On the number of word occurrences in a semi-Markov sequence of letters

Margarita Karaliopoulou

ESAIM: Probability and Statistics Will be set by the publisher
URL: http://www.emath.fr/ps/

ON THE NUMBER OF WORD OCCURRENCES IN A SEMI-MARKOV SEQUENCE OF LETTERS

Karaliopoulou Margarita[1]

Abstract. Let a finite alphabet Ω. We consider a sequence of letters from Ω generated by a discrete time semi-Markov process $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$. We derive the probability of a word occurrence in the sequence. We also obtain results for the mean and variance of the number of overlapping occurrences of a word in a finite discrete time semi-Markov sequence of letters under certain conditions.

1991 Mathematics Subject Classification. 60K10, 60K20, 60C05, 60E05.

The dates will be set by the publisher.

Introduction

Word occurrences in a sequence of letters are known to play an important role in molecular biology, in computer science, telecommunications and generally in applications of sequential analysis. Statistical and probabilistic properties of words have been studied when the underlying model is a bernoulli model and an homogeneous m-order Markov chain, $m \in \mathbb{N}$, $m \geq 1$ ($\mathbb{N} = \{0, 1, 2, \ldots\}$). An overview on these properties of words was recently given by Reinert et al. [5].

In this paper we shall give some results in a probabilistic framework when the underlying model is a discrete time semi-Markov (DTSM) process $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ with finite state space an alphabet Ω. Note that Discrete time semi-Markov processes are processes that generalize discrete time Markov chains and discrete time renewal processes. Under the Markovian hypothesis the distribution of the sojourn time in a state is geometrically distributed. In the semi-Markov case we allow arbitrarily distributed sojourn times in any state.

Recently, Barbu et al. [1] studied a discrete time semi-Markov process $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ in order to derive some specific reliability measurements. Chryssaphinou et al. [2] defining the process $\{U_\gamma;\ \gamma \in \mathbb{N}\}$ to be the backward recurrence times of the DTSM process $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ studied the Markov process $\{(Z_\gamma,\ U_\gamma);\ \gamma \in \mathbb{N}\}$. As an application they considered a finite set of words W of equal length which are produced under the semi-Markovian hypothesis and focused on the waiting time for the first word occurrence from the set W in the semi-Markov sequence of letters. The time spent in the same state (that is, the number of occurrences of the same letter in succesion at the semi-Markov sequence) is included in the formation of the words. We clarify that one step transitions are not allowed from a state to itself in the embedded discrete time Markov chain.

Stefanov [6], [7] studied waiting time problems for discrete time semi-Markov processes with finite number of states. However, the time spent in a state is not included in the formation of the words.

Keywords and phrases: Discrete time Semi-Markov, number of word occurrences

[1] University of Athens , Department of Mathematics, 157 84, Athens, Greece. E-mail: mkaraliop@math.uoa.gr, chronis@ath.forthnet.gr

This research was supported by the program "Platon" between University of Athens and Université de Technologie de Compiègne.

In this paper, we derive the probability of a word occurrence in the semi-Markov sequence when the time spent in a state is included in the formation of the words. This result generalizes the well known formula for the probability of a word occurrence in a Markovian sequence of letters, see [5].

Considering the number of overlapping occurrences of a word we obtain the mean number. The calculation of variance is achieved for the case of having a word which is not a run. The difficulty of the case of having a run arises from the following fact. It is well known that in the semi-Markov case the future depends both on the present state and on the time the process has been in that state. When a not run word occurs we know the backward recurrence time of the last letter of the word. When we have a run we don't have that information. Thus, the calculation of the corresponding covariances is restricted to the case of words which are not runs.

To illustrate our notations we use the DNA alphabet which is convenient. We note that this use does not mean that we examine a specific application of DNA sequences in this paper. However, it is well known that the study of occurrences of words is closely related to reliability problems, as well as to DNA problems. Thus, we conjecture that the theoretical results of this paper could be applied in such problems.

The paper is organized as follows: In section 1, we give the necessary notation for our DTSM model as well as notation relevant to word occurrences. We also present some theorems, which have been proved by Chryssaphinou et al. [2] concerning the Markov process $\{(Z_\gamma,\ U_\gamma);\ \gamma \in \mathbb{N}\}$ and are used in order to prove the new results of this paper. In section 2, we derive the probability of a word occurrence in a DTSM model. We apply this result when the underlying model is a Markovian sequence and we get the expected formula. In section 3 we give explicit formulae for the mean and variance of the number of overlapping occurrences of a word in the discrete time semi-Markov sequence of letters under certain conditions.

1. Notation

1.1. The discrete time semi-Markov model

Let us consider a sequence of outcomes $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ generated by a semi-Markov chain with state space $\Omega = \{\alpha_1,\ \ldots, \alpha_\ell\}$, $2 \le \ell < \infty$. It is $\alpha_i \neq \alpha_j$ for all $\alpha_i,\ \alpha_j \in \Omega,\ \ i \neq j$. We set $S_0 = 0$ and

$$S_{n+1} = \inf\{k > S_n : Z_k \neq Z_{S_n}\}, \qquad n = 1, 2, \ldots\ ,$$

provided that $S_n < \infty$, the jump times of the process.

The process $\{J_n;\ n \in \mathbb{N}\}$, with $J_n = Z_{S_n}$, is the embedded Markov chain of the semi-Markov process.

Example 1.1. Let $\Omega = \{A, C, G, T\}$. Considering the realization of the semi-Markov sequence

$$AAAACCAAAAATGGTTA \ldots\ ,$$

we obtain $J_0 = A,\ S_0 = 0,\ J_1 = C,\ S_1 = 4,\ J_2 = A,\ S_2 = 6,\ J_3 = T,\ S_3 = 11,\ \ldots\ .$

The stochastic process $\{(J_n,\ S_n);\ n \in \mathbb{N}\}$ is the associated Discrete Time Markov Renewal Process (DTMRP) of the semi-Markov chain $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$. We shall consider that $\{(J_n,\ S_n);\ n \in \mathbb{N}\}$ is homogeneous with discrete time semi-Markov kernel $q = \{q(\gamma);\ \gamma \in \mathbb{N}\}$, where $q(\gamma) = (q(\alpha, \alpha', \gamma))_{\alpha,\alpha' \in \Omega}$ and

$$q(\alpha, \alpha', \gamma) := \mathbb{P}(J_{n+1} = \alpha',\ S_{n+1} - S_n = \gamma \,|\, J_n = \alpha), \quad \forall\, n \in \mathbb{N}, \quad \text{for } \gamma = 1,\ 2,\ \ldots. \tag{1.1}$$

Moreover, $q(\alpha, \alpha', 0) = 0$ for every $\alpha, \alpha' \in \Omega$.

From the definition of our model we get that $q(\alpha, \alpha, \gamma) = 0$, for all $\alpha \in \Omega,\ \gamma \in \mathbb{N}$.

The r-fold convolution of the semi-Markov kernel q of a DTMRP $\{(J_n,\ S_n);\ n \in \mathbb{N}\}$ is

$$q^{(r)}(\alpha, \alpha', \gamma) = \mathbb{P}(J_r = \alpha',\ S_r = \gamma \,|\, J_0 = \alpha), \tag{1.2}$$

for all $\alpha, \alpha' \in \Omega$, for all $r = 1,\ 2,\ \ldots$ and $\gamma \in \mathbb{N}$. Moreover,

$$q^{(0)}(\alpha, \alpha', \gamma) = \begin{cases} 1, & \text{for } \gamma = 0 \text{ and } \alpha = \alpha' \\ 0, & \text{elsewhere.} \end{cases} \tag{1.3}$$

We let

$$\psi(\alpha, \alpha', \gamma) = \sum_{r=0}^{\gamma} q^{(r)}(\alpha, \alpha', \gamma)\ , \quad \alpha, \alpha' \in \Omega,\ \gamma \in \mathbb{N}. \tag{1.4}$$

We also consider the quantities

$$H(\alpha, \gamma) = \sum_{n=0}^{\gamma} \sum_{\alpha' \in \Omega} q(\alpha, \alpha', n) \qquad \text{and} \qquad m_\alpha := \sum_{\gamma \geq 0} [1 - H(\alpha, \gamma)], \quad \forall\ \alpha \in \Omega,\ \gamma \in \mathbb{N} \tag{1.5}$$

which express the distribution function of the sojourn time in state α and the corresponding mean respectively.

Furthermore, Chryssaphinou et al. [2] defined $\{U_\gamma;\ \gamma \in \mathbb{N}\}$ the backward recurrence time for the semi-Markov process $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ by

$$U_\gamma = \gamma - [\sup\{u < \gamma :\ Z_u \neq Z_\gamma\} + 1]. \tag{1.6}$$

Generally, we make the convention that

$$\sup\{u < \gamma :\ Z_u \neq Z_\gamma\} = -1 - U_0 \quad \text{for } 0 \leq \gamma < S_1,\ \gamma \in \mathbb{N}$$

and that for $S_0 = 0$ we set $U_0 = 0$. Generally, when observing a semi-Markov sequence at time 0, the variable U_0 could be not equal to 0. For example, if the Markov Renewal Process was a delayed process it would be $U_0 \neq 0$.

We note that if $Z_{\gamma-1} \neq Z_\gamma$ then it is $U_\gamma = 0$. Moreover if $Z_{\gamma-1} = Z_\gamma$ then $U_\gamma = U_{\gamma-1} + 1$ for all $\gamma \in \mathbb{N}$.

The following examples clarify the above definition for the backward recurrence time.

Example 1.2. Let $\Omega = \{A, C, G, T\}$. For $S_0 = 0$, let the sequence of letters $AAAACCAAAAATGGTTA\ldots$. We see that $U_0 = 0,\ U_1 = 1,\ U_2 = 2,\ U_3 = 3,\ U_4 = 0,\ U_5 = 1,\ U_6 = 0,\ U_7 = 1,\ \ldots$.

Example 1.3. Let $\Omega = \{0, 1\}$. For $S_0 = 0$, let the sequence $00111010\ldots$. It is $U_0 = 0,\ U_1 = 1,\ U_2 = 0,\ U_3 = 1,\ U_4 = 2,\ U_5 = 0,\ U_6 = 0,\ U_7 = 0,\ \ldots$.

The stochastic process $\{(Z_\gamma,\ U_\gamma);\ \gamma \in \mathbb{N}\}$ with state space $\Theta = \cup_{\alpha \in \Omega} \Theta\alpha$, where

$$\Theta_\alpha := \{(\alpha, u) \in\ \Omega \times \mathbb{N}\ : H(\alpha, u) < 1\} \quad \forall\ \alpha \in \Omega,$$

is a time homogeneous Markov process with transition probabilities $\hat{p}((\alpha, u), (\alpha', u'))$, $(\alpha, u), (\alpha, u') \in \Theta$. Its first order transition probabilities are given in the following theorem.

Theorem 1.4. *(see [2]) The first order transition probabilities of the Markov chain $\{(Z_\gamma,\ U_\gamma);\ \gamma \in \mathbb{N}\}$ are*

$$\hat{p}((\alpha, u), (\alpha', u')) = \begin{cases} \dfrac{q(\alpha, \alpha', u+1)}{1 - H(\alpha, u)}, & \text{if} \quad u' = 0, \\ \dfrac{I_{\{\alpha = \alpha'\}}\ [1 - H(\alpha, u+1)]}{1 - H(\alpha, u)}, & \text{if} \quad u' = u + 1, \\ 0, & \text{elsewhere} \end{cases}$$

for every $(\alpha, u),\ (\alpha', u') \in \Theta$.

Theorem 1.5. *(see [2]) The transition probabilities of γ order, $\gamma = 2,\ 3, \ldots,$ are*

$$\hat{p}^{\gamma}((\alpha,u),(\alpha',u')) = I_{\{\gamma=u'-u\}} I_{\{\alpha=\alpha'\}} \frac{1-H_(\alpha',u')}{1-H(\alpha,u)} + I_{\{\gamma\geq u'+1\}} \frac{\sum_{c\in\Omega}\sum_{n=1}^{\gamma-u'} q(\alpha,c,n+u)\psi(c,\alpha',\gamma-u'-n)[1-H(\alpha',u')]}{1-H(\alpha,u)},$$

for all $(\alpha,u),(\alpha,u') \in \Theta$.

Moreover, for all $(\alpha,u),(\alpha',u') \in \Theta$, we denote

$$\hat{p}^{0}((\alpha,u),(\alpha',u')) = I_{\{(\alpha,u)=(\alpha',u')\}} \quad \text{and} \quad \hat{p}^{1}((\alpha,u),(\alpha',u')) = \hat{p}((\alpha,u),(\alpha',u')). \tag{1.7}$$

Finally, denoting by μ_α the mean recurrence time of the state α for the semi-Markov process $\{Z_\gamma;\ \gamma\in\mathbb{N}\}$ the following theorem is valid.

Theorem 1.6. *(see [2]) Let the DTMRP $\{(J_n,\ S_n);\ n\in\mathbb{N}\}$ with finite state space Ω be irreducible, aperiodic and $m_\alpha < \infty$ for all $\alpha\in\Omega$. Then the function $\pi(\alpha,u) := \dfrac{1-H(\alpha,u)}{\mu_\alpha}$, $(\alpha,u)\in\Theta$ is a stationary probability distribution for the Markov process $\{(Z_\gamma,\ U_\gamma);\ \gamma\in\mathbb{N}\}$.*

1.2. **The words**

Let us call the set $\Omega = \{\alpha_1,\ \ldots, \alpha_\ell\}$ alphabet and its elements letters. A word over the alphabet Ω is a finite sequence of elements of Ω. Following the notation of Lothaire [4] we define the set of words over the alphabet Ω

$$\Omega^{+} = \{w\ :\ w = c_1\ \ldots c_k, \qquad c_1,\ \ldots,\ c_k\in\Omega, \quad k\in\mathbb{N},\ 1\leq k<\infty\}. \tag{1.8}$$

Note that k denotes the length of word $w\in\Omega^+$.

Let $w\in\Omega^+$. If $w = \underbrace{b\ \ldots\ b}_{k}$ that is a run of b's of length k we shall use this presentation

$$w = b^{(k)}, \quad b\ \in\Omega, \quad k\in\mathbb{N},\ 1\leq k<\infty \tag{1.9}$$

If w is not a run, it can be obtained by concatenating words which are runs. That is word w of length k begins with a run in b_1's of length k_1 followed by another run in b_2's of length k_2 and at the end there is a run of b_η's of length k_η. The variable η, $\eta\in\mathbb{N}$, $1\leq\eta\leq k$, indicates the number of letter renewals in word w. Therefore every word w with $\eta>1$ "contains" η word-runs:

$$w = w(1)w(2)\ldots w(\eta), \tag{1.10}$$

where $w(1) := {b_1}^{(k_1)},\ w(2) := {b_2}^{(k_2)},\ \ldots,\ w(\eta) := {b_\eta}^{(k_\eta)}$.

Thus, in the sequel, we shall use the following presentation of a word w from the set Ω^+, which will be proved useful in the study of word occurrences in a discrete time semi-Markov sequence of letters. We write

$$w = {b_1}^{(k_1)}\ldots {b_\eta}^{(k_\eta)}, \tag{1.11}$$

where $b_1,\ \ldots,\ b_\eta\in\Omega, \quad b_j\neq b_{j+1},$ for $j=1,\ \ldots,\ \eta-1$ and $k_1+\ \ldots+\ k_\eta = k$, where $k_1,\ \ldots,\ k_\eta\ \in\mathbb{N}$.

For example, let $\Omega = \{A, C, G, T\}$ and the word $w = AGGTAAA$, written in the common presentation, where $k = 7$. According to (1.11) it is $\eta = 4$ and $w = A^{(1)}G^{(2)}T^{(1)}A^{(3)}$, with

$$w(1) = A^{(1)},\ b_1 = A,\ k_1 = 1, \quad w(2) = G^{(2)},\ b_2 = G,\ k_2 = 2,$$
$$w(3) = T^{(1)},\ b_3 = T,\ k_3 = 1, \quad w(4) = A^{(3)},\ b_4 = A,\ k_4 = 3.$$

Considering a word $w \in \Omega^+$, let $[w]_\gamma = c_{k-\gamma+1} \dots c_k$ a suffix of the word w of length γ for all $\gamma = 1,\ \dots,\ k$. Note that $[w]_1 = b_k$ and $[w]_k = w$. Obviously, $[w]_\gamma \in \Omega^+$. Therefore, using the representation (1.11) of a word, it is

$$[w]_\gamma = \begin{cases} {b_{\zeta-1}}^{(k^*_\gamma)} {b_\zeta}^{(k_\zeta)} \ \dots {b_\eta}^{(k_\eta)}, & \text{for } \gamma > k_\eta \\ \\ {b_\eta}^{(\gamma)}, & \text{elsewhere,} \end{cases}$$

where $\zeta = \min\{j\ :\ k_j + \dots + k_\eta \le \gamma\}$ and $k^*_\gamma + k_\zeta + \dots + k_\eta = \gamma$ with $0 \le k^*_\gamma < k_{\zeta-1}$.

For example considering the word $w = A^{(1)}G^{(2)}T^{(1)}A^{(3)}$, for $\gamma = 5$ it is $\zeta = 3$ and $k^*_\gamma = \gamma - (k_3 + k_4) = 5 - (1 + 3) = 1$. Thus we have $[w]_5 = G^{(1)}T^{(1)}A^{(3)}$.

Guibas and Odlyzko [3] gave the following useful definition concerning the periods of a word.

Let a word $w = c_1 \dots c_k$ written in the common presentation. The number ρ, $\rho \in \{1,\ \dots,\ k-1\}$, is called a period of word w if $c_i = c_{i+\rho}$, for all $i = 1,\ \dots,\ k - \rho$. The set of all periods of w is denoted $\mathcal{P}(w)$.

2. The probability of a word occurrence in the semi-Markov sequence of letters

Let a word $w \in \Omega^+$. We shall say that the word $w = c_1 \dots c_k$ occurs at time γ in the semi-Markov sequence $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ if and only if

$$Z_{\gamma-k+1} = c_1,\ \dots,\ Z_\gamma = c_k.$$

For every $w \in \Omega^+$ and $\gamma \in \mathbb{N}$ let us consider the event

$$\mathcal{E}_{w;\gamma} := \{w \text{ occurs at } \gamma\} \tag{2.1}$$

and the associated random indicator

$$E_{w;\gamma} := I_{\{w \text{ occurs at } \gamma\}} \tag{2.2}$$

For all $w \in \Omega^+$ we have

$$\mathbb{P}(\mathcal{E}_{w;\gamma}) = \begin{cases} 0, & \text{for } \gamma < k-1 \\ \\ \displaystyle\sum_{(\alpha,u)\in\Theta} \mathbb{P}(\mathcal{E}_{w;\gamma} \,|\, Z_0 = \alpha, U_0 = u)\mathbb{P}(Z_0 = \alpha, U_0 = u), & \text{for } \gamma \ge k-1. \end{cases} \tag{2.3}$$

The following remarks are useful in our study.

Remark 2.1. Let a word $w \in \Omega^+$, $\gamma \ge k-1$. Recall from (1.10) that any word w is a concatenation of η words, which are runs. Thus, let us study at first a word w which is a run.

If w is a run, where $\eta = 1$, it is $w = w(1) = b^{(k)}$. From the definition of the backward recurrence time (1.6), since $Z_{\gamma-k+1} = Z_{\gamma-k+2} = \ \dots = Z_\gamma$ it is $U_{\gamma-k+2} = U_{\gamma-k+1} + 1$, $U_{\gamma-k+3} = U_{\gamma-k+2} + 1,\ \dots,\ U_\gamma = U_{\gamma-1} + 1$. Since we don't know the state of $Z_{\gamma-k}$, we have to consider all the possibilities for the $U_{\gamma-k+1}$, which may be $0, 1, 2, \dots$ thus

$$\mathcal{E}_{w;\gamma} = \cup_{s\ge 0}\mathcal{E}_{w;\gamma}(s), \tag{2.4}$$

where

$$\mathcal{E}_{w;\gamma}(s) := \{Z_{\gamma-k+1} = b,\ U_{\gamma-k+1} = s,\ \ldots,\ Z_\gamma = b,\ U_\gamma = s+k-1\}, \qquad s \in \mathbb{N}, \quad \gamma \geq k-1. \tag{2.5}$$

Now, if w is not a run it is $\eta > 1$. According to (1.10) it is $w = w(1)\ldots w(\eta)$ where $w(1),\ \ldots,\ w(\eta)$ are runs. Thus

$$\mathcal{E}_{w;\gamma} = \mathcal{E}_{w(1);\gamma-k+k_1} \cap \mathcal{E}_{w(2);\gamma-k+k_1+k_2} \cap\ \ldots \cap\ \mathcal{E}_{w(\eta);\gamma}.$$

For the second run, since $Z_{\gamma-k+k_1} = b_1$ and $Z_{\gamma-k+k_1+1} = b_2$, where $b_1 \neq b_2$ it is $Z_{\gamma-k+k_1} \neq Z_{\gamma-k+k_1+1}$. Using the definition of the backward recurrence time (1.6), it is $U_{\gamma-k+k_1+1} = 0$. Similarly $U_{\gamma-k+k_1+k_2+1} = 0,\ \ldots,\ U_{\gamma-k+k_1+\ldots k_{\eta-1}+1} = 0$. But for the first run we don't know the state of $Z_{\gamma-k}$. Therefore, considering (2.4) and (2.5) we write

$$\mathcal{E}_{w;\gamma} = [\cup_{s\geq 0}\mathcal{E}_{w(1);\gamma-k+k_1}(s)] \cap \mathcal{E}_{w(2);\gamma-k+k_1+k_2}(0) \cap\ \ldots \cap\ \mathcal{E}_{w(\eta);\gamma}(0), \tag{2.6}$$

where, for all $j = 1,\ \ldots,\ \eta$,

$$\mathcal{E}_{w(j);\gamma}(s) := \{Z_{\gamma-k_j+1} = b_j,\ U_{\gamma-k_j+1} = s,\ \ldots,\ Z_\gamma = b_j,\ U_\gamma = s+k_j-1\}, \qquad s \in \mathbb{N}, \quad \gamma \geq k_j - 1. \tag{2.7}$$

Remark 2.2. Let $w = b^{(k)}$ a run of b of length k. Using (2.5) it is

$$\mathbb{P}(\mathcal{E}_{w;\gamma}(s)) = \mathbb{P}(Z_{\gamma-k+1} = b,\ U_{\gamma-k+1} = s,\ \ldots,\ Z_\gamma = b,\ U_\gamma = s+k-1). \tag{2.8}$$

If $(b, s+k-1)\ \in \Theta$ then $(b, s+k-1-\kappa)\ \in \Theta$, for all $\kappa = 1,\ \ldots,\ k-1$, where $s \in \mathbb{N}$.

Let $\mathbb{P}(Z_{\gamma-k+1} = b,\ U_{\gamma-k+1} = s) \neq 0$. Since the process $\{(Z_\gamma,\ U_\gamma),\ \gamma \in \mathbb{N}\}$ with state space Θ is a time homogeneous Markov process with first order transition probabilities given by theorem 1.4, for $(b, s+k-1)\ \in \Theta$ we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{w;\gamma}(s)) &= \mathbb{P}(Z_{\gamma-k+1} = b,\ U_{\gamma-k+1} = s)\hat{p}((b,s),(b,s+1))\ldots\hat{p}((b,s+k-2),(b,s+k-1)) = \\
&= \mathbb{P}(Z_{\gamma-k+1} = b,\ U_{\gamma-k+1} = s)\frac{1-H(b,s+1)}{1-H(b,s)}\cdots\frac{1-H(b,s+k-1)}{1-H(b,s+k-2)} \\
&= \mathbb{P}(Z_{\gamma-k+1} = b,\ U_{\gamma-k+1} = s)\frac{1-H(b,s+k-1)}{1-H(b,s)}.
\end{aligned}$$

Therefore, from the above equation we have that the condition $(b, s+k-1)\ \in \Theta$ is necessary for $\mathbb{P}(\mathcal{E}_{w_j;\gamma}(s))$ to be positive.

Thus according to remarks 2.1 and 2.2, if $\eta > 1$ we shall consider words with $(b_2, k_2-1),\ \ldots,\ (b_\eta, k_\eta-1) \in \Theta$. Therefore we define the set of words $\Omega^+_{(\Theta)}$ as follows

$$\Omega^+_{(\Theta)} := \{w \in \Omega^+\ : \eta = 1 \text{ or } (\eta > 1 \text{ and } (b_2, k_2-1),\ \ldots,\ (b_\eta, k_\eta-1) \in \Theta)\}. \tag{2.9}$$

The following lemma is useful in proving our main result.

Lemma 2.3. *Let $w \in \Omega^+_{(\Theta)}$. The probability $\mathbb{P}(\mathcal{E}_{w;r+\gamma}\,|\,Z_r = \alpha,\ U_r = u)$ is independent of r for all $(\alpha, u) \in \Theta$ and $\gamma \geq k-1$.*

Proof. Assume $r \in \mathbb{N}$ such that $\mathbb{P}(Z_r = \alpha,\ U_r = u) > 0$.

I. If w is a run, that is $w = b^{(k)}$ and $\eta = 1$. Using remark 2.1 it is

$$\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, Z_r = \alpha,\ U_r = u) = \sum_{s\geq 0} \mathbb{P}(\mathcal{E}_{w;r+\gamma}(s) \ \,|\, Z_r = \alpha,\ U_r = u)$$

Since the process $\{(Z_\gamma,\ U_\gamma),\ \gamma \in \mathbb{N}\}$ is a time homogeneous with transition probabilities $\hat{p}(\cdot,\cdot)$, we have

$$\begin{aligned}&\mathbb{P}(\mathcal{E}_{w;r+\gamma} \ \,|\, Z_r = \alpha,\ U_r = u) = \\ &= \sum_{\substack{s\geq 0 \\ s:\ (b,s+k-1)\in\Theta}} \hat{p}^{\gamma-k+1}((\alpha,u),(b,s))\hat{p}((b,s),(b,s+1))\dots\hat{p}((b,s+k-2),(b,s+k-1)).\end{aligned} \tag{2.10}$$

II. If w is not a run, that is $w = {b_1}^{(k_1)}\dots{b_\eta}^{(k_\eta)}$, $\eta > 1$, using remark 2.1 we have

$$\begin{aligned}&\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, Z_r = \alpha,\ U_r = u) = \\ &= \sum_{s\geq 0} \mathbb{P}([\mathcal{E}_{w(1);r+\gamma-k+k_1}(s) \cap \mathcal{E}_{w(2);r+\gamma-k+k_1+k_2}(0) \cap\ \dots \cap\ \mathcal{E}_{w(\eta);r+\gamma}(0)] \,|\, Z_r = \alpha,\ U_r = u).\end{aligned}$$

Since the process $\{(Z_\gamma,\ U_\gamma),\ \gamma \in \mathbb{N}\}$ is a time homogeneous with transition probabilities $\hat{p}(\cdot,\cdot)$, we have

$$\begin{aligned}&\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, Z_r = \alpha,\ U_r = u) = \\ &= \sum_{\substack{s\geq 0 \\ s:\ (b_1,s+k_1-1)\ \in\Theta}} \hat{p}^{\gamma-k+1}((\alpha,u),(b_1,s))\hat{p}((b_1,s),(b_1,s+1))\dots\hat{p}((b_1,s+k_1-2),(b_1,s+k_1-1)) \\ &\qquad \hat{p}((b_1,s+k_1-1),(b_2,0))\hat{p}((b_2,0),(b_2,1))\dots\hat{p}((b_2,k_2-2),(b_2,k_2-1)) \\ &\qquad\qquad \dots \\ &\qquad \hat{p}((b_{\eta-1},k_{\eta-1}-1),(b_\eta,0))\hat{p}((b_\eta,0),(b_\eta,1))\dots\hat{p}((b_\eta,k_\eta-2),(b_\eta,k_\eta-1)).\end{aligned} \tag{2.11}$$

□

Since the above probabilities are independent of r we shall write

$$f_{(\alpha,u)}(w;\gamma) := \mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, Z_r = \alpha,\ U_r = u), \quad \forall\,(\alpha,u) \in \Theta, \quad w \in \Omega^+_{(\Theta)} \ \text{ and } \gamma \geq k-1,\ \forall\ r \in \mathbb{N}. \tag{2.12}$$

Now, we proceed to our main result.

Theorem 2.4. *Let $w \in \Omega^+_{(\Theta)}$, $\quad (\alpha,u) \in \Theta$ and $\gamma \geq k-1$. It is*

$$f_{(\alpha,u)}(w;\gamma) = \begin{cases} \Big\{ I_{\{\alpha=b\}}[1-H(b,u+\gamma)]+ \\ \quad +\sum_{c\in\Omega}\sum_{s=0}^{\gamma-k}\sum_{n=1}^{\gamma-k+1-s} q(\alpha,c,n+u)\psi(c,b,\gamma-k+1-s-n) \\ \quad [1-H(b,s+k-1)]\Big\}\dfrac{\sigma(w)}{1-H(\alpha,u)}, & \text{for } \ \eta = 1 \\ \\ \Big\{ I_{\{\alpha=b_1\}}q(b_1,b_2,\gamma-k+1+k_1+u)+ \\ \quad +\sum_{c\in\Omega}\sum_{s=0}^{\gamma-k}\sum_{n=1}^{\gamma-k+1-s} q(\alpha,c,n+u)\psi(c,b_1,\gamma-k+1-s-n) \\ \quad q(b_1,b_2,s+k_1)\Big\}\dfrac{\sigma(w)}{1-H(\alpha,u)}, & \text{for } \ \eta > 1, \end{cases}$$

where

$$\sigma(w) = \begin{cases} q(b_2, b_3, k_2) \cdot \ldots \cdot q(b_{\eta-1}, b_\eta, k_{\eta-1}) \cdot [1 - H(b_\eta, k_\eta - 1)], & \text{for } \eta > 2 \\ 1 - H(b_2, k_2 - 1), & \text{for } \eta = 2 \\ 1, & \text{for } \eta = 1. \end{cases} \tag{2.13}$$

Proof. We consider the cases for $\eta = 1$ and $\eta > 1$ separately. We have

I. For $\eta = 1$ it is $w = b^{(k)}$. Using (2.10) and applying theorem 1.4, which gives the transition probabilities of first order of the homogeneous Markov process $\{(Z_\gamma, U_\gamma);\ \gamma \in \mathbb{N}\}$ with state space Θ, we have

$$\begin{aligned} f_{(a,u)}(w;\gamma) &= \sum_{s\geq 0} \hat{p}^{\gamma-k+1}((\alpha,u),(b,s)) \frac{1-H(b,s+1)}{1-H(b,s)} \cdots \frac{1-H(b,s+k-1)}{1-H(b,s+k-2)} \\ &= \sum_{s\geq 0} \hat{p}^{\gamma-k+1}((\alpha,u),(b,s)) \frac{1-H(b,s+k-1)}{1-H(b,s)}. \end{aligned} \tag{2.14}$$

The transition probabilities of γ order are given by theorem 1.5. Therefore,

$$\begin{aligned} f_{(\alpha,u)}(w;\gamma) &= \sum_{s\geq 0} \Big\{ [I_{\{\alpha=b\}} I_{\gamma-k+1=s-u\}} \frac{1-H(b,s)}{1-H(\alpha,u)} \\ &+ I_{\{\gamma-k\geq s\}} \frac{\sum_{c\in\Omega} \sum_{n=1}^{\gamma-k+1-s} q(\alpha,c,n+u)\psi(c,b,\gamma-k+1-s-n)[1-H(b,s)]}{1-H(\alpha,u)}] \frac{1-H(b,s+k-1)}{1-H(b,s)} \Big\} \\ &= I_{\{\alpha=b\}} \frac{1-H(b,u+\gamma)}{1-H(\alpha,u)} + \\ &+ \frac{\sum_{c\in\Omega} \sum_{s=0}^{\gamma-k} \sum_{n=1}^{\gamma-k+1-s} q(\alpha,c,n+u)\psi(c,b,\gamma-k+1-s-n)[1-H(b,s+k-1)]}{1-H(\alpha,u)}. \end{aligned}$$

II. For $\eta > 1$ using (2.11) and applying theorem 1.4 we obtain the following.

$$\begin{aligned} f_{(\alpha,u)}(w;\gamma) = \sum_{s\geq 0} \hat{p}^{\gamma-k+1}((\alpha,u),(b_1,s)) & \frac{1-H(b_1,s+1)}{1-H(b_1,s)} \cdots \frac{1-H(b_1,s+k_1-1)}{1-H(b_1,s+k_1-2)} \\ & \frac{q(b_1,b_2,s+k_1)}{1-H(b_1,s+k_1-1)} \frac{1-H(b_2,1)}{1-H(b_2,0)} \cdots \frac{1-H(b_2,k_2-1)}{1-H(b_2,k_2-2)} \\ & \cdots \\ & \frac{q(b_{\eta-1},b_\eta,k_{\eta-1})}{1-H(b_{\eta-1},k_{\eta-1}-1)} \frac{1-H(b_\eta,1)}{1-H(b_\eta,0)} \cdots \frac{1-H(b_\eta,k_\eta-1)}{1-H(b_\eta,k_\eta-2)}. \end{aligned}$$

Therefore,

$$f_{(\alpha,u)}(w;\gamma) = \sum_{s\geq 0} \frac{\hat{p}^{\gamma-k+1}((\alpha,u),(b_1,s)) q(b_1,b_2,s+k_1)\sigma(w)}{1-H(b_1,s)}, \tag{2.15}$$

where $\sigma(w)$ is given by (2.13). Applying theorem 1.5 we get

$$\hat{p}^{\gamma-k+1}((\alpha,u),(b_1,s)) = I_{\{\alpha=b_1\}} I_{\{\gamma-k+1=s-u\}} \frac{1-H(b_1,s)}{1-H(\alpha,u)} + I_{\{0\le s\le \gamma-k\}} \frac{\sum_{c\in\Omega}\sum_{n=1}^{\gamma-k+1-s} q(\alpha,c,n+u)\psi(c,b_1,\gamma-k+1-s-n)[1-H(b_1,s)]}{1-H(\alpha,u)}.$$

Therefore,

$$\begin{aligned} f_{(\alpha,u)}(w;\gamma) &= I_{\{\alpha=b_1\}} \frac{1-H(\alpha,u-\gamma+k+1)}{1-H(\alpha,u)} \frac{q(b_1,b_2,\gamma-k+1+k_1+u)\sigma(w)}{1-H(b_1,u-\gamma+k+1)} + \\ &+ \sum_{s=0}^{\gamma-k} \frac{\sum_{c\in\Omega}\sum_{n=1}^{\gamma-k+1-s} q(\alpha,c,n+u)\psi(c,b_1,\gamma-k+1-s-n)[1-H(b_1,s)]}{1-H(\alpha,u)} \\ &\qquad \frac{q(b_1,b_2,s+k_1)\sigma(w)}{1-H(b_1,s)} \\ &= \{I_{\{\alpha=b_1\}} q(b_1,b_2,\gamma-k+1+k_1+u) + \\ &\sum_{s=0}^{\gamma-k}\sum_{c\in\Omega}\sum_{n=1}^{\gamma-k_j+1-s} q(\alpha,c,n+u)\psi(c,b_1,\gamma-k+1-s-n)q(b_1,b_2,s+k_1)\}\frac{\sigma(w)}{1-H(\alpha,u)} \end{aligned}$$

and the proof is complete. □

Theorem 2.5. *Let the DTMRP $\{(J_n,S_n);\ n\in\mathbb{N}\}$ with finite state space Ω be irreducible, aperiodic and $m_\alpha<\infty$ for all $\alpha\in\Omega$. We assume that the homogeneous Markov chain $\{(Z_\gamma,\ U_\gamma),\ \gamma\in\mathbb{N}\}$ is stationary, that is $\mathbb{P}(Z_0=\alpha,\ U_0=u)=\pi(\alpha,u)$ for all $(\alpha,u)\ \in\Theta$. Let $w\in\Omega^+_{(\Theta)}$ then for all $\gamma\ge k-1$, it is*

$$\mathbb{P}(\mathcal{E}_{w;\gamma}) = \pi(w) = \begin{cases} \dfrac{\sum_{s\ge k-1}[1-H(b,s)]}{\mu_b}\sigma(w), & \text{for } \eta=1 \\ \dfrac{\sum_{s\ge k_1} q(b_1,b_2,s)}{\mu_{b_1}}\sigma(w), & \text{for } \eta>1. \end{cases} \tag{2.16}$$

Proof. I. Let w be a run that is $w=b^{(k)}$. Following analogous steps as for the proof of (2.14) we get that

$$\mathbb{P}(\mathcal{E}_{w;\gamma}) = \sum_{s\ge 0}\mathbb{P}(Z_{\gamma-k+1}=b,\ U_{\gamma-k+1}=s)\frac{1-H(b,s+k-1)}{1-H(b,s)}$$

Since the process $\{(Z_\gamma,\ U_\gamma),\ \gamma\in\mathbb{N}\}$ is stationary it is $\mathbb{P}(Z_{\gamma-k+1}=b,\ U_{\gamma-k+1}=s)=\pi(b,s)$. Using theorem 1.6 we obtain

$$\mathbb{P}(\mathcal{E}_{w_j;\gamma}) = \sum_{s\ge 0}\frac{1-H(b,s)}{\mu_b}\frac{1-H(b,s+k-1)}{1-H(b,s)} = \sum_{s\ge k-1}\frac{1-H(b,s)}{\mu_b}.$$

II. If w is not a run , that is $w = w(1)w(2)\dots w(\eta)$, $\eta > 1$, following analogous steps as for the proof of (2.15) we get

$$\mathbb{P}(\mathcal{E}_{w;\gamma}) = \sum_{s\geq 0} \mathbb{P}(Z_{\gamma-k+1} = b_1,\ U_{\gamma-k+1} = s)\frac{q(b_1, b_2, s + k_1)\sigma(w)}{1 - H(b_1, s)}$$

Since the process $\{(Z_\gamma,\ U_\gamma),\ \gamma \in \mathbb{N}\}$ is stationary it is $\mathbb{P}(Z_{\gamma-k+1} = b_1,\ U_{\gamma-k+1} = s) = \pi(b_1, s)$. Using theorem 1.6 we obtain

$$\mathbb{P}(\mathcal{E}_{w;\gamma}) = \sum_{s\geq 0} \frac{1 - H(b_1, s)}{\mu_{b_1}} \frac{q(b_1, b_2, s + k_1)\sigma(w)}{1 - H(b_1, s)} = \sum_{s\geq k_1} \frac{q(b_1, b_2, s)}{\mu_{b_1}}\sigma(w).$$

□

Remark 2.6. In order to clarify the above results we notice the following.

Run: When a run $w = b^{(k)}$ occurs, the term $[1 - H(b, k - 1)]$ concerns the fact that a run of b's of length k occurs and we don't know what is the letter which appears after the last letter b of the word w. Moreover, the summation $\sum_{s\geq k-1}$ is necessary because we do not know the letter which has appeared before the first letter b of the run. That is, because we do not know if this first letter b is a renewal letter. The following figure clarifies our remark for runs.

$$\dots \quad \underset{\substack{\uparrow\\ \text{unknown}}}{.} \quad \underbrace{b\dots b}_{k} \quad \underset{\substack{\uparrow\\ \text{unknown}}}{.} \quad \dots$$

Not run: When a word $w = w(1)w(2)\dots w(\eta), \eta > 1$, which is not a run, occurs the situation is analogous. The term $[1 - H(b_\eta, k_\eta - 1)]$ which is included in the quantity $\sigma(w)$ is about not knowing what is the letter which appears after the last letter b_η of the word w (which is also the last letter of the run word w_η). However we know that the first letter of the run word $w(\eta)$ is a renewal. The summation $\sum_{s\geq k_1}$ concerns the fact that we do not know what is the letter which appears before the first letter of the word w (which is also the first letter of the run word w_1). On the other hand, we know what is the letter which comes after the last letter b_1 of the run word $w(1)$. For the runs $w(2)$, ..., $w(\eta)$ we know that their first letters are renewals. Thus the terms $q(b_2, b_3, k_2) \cdot \dots \cdot q(b_{\eta-1}, b_\eta, k_{\eta-1})$ appear in the quantity $\sigma(w)$. The following figure clarifies our remark for not runs.

$$\dots \quad \underset{\substack{\uparrow\\ \text{unknown}}}{.} \quad \underbrace{b_1\dots b_1}_{k_1} \quad \underbrace{b_2\dots b_2}_{k_2} \quad \dots \quad \underbrace{b_\eta\dots b_\eta}_{k_\eta} \quad \underset{\substack{\uparrow\\ \text{unknown}}}{.} \quad \dots$$

Example 2.7. The Markov case. We notice that if $\{Z_\gamma;\ \gamma \in \mathbb{N}\}$ is a Markov chain with transition probabilities $p^*(\alpha, \alpha')$, $\alpha, \alpha' \in \Omega$, then it is a particular case of a DTMRP with semi-Markov kernel

$$q(\alpha, \alpha', \gamma) = \begin{cases} (p^*(\alpha, \alpha))^{\gamma-1}p^*(\alpha, \alpha'), & \text{if } \alpha \neq \alpha' \text{ and } \gamma \in \mathbb{N},\ \gamma \geq 1 \\ 0, & \text{elsewhere.} \end{cases} \tag{2.17}$$

Let us consider the case $p^*(\alpha,\alpha')>0$, for all $\alpha,\alpha'\in\Omega$. Using (1.5) we have

$$\begin{aligned} H(\alpha,\gamma)&=\sum_{n=0}^{\gamma}\sum_{\alpha'\in\Omega}q(\alpha,\alpha',n)=\sum_{n=1}^{\gamma}\sum_{\substack{\alpha'\in\Omega\\ \alpha'\neq\alpha}}(p^*(\alpha,\alpha))^{n-1}p^*(\alpha,\alpha')=\sum_{n=1}^{\gamma}(p^*(\alpha,\alpha))^{n-1}[1-p^*(\alpha,\alpha)]\\ &=\sum_{n=0}^{\gamma-1}(p^*(\alpha,\alpha))^{n}[1-p^*(\alpha,\alpha)]=\frac{1-(p^*(\alpha,\alpha))^{\gamma}}{1-p^*(\alpha,\alpha)}[1-p^*(\alpha,\alpha)]=1-(p^*(\alpha,\alpha))^{\gamma}. \end{aligned} \tag{2.18}$$

It is $H(\alpha,\gamma)<1$ for all $\alpha\in\Omega$ and all $\gamma\in\mathbb{N}$.

$$m_\alpha=\sum_{\gamma\geq 0}[1-H(\alpha,\gamma)]=\sum_{\gamma\geq 0}(p^*(\alpha,\alpha))^{\gamma}=\frac{1}{1-p^*(\alpha,\alpha)}. \tag{2.19}$$

It is $m_\alpha<1$ for all $\alpha\in\Omega$.

We observe that $q(\alpha,\alpha',\gamma)>0$, for all $\alpha\neq\alpha'$, $\gamma>0$, which implies that the embedded Markov chain $\{J_n;\ n\in\mathbb{N}\}$ is irreducible and aperiodic.

We notice that under the conditions of theorem 2.5 we have

$$\lim_{\gamma\to\infty}\mathbb{P}_i(Z_\gamma=\alpha)=\frac{m_\alpha}{\mu_\alpha} \tag{2.20}$$

(see [1]).

For $p^*(\alpha,\alpha')>0$ the Markov chain $\{Z_\gamma;\ \gamma\in\mathbb{N}\}$ is irreducible and aperiodic with stationary distribution $\pi^*(\alpha)=\dfrac{m_\alpha}{\mu_\alpha}$, $\alpha\in\Omega$.

Using (2.19) we have

$$\mu_\alpha=\frac{1}{(1-p^*(\alpha,\alpha))\pi^*(\alpha)},\ \alpha\in\Omega. \tag{2.21}$$

Now, theorem 2.5 is written as follows.

For $w=b^{(k)}, w\in\Omega^+$ with $\eta=1$ it is

$$\pi(w)=\frac{\displaystyle\sum_{s\geq k-1}[1-H(b,s)]}{\mu_b}=\pi^*(b)[1-p^*(b,b)]\sum_{s\geq k_j-1}(p^*(b,b))^s=\pi^*(b)(p^*(b,b))^{k-1}. \tag{2.22}$$

For $w=w=b_1{}^{(k_1)}\dots b_\eta{}^{(k_\eta)},: w\in\Omega^+$ with $\eta>1$ it is

$$\begin{aligned} \pi(w)&=\frac{\displaystyle\sum_{s\geq k_1}q(b_1,b_2,s)\sigma(w)}{\mu_{b_1}}=\pi^*(b_1)[1-p^*(b_1,b_1)]\sum_{s\geq k_1}(p^*(b_1,b_1))^{s-1}p^*(b_1,b_2)\sigma(w)\\ &=\pi^*(b_1)(p^*(b_1,b_1))^{k_1-1}p^*(b_1,b_2)\sigma(w), \end{aligned} \tag{2.23}$$

where $\sigma(w)=(p^*(b_2,b_2))^{k_1-1}p^*(b_2,b_3)\ \dots\ (p^*(b_{\eta-1},b_{\eta-1}))^{k_{\eta-1}-1}p^*(b_{\eta-1},b_\eta)(p^*(b_\eta,b_\eta))^{k_\eta-1}$.

Relations (2.22) and (2.23) were expected.

Remark 2.8. We note that $\sigma(w)>0$ is a necessary condition for $\pi(w)$ to be positive. We consider the cases

I. For $w = b^{(k)}$ where $\eta = 1$ it is $\sigma(w) = 1 > 0$. Moreover the probability $\pi(w)$ is positive if $H(b, k-1) < 1$. In order to simplify our notation we set

$$S^1 := I_{\{H(b,k-1)<1\}} I_{\{\eta=1\}}.$$

II. For $w = b_1^{(k_1)} \dots b_\eta^{(k_\eta)}$ where $\eta > 1$, according to (2.13) , $\sigma(w)$ is positive if the following are valid

$$\begin{gathered} q(b_2, b_3, k_2) > 0, \ \dots, \ q(b_{\eta-1}, b_\eta, k_{\eta-1}) > 0, \ [1 - H(b_\eta, k_\eta - 1)] > 0, \quad \text{for } \eta > 2 \\ \text{and} \qquad 1 - H(b_2, k_2 - 1) > 0, \quad \text{for } \eta = 2. \end{gathered} \tag{2.24}$$

We notice that the condition $q(b_j, b_{j+1}, k_j) > 0$ implies the condition $H(b_j, k_j - 1) < 1$, which means $(b_j, k_j - 1) \in \Theta, \quad j = 2, \ \dots, \ \eta - 1$.

Moreover, in this case the probability $\pi(w)$ is positive if there exists s, with $s \geq k_1$, such that $q(b_1, b_2, s) > 0$. Analogously, we set

$$S^2 := I_{\{\exists \ s \ : \ s \geq k_1, \ q(b_1,b_2,s)>0\}} I_{\{\eta>1\}}.$$

Therefore in the sequel, we shall consider the set of words

$$W_{\Omega,q} := \{w \in \Omega^+ \ : \ \sigma(w) > 0, \ \text{ and } S^1 + S^2 = 1\}. \tag{2.25}$$

3. The number of overlapping occurrences

Let $w \in W_{\Omega,q}$ of length k. The number of overlapping occurrences of w in the sequence $Z_0 \ \dots \ Z_n$, $n \in \mathbb{N}$ is defined by

$$N(n) = \sum_{\gamma=k-1}^{n} E_{w;\gamma}, \quad n \geq k - 1. \tag{3.1}$$

Reinert et al. [5] have called $N(n)$ count of word w.

In this section we shall derive $\mathbb{E}(N(n))$ and $\mathrm{Var}\,(N(n))$ under certain conditions.

The computation of $\mathbb{E}(N(n))$ is an immediate consequence of theorem 2.5 and relation (3.1). Hence we get the following.

Theorem 3.1. *Let the DTMRP $\{(J_n, S_n); \ n \in \mathbb{N}\}$ with finite state space Ω be irreducible, aperiodic and $m_\alpha < \infty$ for all $\alpha \in \Omega$. We assume that the Markov chain $\{(Z_\gamma, \ U_\gamma), \ \gamma \in \mathbb{N}\}$ is stationary, that is $\mathbb{P}(Z_0 = \alpha, \ U_0 = u) = \pi(\alpha, u)$. It is*

$$\mathbb{E}(N(n)) = (n - k + 2)\pi(w), \quad \forall \ w \in W_{\Omega,q}, \quad n \in \mathbb{N}, \ n \geq k - 1. \tag{3.2}$$

In order to compute the variance of $N(n)$, $n \in \mathbb{N}$, $n \geq k - 1$ we need the probability $\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, \mathcal{E}_{w;r})$, $\quad r \in \mathbb{N}$, $\gamma \in \mathbb{N}$, $\gamma \geq 1$ which is given in the following lemma.

Lemma 3.2. *For every $w \in W_{\Omega,q}$ which is not a run and every $\gamma \in \mathbb{N}, \ \gamma \geq 1$ it is*

$$\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, \mathcal{E}_{w;r}) = \check{\psi}(\gamma), \qquad \forall \ r \in \mathbb{N}, \tag{3.3}$$

where

$$\check{\psi}(\gamma) = I_{\{\gamma<k\}} I_{\{\gamma \in \mathcal{P}(w)\}} f_{(b_\eta, k_\eta - 1)}([w]_\gamma; \gamma) + I_{\{\gamma \geq k\}} f_{(b_\eta, k_\eta - 1)}(w; \gamma). \tag{3.4}$$

Proof. Assume $r \in \mathbb{N}$ such that $\mathbb{P}(\mathcal{E}_{w;r}) > 0$. Using remark 2.1 for $\eta > 1$ we have

$$\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, \mathcal{E}_{w;r}) = \mathbb{P}([\cup_{s\geq 0}\mathcal{E}_{w(1);r+\gamma-k+k_1}(s)] \cap (\mathcal{E}_{w(2);r+\gamma-k+k_1+k_2}(0) \cap \ \ldots \cap \ \mathcal{E}_{w(\eta);r+\gamma}(0)) \\ |\, [\cup_{s\geq 0}\mathcal{E}_{w(1);r-k+k_1}(s)] \cap (\mathcal{E}_{w(2);r-k+k_1+k_2}(0) \cap \ \ldots \cap \ \mathcal{E}_{w(\eta);r}(0)).$$

The term $\mathcal{E}_{w(\eta);r}(0)$ gives us the state of (Z, U) at time r. Thus since the $\{(Z_\gamma,\ U_\gamma);\ \gamma \in \mathbb{N}\}$ is an homogeneous Markov chain we get

- For $\gamma < k$ the two occurrences happen when $\gamma \in \mathcal{P}(w)$, that is, there is an overlap of length $k - \gamma$. Therefore it is

$$\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, \mathcal{E}_{w;r}) = I_{\{\gamma\in\mathcal{P}(w)\}}\mathbb{P}(\mathcal{E}_{[w]_\gamma;r+\gamma} \,|\, Z_r = b_\eta,\ U_r = k_\eta - 1).$$

- For $\gamma \geq k$ there is no overlap. Thus we get

$$\mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, \mathcal{E}_{w;r}) = \mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, Z_r = b_\eta,\ U_r = k_\eta - 1).$$

Using (2.12) the proof is complete. □

Remark 3.3. Considering remark 2.1 we note that if w was a run we would'nt know the states of (Z, U) at time $r,\ r-1,\ \ldots,\ r-k+1$. Thus an analogous lemma in this case could not be obtained.

We give now a theorem for the variance of the count of a word under certain conditions.

Theorem 3.4. *Let the DTMRP $\{(J_n, S_n);\ n \in \mathbb{N}\}$ with finite state space Ω be irreducible, aperiodic and $m_\alpha < \infty$ for all $\alpha \in \Omega$. We assume that the Markov chain $\{(Z_\gamma,\ U_\gamma),\ \gamma \in \mathbb{N}\}$ is stationary, that is $\mathbb{P}(Z_0 = \alpha,\ U_0 = u) = \pi(\alpha, u)$, for all $(\alpha, u) \in \Theta$.*

For all $w \in W_{\Omega,q}$, which is not a run and $n \in \mathbb{N}$, $n \geq k-1$ it is

$$\begin{aligned}\mathrm{Var}\,(N(n)) &= (n-k+2)\pi(w)\big[1-(n-k+2)\pi(w)\big] \\ &\quad + 2\sum_{\gamma\in\mathcal{P}(w)}(n-\gamma-k+2)f_{(b_\eta,k_\eta-1)}([w]_\gamma;\gamma)\pi(w) \\ &\quad + 2\sum_{\gamma=k}^{n-k+1}(n-\gamma-k+2)f_{(b_\eta,k_\eta-1)}(w;\gamma)\pi(w).\end{aligned} \tag{3.5}$$

Proof. Let $n\ \in \mathbb{N}$, $n \geq k-1$. It is $\mathrm{Var}\,(N(n)) = \mathrm{Var}\,(\sum_{\gamma=k-1}^{n} E_{w;\gamma})$. Thus

$$\begin{aligned}\mathrm{Var}\,(N(n)) &= \sum_{\gamma=k-1}^{n}\mathrm{Var}\,(E_{w;\gamma}) + 2\sum_{n_1=k-1}^{n}\sum_{n_2=n_1+1}^{n}\mathrm{cov}\,(E_{w;n_1},\ E_{w;n_2}) \\ &= \sum_{\gamma=k-1}^{n}\mathrm{Var}\,(E_{w;\gamma}) + 2\sum_{\gamma=1}^{n-k+1}\sum_{r=k-1}^{n-\gamma}\mathrm{cov}\,(E_{w;r},\ E_{w;r+\gamma}).\end{aligned}$$

It is $\mathrm{Var}\,(E_{w;\gamma}) = \mathbb{P}(\mathcal{E}_{w;\gamma}) - [\mathbb{P}(\mathcal{E}_{w;\gamma})]^2$ and $\mathrm{cov}\,(E_{w;r},\ E_{w;r+\gamma}) = \mathbb{P}(\mathcal{E}_{w;r}\ \text{ and }\ \mathcal{E}_{w;r+\gamma}) - \mathbb{P}(\mathcal{E}_{w;r})\mathbb{P}(\mathcal{E}_{w;r+\gamma})$.

Since the Markov chain $\{(Z_\gamma,\ U_\gamma),\ \gamma \in \mathbb{N}\}$ is stationary, applying theorem 2.5, it is $\mathrm{Var}\,(E_{w;\gamma}) = \pi(w) - [\pi(w)]^2$ and $\mathrm{cov}\,(E_{w;r},\ E_{w;r+\gamma}) = \mathbb{P}(\mathcal{E}_{w;r}\ \text{ and }\ \mathcal{E}_{w;r+\gamma}) - [\pi(w)]^2$. But

$\mathbb{P}(\mathcal{E}_{w;r}\ \text{ and }\ \mathcal{E}_{w;r+\gamma}) = \mathbb{P}(\mathcal{E}_{w;r+\gamma} \,|\, \mathcal{E}_{w;r})\mathbb{P}(\mathcal{E}_{w;r}) = \check{\psi}(\gamma)\pi(w)$, where $\check{\psi}(\gamma)$ is given by lemma 3.2.

Therefore,

$$\operatorname{cov}\left(E_{w;r},\ E_{w;r+\gamma}\right)=\begin{cases} f_{(b_\eta,k_\eta-1)}([w]_\gamma;\gamma)\pi(w)-[\pi(w)]^2, & \text{for } \gamma<k \text{ and } \gamma\in\mathcal{P}(w),\\ -[\pi(w)]^2, & \text{for } \gamma<k \text{ and } \gamma\notin\mathcal{P}(w),\\ f_{(b_\eta,k_\eta-1)}(w;\gamma)\pi(w)-[\pi(w)]^2, & \text{for } \gamma\geq k. \end{cases} \tag{3.6}$$

Thus,

$$\begin{aligned} 2\sum_{\gamma=1}^{n-k+1}\sum_{r=k-1}^{n-\gamma}\operatorname{cov}\left(E_{w;r},\ E_{w;r+\gamma}\right) &= 2\sum_{\gamma\in\mathcal{P}(w)}(n-\gamma-k+2)\{f_{(b_\eta,k_\eta-1)}([w]_\gamma;\gamma)\pi(w)-[\pi(w)]^2\}\\ &\quad+2\sum_{\substack{\gamma<k\\ \gamma\notin\mathcal{P}(w)}}(n-\gamma-k+2)\{-[\pi(w)]^2\}\\ &\quad+2\sum_{\gamma=k}^{n-k+1}(n-\gamma-k+2)\{f_{(b_\eta,k_\eta-1)}(w;\gamma)\pi(w)-[\pi(w)]^2\}. \end{aligned}$$

Therefore

$$\begin{aligned} \operatorname{Var}\left(N(n)\right) &= \sum_{\gamma=k-1}^{n}\left\{\pi(w)-[\pi(w)]^2\right\}+2\sum_{\gamma\in\mathcal{P}(w)}(n-\gamma-k+2)f_{(b_\eta,k_\eta-1)}([w]_\gamma;\gamma)\pi(w)\\ &\quad+2\sum_{\gamma=k}^{n-k+1}(n-\gamma-k+2)f_{(b_\eta,k_\eta-1)}(w;\gamma)\pi(w)-2\sum_{\gamma=1}^{n-k+1}(n-\gamma-k+2)[\pi(w)]^2\\ &= (n-k+2)\pi(w)\left[1-(n-k+2)\pi(w)\right]+2\sum_{\gamma\in\mathcal{P}(w)}(n-\gamma-k+2)f_{(b_\eta,k_\eta-1)}([w]_\gamma;\gamma)\pi(w)\\ &\quad+2\sum_{\gamma=k}^{n-k+1}(n-\gamma-k+2)f_{(b_\eta,k_\eta-1)}(w;\gamma)\pi(w). \end{aligned}$$

□

A note. As we mentioned above under the Markovian hypothesis the distribution of the sojourn time in a state is geometrically distributed. In the semi-Markov case we allow arbitrarily distributed sojourn times in any state. Under this general assumption our first aim was to derive the occurrence probability of a word as well as the mean and variance of the word count. According to our opinion, we believe that these results will be useful in generalizing further results which have been derived under the Markovian hypothesis.

I would like to thank Professor Ourania Chryssaphinou for many helpful comments and stimulating discussions.

I would also like to thank the Associate Editor and the reviewers for their useful comments and suggestions which have improved this paper.

References

[1] V. Barbu, M. Boussemart and N. Limnios, Discrete time semi-Markov processes for reliability and survival analysis. *Communications in Statistics-Theory and Methods* **33**(11) (2004) 2833-2868.

[2] O. Chryssaphinou, M. Karaliopoulou and N. Limnios, On Discrete Time semi-Markov chains and applications in words occurrences. *Communications in Statistics-Theory and Methods* **37**(8) (2008) 1306-1322.
[3] L.J. Guibas and A.M. Odlyzko, String Overlaps, pattern matching and nontransitive games. *Journal of combinatorial Theory* Series A, **30** (1981) 183-208.
[4] M. Lothaire, *Combinatorics on Words.* Addison-Wesley (1983).
[5] G. Reinert, S. Schbath and M.S. Waterman, Probabilistic and Statistical Properties of Finite Words in Finite Sequences. In: *Lothaire: Applied Combinatorics on Words.* Cambridge University Press, J. Berstel and D. Perrin, eds. (2005).
[6] V.T. Stefanov, On Some Waiting Time Problems. *J. Appl. Prob.* **37**(3) (2000) 756-764.
[7] V.T. Stefanov, The intersite distances between pattern occurrences in strings generated by general discrete-and continuous-time models: an algorithmic approach. *J. Appl. Prob.* **40**(4) (2003) 881-892.