Average Value and Variance of Pattern Statistics in Rational Models [⋆] [⋆⋆]

Massimiliano Goldwurm and Roberto Radicioni

Università degli Studi di Milano
Dipartimento di Scienze dell'Informazione
Via Comelico 39, 20135 Milano, Italy
goldwurm,radicioni@dsi.unimi.it

Abstract. We study the pattern statistics representing the number of occurrences of a given string in a word of length n generated at random by rational stochastic models, defined by means of weighted finite automata. We get asymptotic estimations for the mean value and the variance of these statistics under the hypothesis that the matrix of all transition weights is primitive. Our results extend previous evaluations obtained by assuming ergodic stationary Markovian sources and they yield a general framework to determine analogous estimations under several stochastic models. In particular they show the role of the stationarity hypothesis in such models.

1 Introduction

The classical problem of evaluating the number of occurrences of a given string (usually called pattern) in a random text has been mainly studied assuming the text generated by a Markovian source [12,9,10]. Here we assume more general stochastic models, called rational, which were first considered in [2] and studied in details in [6]. The rational models are defined by means of weighted finite automata and are able to generate a random string of given length in a regular language under uniform distribution. In this work we determine asymptotic expressions of average value and variance of the number of occurrences of a pattern in a string of length n generated at random in such models. We compare our results with analogous evaluations obtained in [12,9,2]. Our approach yields a general framework where the previous evaluations appear as special cases. We also relax the stationarity hypothesis assumed in [12,9] and show how such a condition affects the evaluations of mean value and variance of our statistics.

2 Preliminary Notions

Given a set X and an integer $m > 0$, we denote by X^m and $X^{m \times m}$, respectively, the set of all vectors and the set of all square matrices of size m with coefficients in X.

[⋆] Appeared in revised form in *Proc. CIAA 2007*, 12th Conference on Implementation and Application of Automata, Lecture Notes in Computer Science n. 4783, 62-72, Springer-Verlag, 2007.

[⋆⋆] This work has been supported by Project MIUR PRIN 2005-2007 "Automata and Formal Languages: mathematical and applicative aspects".

Any $x \in X^m$ is considered as a column vector, while x' is its transposed (row) vector. Denoting by $\mathbb{R}_+$ the set of nonnegative real numbers, we recall that a matrix $M \in \mathbb{R}_+^{m \times m}$ is called primitive if $M^n > 0$ for some integer $n > 0$, meaning that all entries of M^n are greater than 0. By the Perron–Frobenius theorem [13, Sect.1], it is well-known that every primitive matrix $M \in \mathbb{R}_+^{m \times m}$ admits a real positive eigenvalue λ, called the Perron–Frobenius eigenvalue of M, which is a simple root of the characteristic polynomial of M, such that $|\nu| < \lambda$ for every eigenvalue $\nu \neq \lambda$.

The properties of nonnegative matrices are widely used to study the behaviour of Markov chains [5,8,13]. We recall that a real vector $\pi' = (\pi_1, \pi_2, \ldots, \pi_m)$ is stochastic if $0 \leq \pi_i \leq 1$ for every i and $\sum_{i=1}^{n} \pi_i = 1$. A matrix $P \in \mathbb{R}^{m \times m}$ is stochastic if all its rows are stochastic vectors. It is easy to see that any stochastic matrix P has eigenvalue 1, with a corresponding right eigenvector $\underline{1}' = (1, 1, \ldots, 1)$, while $|\gamma| \leq 1$ for any other eigenvalue γ of P.

A stochastic vector π and a stochastic matrix P of same size m allows us to define a Markov chain over the set of states $Q = \{1, 2, \ldots, m\}$, i.e. a sequence of random variables $\{X_n\}_{n \in N}$ taking on values in Q, such that $\Pr(X_0 = i) = \pi_i$ and

$$\Pr(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \Pr(X_{n+1} = j \mid X_n = i) = P_{ij}$$

for every $n \in \mathbb{N}$, and any tuple of states $j, i, i_0 \ldots, i_{n-1} \in Q$. The arrays π and P are called, respectively, the initial vector and the transition matrix of the Markov chain. Note that $\Pr(X_n = j) = (\pi' P^n)_j$, for each $j \in Q$ and every $n \in \mathbb{N}$. Moreover, if P is primitive, by the Perron–Frobenius theorem one can prove that

$$P^n = \underline{1} v' + \mathrm{O}(\varepsilon^n) , \tag{1}$$

where $0 \leq \varepsilon < 1$ and v' is the left eigenvector of P corresponding to the eigenvalue 1 such that $v' \underline{1} = 1$. Observe that $\underline{1} v'$ is a stable matrix, i.e. all its rows equal v'; moreover, v' is a stochastic vector, called the stationary vector of the chain, and it is the unique stochastic vector such that $v' P = v'$. If further $\pi = v$ then the Markov chain is called stationary, since $\pi' P^n = \pi'$ for every $n \in \mathbb{N}$, and hence $\Pr(X_n = j) = \pi_j$ for any state j.

Now, let us fix our notation on words and formal series. Given a finite alphabet A, for every $x \in A^*$, $|x|$ is the length of x and $|x|_a$ is the number of occurrences of a symbol $a \in A$ in x. We also denote by A^n the set $\{x \in A^* \mid |x| = n\}$ for every $n \in \mathbb{N}$. A formal series over A with coefficients in $\mathbb{R}_+$ is a function $r : A^* \longrightarrow \mathbb{R}_+$, usually represented in the form $r = \sum_{x \in A^*} r(x) \cdot x$, where $r(x)$ denotes the value of r at $x \in A^*$. We denote by $\mathbb{R}_+\langle\langle A \rangle\rangle$ the family of all formal series over A with coefficients in $\mathbb{R}_+$. This set forms a semiring with respect to the traditional operations of sum and Cauchy product. As an example of formal series in $\mathbb{R}_+\langle\langle A \rangle\rangle$, we recall the notion of *probability measure* on A^*, defined as a map $f : A^* \longrightarrow [0, 1]$, such that $f(\epsilon) = 1$ and $\sum_{a \in A} f(xa) = f(x)$, for every $x \in A^*$ [7].

A formal series $r \in \mathbb{R}_+\langle\langle A \rangle\rangle$ is called rational if it admits a linear representation, that is a triple $\langle \xi, \mu, \eta \rangle$ where, for some integer $m > 0$, ξ and η are (column) vectors in $\mathbb{R}_+^m$ and $\mu : A^* \longrightarrow \mathbb{R}_+^{m \times m}$ is a monoid morphism, such that $r(x) = \xi' \mu(x) \eta$ holds for each $x \in A^*$. We say that m is the *size* of the representation. Such a triple $\langle \xi, \mu, \eta \rangle$ can be interpreted as a weighted nondeterministic automaton, where the set of states is given by $\{1, 2, \ldots, m\}$ and the transitions, the initial and the final states are assigned weights in $\mathbb{R}_+$ by μ, ξ and η, respectively. To avoid redundancy it is convenient to assume that $\langle \xi, \mu, \eta \rangle$ is trim, i.e. for every index i there are two indexes p, q and two words $x, y \in A^*$ such that $\xi_p \mu(x)_{pi} \neq 0$ and $\mu(y)_{iq} \eta_q \neq 0$. The total transition matrix M of $\langle \xi, \mu, \eta \rangle$ is defined by $M = \sum_{a \in A} \mu(a)$. We say that $\langle \xi, \mu, \eta \rangle$ is primitive if such a matrix M is primitive.

Several properties of the formal series in $\mathbb{R}_+\langle\langle A \rangle\rangle$ can be studied by considering their commutative image. To define it formally, consider the canonical morphism $\Phi : A^* \longrightarrow \mathcal{M}(\mathcal{A})$, where $\mathcal{M}(\mathcal{A})$ is the free totally commutative monoid over A. Such a monoid morphism extends to a semiring morphism from $\mathbb{R}_+\langle\langle A \rangle\rangle$ to the traditional ring $\mathbb{R}[[A]]$ of formal series with real coefficients and commutative variables in A. We recall that, if $r \in \mathbb{R}_+\langle\langle A \rangle\rangle$ is rational, then also $\Phi(r)$ is rational in $\mathbb{R}[\![A]\!]$, i.e. $\Phi(r) = pq^{-1}$ for two polynomials $p, q \in \mathbb{R}[A]$.

3 Stochastic Models on Words

Several stochastic models have been proposed in the literature to study probability measures on free monoids [11,7]. Here, we intuitively consider a stochastic (probabilistic) model over a finite alphabet A as a device to define a probability function on the set A^n for every integer $n > 0$, equipped with an effective procedure to generate on input n a word in A^n with the prescribed probability.

In this section we discuss three types of probabilistic models introduced in [6] and called, respectively, Markovian, sequential and rational models. Here, we recall their main properties and differences. These models include the classical Markovian sequences of any order and the rational probability measure studied in [7]. They can be seen as special cases of more general probabilistic devices studied in [11].

The simplest probabilistic model on words is the well-known Bernoullian model. A *Bernoullian* model $\mathcal{B}$ over A is defined by a function $p : A \to [0, 1]$ such that $\sum_{a \in A} p(a) = 1$. A word $x \in A^+$ is generated in this model by choosing each letter of x under the distribution defined by p independently of one another. Thus, the probability of $x = x_1 x_2 \cdots x_n$, where $x_i \in A$ for each i, is given by $\mathrm{Pr}_{\mathcal{B}}(x) = p(x_1)p(x_2)\cdots p(x_n)$, which clearly defines a probability function over A^n for every integer $n > 0$.

3.1 Markovian Models

A *Markovian* model over A is defined as a pair $\mathcal{M} = (\pi, \mu)$ where, for some integer $k > 0$, $\pi \in [0, 1]^k$ is a stochastic vector and μ is a function $\mu : A \to [0, 1]^{k \times k}$ such that, for every $a \in A$, each row of $M(a)$ has at most one non-null entry and the matrix $M = \sum_{a \in A} \mu(a)$ is stochastic.

The probability of a word $x = x_1 x_2 \cdots x_n$, where $x_i \in A$ for each $i = 1, 2, \ldots, n$, is given by

$$\mathrm{Pr}_{\mathcal{M}}(x) = \pi' \mu(x_1) \mu(x_2) \cdots \mu(x_n) \underline{1}.$$

Thus, $\mathrm{Pr}_{\mathcal{M}}$ is a rational formal series in $\mathbb{R}_+\langle\langle A \rangle\rangle$ with linear representation $\langle \pi, \mu, \underline{1} \rangle$. Also, since both π and M are stochastic arrays, $\mathrm{Pr}_{\mathcal{M}}$ defines a probability function over

A^n for each positive integer n. This model implicitly defines a Markov chain taking on values in the set of states $\{1, 2, \ldots, k\}$, that has initial vector π and transition matrix M; we may call it the *underlying* Markov chain of $\mathcal{M}$.

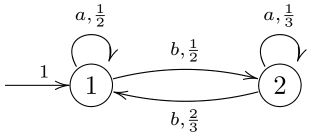
Note that every Bernoullian model is a Markovian model. Moreover, the pair $\mathcal{M} = (\pi, \mu)$ defines a deterministic finite state automaton where transitions are weighted by probabilities: the set of states is $Q = \{1, 2, \ldots, k\}$, the transition function $\delta_{\mathcal{M}} : Q \times A \longrightarrow Q \cup \{\bot\}$ is defined so that for every $i \in Q$ and every $a \in A$, $\delta_{\mathcal{M}}(i, a) = j$ if $\mu(a)_{ij} \neq 0$, and the same value $\mu(a)_{ij}$ is the weight of the transition, while $\delta_{\mathcal{M}}(i, a) = \bot$ if $\mu(a)_{ij} = 0$. Clearly, $\delta_{\mathcal{M}}$ can be extended to all words in A^*. Thus, the sum of weights of all transitions outgoing from any state equals 1 and, since the automaton is deterministic, for every word $x \in A^*$ and every $i \in Q$ there exists at most one path labeled by x starting from i. These properties lead to prove the following lemma, which gives an asymptotic property of the probabilities defined in Markovian models.

Lemma 1. *[6] Let $\mathcal{M} = (\pi, \mu)$ be a Markovian model of size k over the alphabet A and let $x \in A^+$. Then, there exists $0 \leq \beta \leq 1$ such that $\mathrm{Pr}_{\mathcal{M}}(x^n) = \Theta(\beta^n)$, as n tends to $+\infty$* [1].

This lemma plays a role similar to classical pumping lemma in formal languages in the sense that it can be used to show that a given probabilistic model on A is not Markovian simply by showing that, for a word $x \in A^+$, the probability of x^n is not of the order $\Theta(\beta^n)$ for any constant $\beta \geq 0$.

Observe that the Markovian models can generate the traditional Markov sequences of order m over A (for any $m \in \mathbb{N}$), where the probability of the next symbol occurrence only depends on the previous m symbols. To define these sources in our context we say that a Markovian model $\mathcal{M}$ over A is of *order* m if for every word $w \in A^m$ either there exists j such that $\delta_{\mathcal{M}}(i, w) = j$ for every $i \in Q$ or $\delta_{\mathcal{M}}(i, w) = \bot$ for every $i \in Q$, and m is the smallest integer with such a property.

A relevant case occurs when $m = 1$. In this case, the set of states Q can be reduced to A and $\mathrm{Pr}_{\mathcal{M}}$ is called Markov probability measure in [7]. Also observe that there exist Markovian models that are not of order m, for any $m \in N$. For instance, if $\mathcal{M}$ is defined by the following (weighted) finite automaton, then $\delta_{\mathcal{M}}(1, a^n b) \neq \delta_{\mathcal{M}}(2, a^n b)$ for every $n \in \mathbb{N}$.



Hence our notion of Markovian model properly includes the traditional sources of Markovian sequences.

3.2 Sequential Models

A natural proper extension of the previous model can be obtained by allowing nondeterminism in the corresponding finite state device. In this way the model corresponds

[1] This means that for some positive constants c_1, c_2, the relation $c_1\beta^n \leq Pr_{\mathcal{M}}(x^n) \leq c_2\beta^n$ holds for any n large enough.

to a stochastic sequential machine, as defined in [11], with a unary input alphabet. Moreover, it is characterized by the rational probability measures, i.e. the probability measures on A^* that are rational formal series in $\mathbb{R}_+\langle\langle A\rangle\rangle$ [7].

Formally, we define a *sequential* stochastic model over A as a pair $\mathcal{Q} = (\pi, \mu)$ where $\pi \in [0,1]^k$ is a stochastic vector and μ is a function $\mu : A \to [0,1]^{k\times k}$ such that $M = \sum_{a\in A} \mu(a)$ is a stochastic matrix. Clearly, any Markovian model is also a sequential model. In particular, as in the Markovian models, μ defines a monoid morphism from A^* to $[0,1]^{k\times k}$, and the probability of a word $x = x_1x_2\cdots x_n \in A^*$ is given by

$$\mathrm{Pr}_{\mathcal{Q}}(x) = \pi'\mu(x)\underline{1} = \pi'\mu(x_1)\mu(x_2)\cdots\mu(x_n)\underline{1}$$

Also in this case, $\mathrm{Pr}_{\mathcal{Q}}$ is a rational formal series, taking on values in $[0,1]$, that admits the linear representation $\langle \pi, \mu, \underline{1}\rangle$ and defines a probability function over A^n, for every positive integer n. Furthermore, it is easy to see that $\mathrm{Pr}_{\mathcal{Q}}$ is a rational probability measure on A^*. Actually, that is a characterization of the sequential models, in the sense that, as proved in [7], for every rational probability measure f on A^* there exists a sequential model $\mathcal{Q}$ such that $f = \mathrm{Pr}_{\mathcal{Q}}$.

Analogously, one can define the underlying Markov chain on the set of states $Q = \{1, 2, \ldots, k\}$ with initial vector π and transition matrix M. We say that the sequential model is primitive if M is a primitive matrix; if further π is the stationary vector then the model is said to be stationary too. Moreover, the pair $\mathcal{Q} = (\pi, \mu)$ can be interpreted as a finite state automaton equipped with probabilities associated with transitions; the main difference, with respect to the Markovian models, is that now the automaton is nondeterministic. For any $a \in A$, every non-null entry $\mu(a)_{ij}$ is the weight of the transition from i to j labeled by a and, for every word x, $\mathrm{Pr}_{\mathcal{Q}}(x)$ is the sum of the weights of all paths labeled by x in the corresponding transition diagram.

However, in spite of these similarities, the sequential models seem to be much more general than Markovian models. In particular, their probability functions do not satisfy Lemma 1. In fact, it is easy to find a sequential model $\mathcal{Q}$ such that $\mathrm{Pr}_{\mathcal{Q}}(a^n) = \Theta(n\varepsilon^n)$ for some $a \in A$ and $0 < \varepsilon < 1$.

Further properties of sequential models concern the commutative image of the probability functions. Since $\mathrm{Pr}_{\mathcal{Q}}$ is rational, also its commutative image $\Phi(\mathrm{Pr}_{\mathcal{Q}})$ is rational in $\mathbb{R}[\![A]\!]$; such a series is given by $\Phi(\mathrm{Pr}_{\mathcal{Q}}) = \pi'(I - \sum_{a\in A}\mu(a)a)^{-1}\underline{1}$ and it represents the generating function of the probabilities of occurrences of symbols in A. In other words, setting $A = \{a_1, a_2, \ldots, a_s\}$, we have

$$\Phi(\mathrm{Pr}_{\mathcal{Q}}) \;=\; \sum_{i\in\mathbb{N}^s} \;\sum_{|x|_{a_1}=i_1,\ldots,|x|_{a_s}=i_s} \mathrm{Pr}_{\mathcal{Q}}(x)\, a_1^{i_1}\cdots a_s^{i_s} \;.$$

3.3 Rational Models

Consider a rational formal series $r \in \mathbb{R}_+\langle\langle A\rangle\rangle$ and, for every positive integer n, assume $r(w) \neq 0$ for some $w \in A^n$. Then r defines a probability function over A^n, given by

$$\mathrm{Pr}_r(x) = \frac{r(x)}{\sum_{w\in A^n} r(w)} \qquad \text{for every } x \in A^n \;. \tag{2}$$

Observe that if r is the characteristic series χ_L of a regular language $L \subseteq A^*$, then Pr_r represents the uniform probability function over $L \cap A^n$, for each n.

Since r is rational it admits a linear representation (ξ, μ, η) and hence

$$\mathrm{Pr}_r(x) = \frac{\xi' \mu(x) \eta}{\xi' M^n \eta} \qquad \text{for every } x \in A^n, \tag{3}$$

where $M = \sum_{a \in A} \mu(a)$. Thus any linear representation (ξ, μ, η) defines a rational model; we say that the model is primitive if M is a primitive matrix. Also observe that Pr_r is a sort of Hadamard division of two rational formal series. Well-known algorithms for the generation of random words in regular languages can be easily modified in order to compute, for an input $n \in \mathbb{N}$, a word x with probability $\mathrm{Pr}_r(x)$ [4,3].

It is clear that every sequential model over A is a rational model over the same alphabet. However, in this case $\Phi(\mathrm{Pr}_r)$ is not always a rational series. As shown in [6], such a function may even be non-holonomic (and hence transcendental). This occurs for rather simple r as, for instance, the characteristic series of the language $(b + ab)^*$. The key property here is that the Hadamard division of two rational series is not necessarily rational. This proves that rational models are a proper extension of the sequential models.

Thus, we can summarize the discussion presented in this section by the following statement.

Proposition 1. *The chain of inclusions*

Markovian models $\subset$ *Sequential models* $\subset$ *Rational models*

is strict. Moreover, the Markovian models strictly include the Bernoullian models and can generate the Markovian sequences of order m*, for every integer* $m \geq 1$.

4 Average Number of Pattern Occurrences

In this section we evaluate the average number of occurrences of a pattern $w \in A^*$ in a string $x \in A^*$ of length n, generated at random in a rational model, assuming that the corresponding series admits a primitive linear representation.

Let (ξ, μ, η) be a linear representation of size m over the alphabet A as defined in Section 2. We assume the matrix $M = \sum_{a \in A} \mu(a)$ is primitive. Let λ be its Perron-Frobenius eigenvalue and denote by v and u, respectively, the left and right (strictly positive) eigenvector of M corresponding to λ such that $v'u = 1$. We know (see for instance [13]) that for every $n \in \mathbb{N}$

$$M^n = \lambda^n (uv' + C(n)) \tag{4}$$

where $C(n)$ is a real matrix such that $C(n) = \mathrm{O}(\varepsilon^n)$, for some $0 < \varepsilon < 1$. Thus, we can define the matrix $C = \sum_{n=0}^{+\infty} C(n)$, which turns out to be an analogous of the fundamental matrix in Markov chains [8, Sect. 4.3]. In fact, the following identities are easy to prove:

$$C = \left(I - \left(\frac{M}{\lambda} - uv'\right)\right)^{-1} - uv' \ , \ \ v'C = Cu = 0 \ ,$$

$$CM = MC = \lambda(C - I + uv') \ , \ \sum_{n=0}^{+\infty} nC(n) = \frac{CMC}{\lambda} = C^2 - C \ .$$

Now, let $w \in A^*$ be a pattern of length $m \geq 1$ and let $x \in A^*$ be a word of length $n \geq m$, generated at random in the rational model defined by the linear representation (ξ, μ, η). We can consider the random variable O_n representing the number of occurrences of w in x, i.e. $O_n = |x|_w$. Clearly, O_n takes on values in $\{0, 1, \ldots, n - m + 1\}$ and for each i in that set we have

$$\Pr(O_n = i) \ = \sum_{y \in A^n, |y|_w = i} \frac{\xi'\mu(y)\eta}{\xi' M^n \eta}$$

Setting $x = x_1 \cdots x_n$ with $x_i \in A$ for each i, we can consider O_n as a sum of random variables of the form

$$O_n = I_m + I_{m+1} + \cdots + I_n \tag{5}$$

where, for every $j = m, m+1, \ldots, n$

$$I_j = \begin{cases} 1 \text{ if } x_1 x_2 \cdots x_j \in A^* w \\ 0 \text{ otherwise} \end{cases}$$

Note that each I_j is a Bernoullian random variable such that

$$\Pr(I_j = 1) = \frac{\xi' M^{j-m} \mu(w) M^{n-j} \eta}{\xi' M^n \eta}$$

Proposition 2. *Let O_n be the number of occurrence of a nonempty pattern $w \in A^m$ in a string of length $n \geq m$ generated at random in a rational model defined by a primitive linear representation (ξ, μ, η) of size m. Then, its average value is given by*

$$E(O_n) = \beta(n - m - 1) + a + b + O(\varepsilon^n) \ , \qquad (|\varepsilon| < 1)$$

where β, a and b are real constants defined by

$$\beta = \frac{v'\mu(w)u}{\lambda^m}, \quad a = \frac{\xi' C \mu(w) u}{\lambda^m \xi' u}, \quad b = \frac{v'\mu(w) C \eta}{\lambda^m v' \eta}$$

with λ, v, u and C defined as above.

Proof. From (4) it is easy to derive the equations

$$\xi' M^n \eta = \lambda^n (\xi' u v' \eta + \mathrm{O}(\varepsilon^n)) \tag{6}$$

and

$$\sum_{j=m}^{n} \xi' M^{j-m} \mu(w) M^{n-j} \eta =$$
$$\lambda^{n-m} \left\{ \xi' \left[(n-m+1) u v' \mu(w) u v' + C \mu(w) u v' + u v' \mu(w) C \right] \eta + \mathrm{O}(\varepsilon^n) \right\} \quad (7)$$

Thus, the result follows by replacing the right hand sides of (6) and (7) into the expression

$$E(O_n) = \sum_{j=m}^{n} \frac{\xi' M^{j-m} \mu(w) M^{n-j} \eta}{\xi' M^n \eta}$$

As a comment to the previous result we now make the following remarks:

1. If $m = 1$ the pattern w is reduced to an element of A and we get the average number of symbol occurrences in a primitive rational model obtained in [2].
2. If (ξ, μ, η) is a sequential model then M is a stochastic matrix. Therefore, $\lambda = 1$, $\eta = u = \underline{1}$ and v is the stationary distribution of the underlying Markov chain, here defined by the initial vector ξ and the transition matrix M. As a consequence, we also have $C\eta = 0$, and hence

$$E(O_n) = v' \mu(w) \underline{1}\ (n - m + 1) + \xi' C \mu(w) \underline{1} + \mathrm{O}(\varepsilon^n)\ . \quad (8)$$

   In this case the leading constant $\beta = v' \mu(w) \underline{1}$ is the probability of generating w in the stationary sequential model (v, μ) (i.e. a sort of stationary probability of w).
3. If (ξ, μ, η) is a stationary sequential model (i.e. $\xi = v$), then $\xi' C = 0$ and we get

$$E(O_n) = v' \mu(w) \underline{1}\ (n - m + 1)\ , \quad (9)$$

   which is the equation obtained in [12] (see also [9] Eq. (7.2.25)) in a stationary (primitive) Markovian model of order 1. Thus, our proposition extends the same equality to all stationary (primitive) sequential models.
4. Note that Equation (9) is not true if the sequential model is not stationary (i.e. $\xi \neq v$); in this case constant $\xi' C \mu(w) \underline{1}$ of Equation (8) is not null in general. This means that the stationarity hypothesis is necessary to get (9), even in the Markovian models of order 1.

5 Analysis of the Variance

In this section we study the variance of O_n under the same assumptions of the previous section. Our goal is to determine an asymptotic expression of the form $Var(O_n) = \gamma n + \mathrm{O}(1)$ where γ is a real constant depending on the pattern w and the linear representation (ξ, μ, η).

It turns out that γ is also related to the autocorrelation set of w, a classical notion we here define following the approach proposed in [9]. Assume $w = w_1 \cdots w_m$, where

$w_i \in A$ for each i. For every $1 \leq i \leq j \leq m$ let $w_i^j \in A^+$ be given by $w_i^j = w_i \cdots w_j$. Then, we define the set of indices S and the matrix $P(w)$ given by

$$S = \{k \in \{1, 2, \ldots, m-1\} \mid w_1^k = w_{m-k+1}^m\} , \quad P(w) = \sum_{k \in S} \lambda^k \mu(w_{k+1}^m)$$

Clearly, if $m = 1$ then $S = \emptyset$ and $P(w) = 0$.

Proposition 3. *Under the assumptions of Proposition 2 we have*

$$Var(O_n) = \gamma n + c + O(\varepsilon^n) , \qquad (|\varepsilon| < 1)$$

where γ and c are real constants, the first one being given by

$$\gamma = \beta - (2m-1)\beta^2 + 2\frac{v'\mu(w)\left[C\mu(w) + P(w)\right] u}{\lambda^{2m}} \tag{10}$$

Proof. By Equation (5) and Proposition 2 we have

$$\begin{aligned} E(O_n^2) &= \sum_{j=m}^{n} E(I_j^2) + 2 \sum_{i=m}^{n-1} \sum_{j=i+1}^{n} E(I_i I_j) = \\ &= (n-m-1)\beta + a + b + \\ &+ 2 \sum_{i=m}^{n-m} \sum_{j=i+m}^{n} E(I_i I_j) + 2 \sum_{i=m}^{n-1} \sum_{j=i+1}^{\min\{i+m-1,n\}} E(I_i I_j) \end{aligned} \tag{11}$$

Observe that $E(I_i I_j)$ is easy to evaluate when $i + m \leq j$ since in this case there is no overlap in the occurrences of w at positions i and j. Hence, for every $i = m, \ldots, n-m$ and every $j = i+m, \ldots, n$, we have

$$E(I_i I_j) = \Pr(I_i = 1, I_j = 1) = \frac{\xi' M^{i-m} \mu(w) M^{j-m-i} \mu(w) M^{n-j} \eta}{\xi' M^n \eta}$$

Thus, replacing (4) in the previous equation, by some computation one obtains

$$\begin{aligned} & 2 \sum_{i=m}^{n-m} \sum_{j=i+m}^{n} E(I_i I_j) = \\ & = n^2 \beta^2 + n \left(2\beta(a+b) - \beta^2(4m-3) + 2\frac{v'\mu(w)C\mu(w)u}{\lambda^{2m}} \right) + \mathrm{O}(1) \end{aligned} \tag{12}$$

Now consider the last sum in the right hand side of (11); this term exists only if $m > 1$ and in this case (being $m \notin S$) it can be proved that

$$2 \sum_{i=m}^{n-1} \sum_{j=i+1}^{\min\{i+m-1,n\}} E(I_i I_j) = 2 \sum_{i=m}^{n-m+1} \sum_{k \in S} E(I_i I_{i+m-k}) + \mathrm{O}(1)$$

Again applying (4) we get

$$2\sum_{i=m}^{n-m+1}\sum_{k\in S}E(I_iI_{i+m-k}) = 2n\frac{v'\mu(w)P(w)u}{\lambda^{2m}} + \mathrm{O}(1) \tag{13}$$

Thus, replacing (12) and (13) in (11) and recalling that $Var(O_n) = E(O_n^2) - E(O_n)^2$, the result follows.

Now, let us discuss the previous result in some special cases.

1. In the case $m = 1$ we get the same evaluation of the variance obtained in [2], with $\gamma = \beta - \beta^2 + 2(v'\mu(w)C\mu(w)u)/\lambda^2$.
2. If (ξ, μ, η) is a sequential model (and hence $\beta = v'\mu(w)\underline{1}$) we get

$$\gamma = \beta - (2m-1)\beta^2 + 2v'\mu(w)\left[C\mu(w) + \sum_{k\in S}\mu(w_{k+1}^m)\right]\underline{1}$$

   which generalizes and extends the evaluation of the leading term of the variance obtained in [12] in stationary (primitive) Markovian models of order 1(see also [9, Th. 7.2.8]).
3. If (ξ, μ, η) is just a Markovian model of order 1, Equation (10) is equivalent to Equation (7.2.27) in [9]. Note that the leading term γ does not depend on the initial distribution ξ and is the same as in the stationary case.
4. Also the constant c can be computed explicitly as a function of w and (ξ, μ, η). In particular, in the sequential model, reasoning as in Proposition 3 one gets

$$\begin{aligned} c = {} & (3m^2 - 4m + 1)\beta^2 - 2(2m-1)v'\mu(w)C\mu(w)\underline{1} - 2v'\mu(w)CMC\mu(w)\underline{1} + \\ & -(m-1)\beta - 2(m-1)v'\mu(w)P(w)\underline{1} + 2v'\mu(w)\sum_{\ell=2}^{m-1}\sum_{k\in S, k\geq \ell}\mu(w_{k+1}^m)\underline{1} + \\ & +a - a^2 - 2\beta\xi'CMC\mu(w)\underline{1} + 2\xi'C\mu(w)C\mu(w)\underline{1} + 2\xi'C\mu(w)P(w)\underline{1}. \end{aligned}$$

   Note that, as for the average value, the constant term of the variance depends on the initial distribution ξ. If the sequential model is stationary, the above expression of c further simplifies and the equation reduces to the first two rows, all terms of the third one being null. Hence, the terms in the third row represent the contribution given to the variance by the non-stationary hypothesis.

References

1. J. Berstel and C. Reutenauer. *Rational Series and their Languages*, Springer-Verlag, New York - Heidelberg - Berlin, 1988.
2. A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati. On the number of occurrences of a symbol in words of regular languages. *Theoret. Comput. Sci.*, 302(1-3):431–456, 2003.
3. A. Denise. Génération aléatoire et uniforme de mots de langages rationnels. *Theoret. Comput. Sci.*, 159(1):43–63, 1996.

4. P. Flajolet, P. Zimmerman, and B. Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theoret. Comput. Sci.*, 132(1-2):1–35, 1994.
5. F. R. Gantmacher, *The theory of matrices*, Vol. 2. Chelsea Publishing Co. New York, 1959.
6. M. Goldwurm and R. Radicioni. Probabilistic models for pattern statistics. *RAIRO-Info. Theor. Appl.*, 40:207-225, 2006.
7. G. Hansel and D. Perrin. Rational probability measures. *Theoret. Comput. Sci.*, 65 : 171–188, 1989 (french version in *Mots*, M. Lothaire ed., Hermes, 1990, pp. 335–357).
8. M. Iosifescu. *Finite Markov Processes and Their Applications*, J. Wiley and Sons, 1980.
9. P. Jacket and W. Szpankowski. Analytic approach to pattern matching, Ch.7 in M. Lothaire, *Applied Combinatorics on Words*, Cambridge University Press, 2005.
10. P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoret. Comput. Sci.*, 287(2):593–617, 2002.
11. A. Paz. *Introduction to Probabilistic Automata*, Academic Press, 1971.
12. M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.
13. E. Seneta. *Non-negative Matrices and Markov Chains*, Springer–Verlag, 1981.