all moving windows of length \( s \).

We denote

\[ A_i = \begin{cases} \sum_{j=i}^{i+s-1} I\{X_j > u\} = l - 1, & \sum_{j=i+1}^{i+s} I\{X_j > u\} = l \end{cases} \]

Then

\[ \{T_{n,s}(u) < l\} = \left\{ \sum_{i=1}^{n-s} I_{A_i} = 0 \right\} \]

The distributions (both exact and asymptotic) of scan statistics for a sequence of independent random variables have been well studied [10], and their applications in various applied problems [10, 2] are also widely discussed.

In this paper, we estimate the total variation distance between the distribution of the number of appearances of events \( A_j \) in a segment of Markov chain and the accompanying Poisson distribution. As a result the Poisson and normal limit theorem for the random variable under consideration will be derived.

2. Main Results

Let \( \{X_j, j = 1, \ldots, n\} \) be a strongly ergodic stationary Markov chain with the states set \( A_N = \{1, \ldots, N\}, \ N \geq 2 \), with the transition probability matrix \( P = ||p_{a,b}||_{a,b \in A_N} \) and stationary distribution \( \{\pi_a, a \in A_N\} \). The elements of the matrix \( P^n \) are denoted by \( p_{a,b}^{(n)} \), for \( n = 1 \) \( p_{a,b}^{(1)} = p_{a,b} \).
According to [12, Cor. 4.1.5, p. 71] there are constants $C, \gamma > 0$ such as
\[ |\mathbb{P}^{(n)}_{a,b} - \pi_b| \leq C \pi_b e^{-\gamma n}, \quad n \geq 1. \]  

(1)

We assume that the random event $A_j$ depends only on the random variables $X_j, \ldots, X_{j+s}$, $s \geq 1$, and that the set of events $\{ A_j, j = 1, \ldots, n - s \}$ is homogeneous and has the property $\mathbb{P}\{ A_i | A_j \} = 0$ for $|i - j| \leq s$.

Let $\Gamma = \{ 1, \ldots, n - s \}$, $\{ \alpha_j = I_{A_j} : j \in \Gamma \}$ be the set of random indicators corresponding to the random events $\{ A_j, j \in \Gamma \}$, and $Q_s = \mathbb{P}\{ A_j \}$ be the probability of any event from the set $\{ A_j, j \in \Gamma \}$.

We define the random variable $\xi = \sum_{j=1}^{n-s} \alpha_j$, which is equal to the number of appearances of events $A_j$ in $\{ X_j, j = 1, \ldots, n \}$, and its expectation
\[ \lambda_s = \mathbb{E}\xi = (n - s)Q_s. \]  

(2)

The following notation will be used: $\mathcal{L}(\xi)$ for the distribution of the random variable $\xi$, $\text{Pois}(\lambda)$ for the Poisson distribution with parameter $\lambda$, $\mathcal{N}(0,1)$ for the standard normal distribution, and $\rho(\mathcal{L}(\xi), \mathcal{L}(\eta))$ for the total variation distance between $\mathcal{L}(\xi)$ and $\mathcal{L}(\eta)$, respectively. For non-negative integer random variables $\xi$ and $\eta$ the total variation distance is given by the formula
\[ \rho(\mathcal{L}(\xi), \mathcal{L}(\eta)) = \frac{1}{2} \sum_{u=0}^{\infty} |\mathbb{P}\{ \xi = u \} - \mathbb{P}\{ \eta = u \}|. \]

**Theorem 1.** Let $s, m \geq 1$ and $\lambda_s \geq 1$. Then
\[ \rho(\mathcal{L}(\xi), \text{Pois}(\lambda_s)) \leq \left( 2(s + 2m) + 1 + \frac{2C}{e^{\gamma} - 1} \right) \frac{\lambda_s}{n - s} \]
\[ + Ce^{-\gamma(m+1)} \sqrt{\lambda_s} \left( 2 + Ce^{-\gamma(m+1)} + e^{-\gamma(s+m+1)} \right), \]
where the constants $C$ and $\gamma$ are defined in (1).

**Corollary 2.** Let $s, n \to \infty$, such that $s/n \to 0$, $Q_s \to 0$, $\lambda_s \to \lambda \in (0, \infty)$. Then $\mathcal{L}(\xi) \to \text{Pois}(\lambda)$.

**Corollary 3.** Let $s, n \to \infty$, such that $\lambda_s \to \infty$, $s\lambda_s/n \to 0$, $\lambda_s = o(e^{\gamma s})$. Then $\mathcal{L}((\xi - \lambda_s)/\sqrt{\lambda_s}) \to \mathcal{N}(0,1)$.

Corollaries 2 and 3 follow immediately from Theorem 1.
3. Proof of Theorem 1

We will use Theorem 1.A of [3, p. 9] and the proof scheme proposed by Mikhailov and Shoitov [21] and Minakov [22] to estimate the total variation distance between the distribution of the random variable $\xi$ and the accompanying Poisson distribution (i.e., the Poisson distribution with parameter $\lambda_s$).

For each $i \in \Gamma$ we define the set of indices $\Gamma^s_i = \{ j \in \Gamma \setminus \{i\} \}$ such that $\alpha_i$ and $\alpha_j$ are strongly dependent. The remaining indices are assigned to the set $\Gamma^w_i = \Gamma \setminus \{ \Gamma^s_i \cup \{i\} \}$, which is called a set of weak dependence for a random indicator $\alpha_i$. We present the formulation of Theorem 1.A from [3] for our case.

**Theorem 4.** For each $i \in \Gamma$, let the set $\Gamma \setminus \{i\}$ be split into the disjoint subsets $\Gamma^s_i$ and $\Gamma^w_i$. Then

$$\rho(\mathcal{L}(\xi), \text{Pois}(\lambda_s)) \leq \min \left\{ 1, \frac{1}{\lambda_s} \right\} (S_1 + S_2) + \min \left\{ 1, \frac{1}{\sqrt{\lambda_s}} \right\} S_3,$$

where

$$S_1 = \sum_{i \in \Gamma} \sum_{j \in \Gamma^s_i \cup \{i\}} \mathbb{E}\alpha_i \mathbb{E}\alpha_j,$$

$$S_2 = \sum_{i \in \Gamma} \sum_{j \in \Gamma^s_i} \mathbb{E}\alpha_i \mathbb{E}\alpha_j,$$

$$S_3 = \sum_{i \in \Gamma} \mathbb{E}\left| \mathbb{E}\alpha_i - \mathbb{E}(\alpha_i | \{\alpha_j, j \in \Gamma^w_i\}) \right|.$$

Let $m \geq 1$. We put $\Gamma^s_i = \{ j \in \Gamma : 1 \leq |i - j| \leq s + m \}$, $\Gamma^w_i = \Gamma \setminus (\{i\} \cup \Gamma^s_i)$, and estimate all summands in (3).

Let us begin with the sum $S_1$, which is given by (4). Since $|\Gamma^s_i \cup \{i\}| \leq 2(s + m) + 1$ and $|\Gamma| = n - s$, then considering the definition (2), we derive

$$S_1 \leq (2(s + m) + 1)(n - s)Q_s^2 = (2(s + m) + 1)\lambda_s Q_s.$$  

(7)

Next, we turn to $S_2$ (see (5)). The incompatibility of events $A_i$ and $A_j$ with $|i - j| \leq s$ leads to

$$S_2 \leq 2 \sum_{i \in \Gamma} \sum_{j = i + s + 1}^{i + s + m} \mathbb{E}\alpha_i \alpha_j.$$

Due to the Markov property we obtain

$$\mathbb{E}\alpha_i \alpha_j = \mathbb{P}\{A_i A_j\} = \mathbb{P}\{A_i\} \mathbb{P}\{A_j|A_i\} \leq Q_s \max_{a \in A_N} \mathbb{P}\{A_j|X_{i+s} = a\}.$$  

(8)
Then
\[ P\{A_j|X_{i+s} = a\} \leq \max_{b \in \mathcal{A}_N} p_{a,b}^{(j-i-s)} \frac{1}{\pi_b} Q_s \]
\[ \leq \max_{b \in \mathcal{A}_N} \pi_b (1 + C e^{-\gamma(j-i-s)}) \frac{1}{\pi_b} Q_s = (1 + C e^{-\gamma(j-i-s)}) Q_s. \]

Substituting this estimate into (8), we obtain
\[ E\alpha_i\alpha_j \leq (1 + C e^{-\gamma(j-i-s)}) Q_s^2. \]

Then, it follows from (5) that
\[ S_2 \leq 2(n-s)Q_s^2 \sum_{k=1}^{m} (1 + C e^{-\gamma k}) = 2(n-s)Q_s^2 \left( m + C \frac{1 - e^{-\gamma m}}{e^\gamma - 1} \right). \]

Applying definition (2) to the last estimate produces
\[ S_2 \leq 2\lambda_s Q_s \left( m + \frac{C}{e^\gamma - 1} \right). \tag{9} \]

To estimate the sum $S_3$ we use the scheme proposed by Mikhailov and Shoitov [21] and Minakov [22]. We start with a separate summand of $S_3$, which we denote as $s_{3,i}$ as follows:
\[ s_{3,i} = E|Q_s - E(\alpha_i|\{\alpha_j, j \in \Gamma_i^w\})|. \]

According to the Markov property there are three possible cases:

a) for $m+2 \leq i \leq n-s-m-1$
\[ s_{3,i} = E|Q_s - E(\alpha_i|\{X_{i-m-1}, X_{i+s+m+1}\})|; \]

b) for $1 \leq i \leq m+1$
\[ s_{3,i} = E|Q_s - E(\alpha_i|X_{i+s+m+1})|; \]

c) for $n-s-m \leq i \leq n-s$
\[ s_{3,i} = E|Q_s - E(\alpha_i|X_{i-m-1})|. \]

First, for case a) we have
\[ s_{3,i} = \sum_{a,b \in \mathcal{A}_N} |Q_s - E(\alpha_i|X_{i-m-1} = a, X_{i+s+m+1} = b)| \times \]
\[ \times P\{X_{i-m-1} = a, X_{i+s+m+1} = b\} = \]
ON THE NUMBER OF EVENT APPEARANCES...

\[
\sum_{a,b \in A_N} |Q_s P\{X_{i-m-1} = a, X_{i+s+m+1} = b\} - P\{A_i, X_{i-m-1} = a, X_{i+s+m+1} = b\}|. \tag{10}
\]

Since \(P\{X_{i-m-1} = a, X_{i+s+m+1} = b\} = \pi_a p_{a,b}^{s+2(m+1)}\), then the use of (1) leads to the following estimate

\[
|p_{a,b}^{s+2(m+1)} - \pi_b| \leq C \pi_b e^{-\gamma(s+2(m+1))}.
\]

Now we can find the upper and lower bounds for the second probability in (10):

\[
P\{A_i, X_{i-m-1} = a, X_{i+s+m+1} = b\} \leq \pi_a \max_{u,v \in A_N} p_{u,v}^{(m+1)} \pi_u Q_s p_{v,b}^{(m+1)}
\leq \pi_a Q_s \max_{u,v \in A_N} \pi_u (1 + C e^{-\gamma(m+1)}) \frac{1}{\pi_u} \pi_b (1 + C e^{-\gamma(m+1)})
= \pi_a \pi_b Q_s (1 + C e^{-\gamma(m+1)})^2. \tag{11}
\]

Analogously,

\[
P\{A_i, X_{i-m-1} = a, X_{i+s+m+1} = b\} \geq \pi_a \min_{u,v \in A_N} p_{u,v}^{(m+1)} \pi_u Q_s p_{v,b}^{(m+1)} q
\geq \pi_a Q_s \min_{u,v \in A_N} \pi_u (1 - C e^{-\gamma(m+1)}) \frac{1}{\pi_u} \pi_b (1 - C e^{-\gamma(m+1)})
= \pi_a \pi_b Q_s (1 - C e^{-\gamma(m+1)})^2. \tag{12}
\]

Substituting (11) and (12) into (10), we obtain

\[
s_{3,i} \leq \sum_{a,b \in A_N} \pi_a \pi_b Q_s \left[ (1 + C e^{-\gamma(m+1)})^2 - (1 - C e^{-\gamma(s+2(m+1))}) \right]
\leq C e^{-\gamma(m+1)} Q_s \left[ 2 + C e^{-\gamma(m+1)} + e^{-\gamma(s+m+1)} \right]. \tag{13}
\]

The number of such summands in the sum \(S_3\) is \(n - s - 2m - 2\).

We turn to case b). Similarly to case a) we derive

\[
s_{3,i} = \sum_{a \in A_N} |Q_s - E(\alpha_i | X_{i+s+m+1} = a)| P\{X_{i+s+m+1} = a\}
= \sum_{a \in A_N} |Q_s P\{X_{i+s+m+1} = a\} - P\{A_i, X_{i+s+m+1} = a\}|
\]
Now we write the upper and lower bounds for the second probability in (14):

\[ P\{A_i, X_{i+s+m+1} = a\} \leq \max_{v \in \mathcal{A}_N} Q_s p_{v,a}^{(m+1)} \]

\[ \leq Q_s \max_{u,v \in \mathcal{A}_N} \pi_a (1 + C e^{-\gamma(m+1)}) = \pi_a Q_s (1 + C e^{-\gamma(m+1)}), \]

\[ P\{A_i, X_{i+s+m+1} = a\} \geq \min_{v \in \mathcal{A}_N} Q_s p_{v,a}^{(m+1)} \geq \pi_a Q_s (1 - C e^{-\gamma(m+1)}). \]

From the last two formulas we derive

\[ s_{3,i} \leq \sum_{a \in \mathcal{A}_N} \pi_a Q_s \left[ (1 + C e^{-\gamma(m+1)}) - 1 \right] = C e^{-\gamma(m+1)} Q_s. \tag{15} \]

Similar calculations in case c) lead to the same estimate. The total number of summands in cases b) and c) is \(2m + 2\). Substituting the estimates of the terms (13) and (15) into (6), we obtain the following estimator:

\[ S_3 \leq C e^{-\gamma(m+1)} Q_s (n - s - 2m - 2) \left( 2 + C e^{-\gamma(m+1)} + e^{-\gamma(s+m+1)} \right) \]

\[ + C e^{-\gamma(m+1)} Q_s (2m + 2) \]

\[ \leq C e^{-\gamma(m+1)} \lambda_s \left( 2 + C e^{-\gamma(m+1)} + e^{-\gamma(s+m+1)} \right). \tag{16} \]

The statement of the theorem arises from the formulas (7), (9) and (16). The proof of Theorem 1 is complete.

Acknowledgements

The author is grateful to V.G. Mikhailov for the helpful suggestions and remarks.

References

ON THE NUMBER OF EVENT APPEARANCES... 545


