Pattern statistics in bicomponent stochastic models*

Diego de Falco, Massimiliano Goldwurm, Violetta Lonati

Università degli Studi di Milano,
Dipartimento di Scienze dell'Informazione,
via Comelico 39, 20135 Milano, Italy
{defalco,goldwurm,lonati}@dsi.unimi.it

Abstract

We give asymptotic estimates of the frequency of occurrences of a symbol in a random word generated by any (non-ergodic) bicomponent stochastic model. More precisely, we consider the random variable Y_n representing the number of occurrences of a given symbol in a word of length n generated at random; the stochastic model is defined by a rational formal series r having a linear representation with two primitive components. This model includes the case when r is the product or the sum of two primitive rational formal series. We obtain asymptotic evaluations for the mean and the variance of Y_n and its limit distribution. These results improve the analysis presented in a recent work dealing with the particular case where r is the product of two primitive rational formal series [5].

Keywords: Frequencies of pattern occurrences, automata and formal languages, limit distributions, Perron–Frobenius theory, rational formal series.

1 Introduction

Estimating the frequency of given patterns in a random text is a classical problem studied in several research areas of computer science and mathematics that has well-known applications in molecular biology [11, 16, 9, 15, 18]. Pattern statistics studies this problem in a probabilistic framework: one or more patterns are fixed and a text of length n is randomly generated by a memoryless source (also called *Bernoulli model*) or a Markovian source (the *Markovian model*) where the probability of a symbol in any position only depends on a finite number of previous occurrences [12, 16, 14]. Among the main goals of the research in this context we recall the asymptotic expressions of mean and variance of the number of pattern occurrences in the text and its limit distribution. Many results show a gaussian limit distribution of the number of pattern occurrences in the sense of the central or local limit theorem [1]. In particular in [14] properties of this kind are obtained for a pattern statistics which represents the number of (positions of) occurrences of words from a regular language in a random string of length n generated in a Bernoulli or a Markovian model.

This approach has been extended in [3, 4, 5] to the so-called *rational stochastic model*, where the pattern is reduced to a single symbol and the text is randomly generated by means of a rational formal series in two non-commutative variables. There are well-known linear time algorithms that generate a random word of given length in such a model [7]. It is proved that the symbol frequency problem in the rational model includes, as a special case, the general frequency problem of regular patterns in the Markovian model (studied in [14]) and it is also known that the two models are not equivalent [3]. The symbol frequency problem in the rational model is studied in [3, 4] in the ergodic case, i.e. when the matrix associated with the rational formal series (counting the transitions between states) is primitive. In [5] we have studied the same problem in a simple non-ergodic rational model, where the formal series is given by the Cauchy product of two primitive rational formal series (the *product model*).

*This work has been supported by the Project M.I.U.R. COFIN "Formal languages and automata: theory and applications".

In the present paper we carry on the analysis considering *bicomponent rational models*, defined by a formal series which admits a linear representation with two primitive components. We obtain asymptotic evaluations for the mean value and the variance of the number of symbol occurrences and its limit distribution: the results strongly depend on whether the matrix defining the transition from the first component to the second (*communication matrix*) is null or not. If it is not null, the results extend those obtained in the product model [5], which occurs when the communication matrix has a special form.

On the other hand, if the communication matrix is null, then the formal series defining the model is simply the sum of two primitive rational formal series (*sum model*) and the results we get are quite different from the previous case. In this paper we present in detail the proofs concerning the sum model and only state the other results. All proofs can be found in the extended version of this work [6].

The material we present is organized as follows. After recalling some preliminaries in Section 2 and the rational stochastic model in Section 3, we revisit the primitive case in Section 4. In Section 5 we introduce the bicomponent rational model; the special case of the sum is studied in Section 6 while in the last one we present the analysis of our statistics in the general bicomponent model.

2 Preliminaries

We summarize some notions of probability theory used in the subsequent sections.

Let X be an integer valued random variable (r.v.), such that $\Pr\{X = k\} = p_k$ for every $k \in \mathbb{N}$. We denote by F_X its distribution function, i.e. $F_X(\tau) = \Pr\{X \leq \tau\}$ for every $\tau \in \mathbb{R}$. If the set of indices $\{k \mid p_k \neq 0\}$ is finite we can consider the moment generating function of X, given by $\Psi_X(z) = \sum_{k\in\mathbb{N}} p_k e^{zk}$ for every $z \in \mathbb{C}$. In this case the first two moments of X can be computed by

$$\mathbb{E}(X) = \Psi'_X(0) \ , \qquad \mathbb{E}(X^2) = \Psi''_X(0) \ . \tag{1}$$

Moreover, the characteristic function of X is defined by

$$\Phi_X(t) = \mathbb{E}(e^{itX}) = \Psi_X(it)$$

The function Φ_X is always well-defined for every $t \in \mathbb{R}$, it is periodic of period 2π and it completely characterizes the function F_X. Moreover it represents the classical tool to prove convergence in distribution. Given a sequence of random variables $\{X_n\}_n$ and a random variable X we say that X_n converges to X *in distribution* (or *in law*) if $\lim_{n\to\infty} F_{X_n}(\tau) = F_X(\tau)$ for every point $\tau \in \mathbb{R}$ of continuity for F_X. It is well-known that X_n converges to X in distribution if and only if $\Phi_{X_n}(t)$ tends to $\Phi_X(t)$ for every $t \in \mathbb{R}$. Several forms of the central limit theorem are classically proved in this way [10, 8].

A convenient approach to prove the convergence in law to a Gaussian random variable relies on the so called "quasi-power" theorems introduced in [13] (see also [8]) and implicitely used in the previous literature [1]. For our purpose it is convenient to recall such a theorem in a simple form.

To this end, let $\{X_n\}$ be a sequence of random variables, where each X_n takes values in $\{0, 1, \ldots, n\}$, defined by a family of non-negative real coefficients $\{c_k^{(n)} \mid n, k \in \mathbb{N}\}$ so that, for every k, n,

$$\Pr\{X_n = k\} = \frac{c_k^{(n)}}{\sum_{j=0}^{n} c_j^{(n)}} \ .$$

Define the function $h_n(z) = \sum_{k=0}^{n} c_k^{(n)} e^{kz}$ and observe that $\Psi_{X_n}(z) = \frac{h_n(z)}{h_n(0)}$. Then, the following property holds (for the proof see [8, Theorem 9.6] or [1, Theorem 1]).

Theorem 1 *Let $\{X_n\}$ and $\{h_n\}$ be defined as above and assume there exist two functions $r(z)$, $u(z)$, both analytic and non-null at $z = 0$, and two positive constants c, ρ, such that for every $|z| < c$*

$$h_n(z) = r(z) \cdot u(z)^n + O(\rho^n) \qquad \textit{and} \qquad \rho < |u(z)|.$$

Also set

$$\mu = \frac{u'(0)}{u(0)} \qquad \textit{and} \qquad \sigma = \frac{u''(0)}{u(0)} - \left(\frac{u'(0)}{u(0)}\right)^2$$

and assume $\sigma > 0$ *(variability condition). Then* $\frac{X_n - \mu n}{\sqrt{\sigma n}}$ *converges in distribution to the normal random variable of mean* 0 *and variance* 1*, i.e. for every* $x \in \mathbb{R}$

$$\lim_{n \longrightarrow +\infty} Pr\left\{\frac{X_n - \mu n}{\sqrt{\sigma n}} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \; .$$

At last, we recall that a sequence of random variable $\{X_n\}$ converges *in probability* to a random variable X if, for every $\varepsilon > 0$, $\Pr\{|X_n - X| > \varepsilon\}$ tends to 0 as n goes to $+\infty$. It is well-known that convergence in probability implies convergence in law.

3 The rational stochastic model

The stochastic model we consider in this work is defined by using the notion of linear representation [2]. Let $\mathbb{R}_+$ be the semiring of non-negative real numbers. A *linear representation* over a binary alphabet $\{a, b\}$ is a triple (ξ, μ, η) such that, for some integer $m > 0$, ξ and η are (column) vectors in $\mathbb{R}_+^m$ and $\mu : \{a, b\}^* \longrightarrow \mathbb{R}_+^{m \times m}$ is a monoid morphism. We say that m is the size of (ξ, μ, η) and, for sake of brevity, we set $A = \mu(a)$ and $B = \mu(b)$ and denote by M the matrix $A + B$.

Such a linear representation defines a rational formal series r in the non-commutative variables a, b, with coefficients in $\mathbb{R}_+$, i.e. a function $r : \{a, b\}^* \longrightarrow \mathbb{R}_+$, such that for any word $w \in \{a, b\}^*$ the value of r at w is $(r, w) = \xi' \mu(w) \eta$, where ξ' denotes the transpose of ξ.

Moreover, for every positive integer n, we can define a probability space as follows. Let us define a *computation path* of length n as a string ℓ of the form

$$\ell = q_0 x_1 q_1 x_2 q_2 \cdots q_{n-1} x_n q_n \tag{2}$$

where $q_j \in \{1, 2, \ldots, m\}$ and $x_i \in \{a, b\}$ for every $j = 0, 1, \ldots, n$ and every $i = 1, 2, \ldots, n$. We denote by Ω_n the set of all computation paths of length n and, for each $\ell \in \Omega_n$ of the form (2), we define the probability of ℓ as

$$\Pr\{\ell\} = \frac{\xi_{q_0} \mu(x_1)_{q_0 q_1} \mu(x_2)_{q_1 q_2} \cdots \mu(x_n)_{q_{n-1} q_n} \eta_{q_n}}{\xi' M^n \eta}$$

Denoting by $\mathcal{P}(\Omega_n)$ the family of all subsets of Ω_n, it is clear that $\langle \Omega_n, \mathcal{P}(\Omega_n), \Pr \rangle$ is a probability space.

Now, let us consider the random variable $Y_n : \Omega_n \to \{0, 1, \ldots, n\}$ such that $Y_n(\ell)$ is the number of a occurring in ℓ, for each $\ell \in \Omega_n$. It is clear that, for every integer $0 \leq k \leq n$, setting

$$\varphi_k^{(n)} = \sum_{|w|=n, |w|_a = k} \xi' \mu(w) \eta \tag{3}$$

we have

$$\Pr\{Y_n = k\} = \frac{\varphi_k^{(n)}}{\sum_{j=0}^{n} \varphi_j^{(n)}}. \tag{4}$$

To study the asymptotic behaviour of Y_n, one should consider the moment generating function of the random variable Y_n which is defined as

$$\Psi_{Y_n}(z) = \frac{h_n(z)}{h_n(0)} \qquad \text{where} \qquad h_n(z) = \sum_{k=0}^{n} \varphi_k^{(n)} e^{zk} = \xi'(Ae^z + B)^n \eta \tag{5}$$

and observe that by (1) we have

$$\mathbb{E}(Y_n) = \frac{h_n'(0)}{h_n(0)} \quad \text{and} \quad \mathbb{V}ar(Y_n) = \frac{h_n''(0) \cdot h_n(0) - [h_n'(0)]^2}{[h_n(0)]^2}. \tag{6}$$

Finally, the characteristic function of the random variable Y_n is given by

$$\Phi_{Y_n}(t) = \mathbb{E}(e^{itY_n}) = \frac{h_n(it)}{h_n(0)} \; .$$

4 The primitive case

In [3, 4] the moments and the limit distribution of Y_n are obtained, in the case when r admits a *primitive* linear representation, i.e. the matrix $M = \mu(a) + \mu(b)$ is primitive. We recall that a nonnegative matrix T is called *primitive* if there exists $p \in \mathbb{N}$ such that all entries of T^p are strictly positive (see for instance [17]). In this section, we recall those results and the main steps of their proofs, which will be useful in subsequent sections.

First of all, observe that under this hypothesis, by Perron–Frobenius Theorem (see [17]) there exists a unique eigenvalue λ of M of maximum modulus which is real and positive. Furthermore, one can associate with λ strictly positive left and right eigenvectors v and u, normed so that $v'u = 1$ and one can prove that, for each $n \in \mathbb{N}$,

$$M^n = \lambda^n \left(uv' + C(n)\right)$$

where $C(n)$ is a real matrix such that $|C(n)_{ij}| = \mathrm{O}(\varepsilon^n)$, for some $0 \leq \varepsilon < 1$ and for any i, j and all n large enough. Moreover, the matrix $C = \sum_{n=0}^{\infty} C(n)$ is well-defined and $v'C = Cu = 0$.

Proposition 2 *If M is primitive and λ is its Perron–Frobenius eigenvalue, then the generating function $h_n(z)$ defined in (5) satisfies the following relations*

$$\begin{aligned} h_n(0) &= \lambda^n \cdot \alpha + O(\rho^n) \\ h_n'(0) &= n\lambda^n \cdot \alpha\beta + \lambda^n \delta + O(\rho^n) \\ h_n''(0) &= n^2\lambda^n \cdot \alpha\beta^2 + n\lambda^n \cdot (\alpha\gamma + 2\beta\delta) + O(\lambda^n) \end{aligned} \tag{7}$$

where $|\rho| < \lambda$ gives the contribution of smaller eigenvalues of M and the constants $\alpha, \beta, \gamma, \delta$ are given by

$$\alpha = \xi' uv' \eta \ , \quad \beta = \frac{v'Au}{\lambda} \ , \quad \gamma = \beta - \beta^2 + 2\frac{v'ACAu}{\lambda^2} \ , \quad \delta = \xi' \frac{CA}{\lambda} uv'\eta + \xi' uv' \frac{AC}{\lambda} \eta \ . \tag{8}$$

From the previous proposition and equation (6) it is easy to prove the following theorem.

Theorem 3 *The mean value and the variance of Y_n satisfy the relations*

$$\mathbb{E}(Y_n) = \beta n + \frac{\delta}{\alpha} + O(\varepsilon^n) \, , \qquad \mathbb{V}ar(Y_n) = \gamma n + O(1) \ , \tag{9}$$

where $0 < \varepsilon < 1$ and $\alpha, \beta, \gamma, \delta$ are defined in (8).

Notice that $B = 0$ implies $\beta = 1$ and $\gamma = \delta = 0$, while $A = 0$ implies $\beta = \gamma = \delta = 0$; on the contrary, if $A \neq 0 \neq B$ then clearly $0 < \beta < 1$.

As far as the limit distribution is concerned, in [3] it is proved that, when M is primitive and $A \neq 0 \neq B$, Y_n converges in law to a Gaussian random variable. To present this result, note that by the Perron-Frobenius Theorem the equation

$$\det\left(uI - Ae^z - B\right) = 0$$

defines an implicit function $u = u(z)$ analytic in a neighbourhood of $z = 0$ such that $u(0) = \lambda$ and $u'(0) \neq 0$. Moreover, the following proposition holds.

Proposition 4 *For every z near 0, as n tends to infinity we have*

$$h_n(z) = r(z) \cdot u(z)^n + O(\rho^n),$$

where $\rho < |u(z)|$ and $r(z)$ is a rational function with respect to e^z and $u(z)$, analytic and non-null at $z = 0$.

Note that from the previous result one can express the moments of Y_n as function of $u(z)$, obtaining

$$\beta = \frac{u'(0)}{\lambda} \ , \quad \gamma = \frac{u''(0)}{\lambda} - \left(\frac{u'(0)}{\lambda}\right)^2 \tag{10}$$

where $\lambda = u(0)$. Finally, in [3] it is shown that if $A \neq 0 \neq B$ then $\gamma > 0$, and hence Theorem 1 applies, yielding

Theorem 5 *If M is primitive and $A \neq 0 \neq B$, then the distribution of $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ converges to the standard normal distribution.*

5 The bicomponent model

Here we consider a linear representation (ξ, μ, η) where the matrix $\mu(a) + \mu(b)$ consists of two primitive components. More formally, we consider a triple (ξ, μ, η) such that there exist two primitive linear representations (ξ_1, μ_1, η_1) and (ξ_2, μ_2, η_2), of size s and t respectively, satisfying the following relations:

$$\xi' = (\xi_1', \xi_2'), \qquad \mu(x) = \begin{pmatrix} \mu_1(x) & \mu_0(x) \\ 0 & \mu_2(x) \end{pmatrix}, \qquad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \tag{11}$$

where $\mu_0(x) \in \mathbb{R}_+^{s\times t}$ for every $x \in \{a, b\}$. In the sequel, we say that (ξ, μ, η) is a *bicomponent* linear representation.

For sake of brevity we use the notations $A_j = \mu_j(a)$, $B_j = \mu_j(b)$ and $M_j = A_j + B_j$ for $j = 0, 1, 2$. Hence, we have

$$A = \mu(a) = \begin{pmatrix} A_1 & A_0 \\ 0 & A_2 \end{pmatrix}, \quad B = \mu(b) = \begin{pmatrix} B_1 & B_0 \\ 0 & B_2 \end{pmatrix}, \quad M = A + B = \begin{pmatrix} M_1 & M_0 \\ 0 & M_2 \end{pmatrix}. \tag{12}$$

To avoid trivial cases, from now on we assume $A \neq 0 \neq B$ and $\xi_1 \neq 0 \neq \eta_2$.

Intuitively, this linear representation corresponds to a weighted non-deterministic finite state automaton (which may have more than one initial state) such that its state diagram consists of two disjoint strongly connected subgraphs, possibly equipped with some further arrows from the first component to the second one. Here a computation path $\ell = q_0 x_1 q_1 x_2 q_2 \cdots q_{n-1} x_n q_n$ can be of three different kinds:

1. All q_j's are in the first component (in which case we say that ℓ is *contained* in the first component);
2. There is an index $0 \leq s < n$ such that the indices $q_0, q_1, \dots, q_s$ are in the first component while $q_{s+1}, \dots, q_n$ are in the second one. In this case x_{s+1} is the label of the transition from the first to the second component;
3. All q_j's are in the second component (in which case we say that ℓ is *contained* in the second component).

Using the notation introduced in the previous section, from now on the function $h_n(z)$ defined in (5) is referred to the linear representation (ξ, μ, η). From the decomposition (12) it is easy to see that $h_n(z)$ can be written in the form

$$h_n(z) = h_n^{(1)}(z) + g_n(z) + h_n^{(2)}(z)$$

where $h_n^{(1)}$, g_n and $h_n^{(2)}$ correspond to the three kinds of computation paths of the automaton. More precisely, for $j = 1, 2$, we have

$$h_n^{(j)}(z) = \xi_j'(A_j e^z + B_j)^n \eta_j$$

that is $h_n^{(j)}$ is the generating function of the primitive component (ξ_j, μ_j, η_j) and hence it satisfies the properties of Section 4. Moreover

$$g_n(z) = \sum_{i=0}^{n-1} \xi_1'(A_1 e^z + B_1)^i (A_0 e^z + B_0)(A_2 e^z + B_2)^{n-1-i} \eta_2$$

The bicomponent model introduced so far includes two special cases which occur respectively when the formal series defined by (ξ, μ, η) is the sum or the product of two rational formal series having primitive linear representation.

Sum model: let r be the series defined by

$$(r, \omega) = \xi_1' \mu_1(\omega) \eta_1 + \xi_2' \mu_2(\omega) \eta_2 \qquad \forall \omega \in \{a, b\}^*$$

where (ξ_j, μ_j, η_j) is a primitive linear representation for $j = 1, 2$. Clearly, r admits a bicomponent linear representation (ξ, μ, η) which satisfies (11) and such that $M_0 = 0$. As a consequence, the computation paths of type 2 cannot occur and hence

$$h_n(z) = h_n^{(1)}(z) + h_n^{(2)}(z)$$

Product model: consider the formal series

$$(r, \omega) = \sum_{\omega = xy} \pi_1' \, \nu_1(x) \, \tau_1 \cdot \pi_2' \, \nu_2(y) \, \tau_2 \qquad \forall \omega \in \{a, b\}^*$$

where (π_j, ν_j, τ_j) is a primitive linear representation for $j = 1, 2$. Then, r admits a bicomponent linear representation (ξ, μ, η) such that

$$\xi' = (\pi_1', 0), \qquad \mu(x) = \begin{pmatrix} \nu_1(x) & \tau_1 \, \pi_2' \, \nu_2(x) \\ 0 & \nu_2(x) \end{pmatrix}, \qquad \eta = \begin{pmatrix} \tau_1 \, \pi_2' \, \tau_2 \\ \tau_2 \end{pmatrix}.$$

In this case, the three terms of $h_n(z)$ can be merged in a unique convolution

$$h_n(z) = \sum_{i=0}^{n} \xi_1'(A_1 e^z + B_1)^i \, \tau_1 \, \pi_2' \, (A_2 e^z + B_2)^{n-i} \eta_2$$

To study the random variable Y_n, one can consider the bivariate generating function $\sum_{n=0}^{\infty} h_n(z) w^n$ of the sequence $\{h_n(z)\}_n$ and analyse its singularities. It turns out that the main contribution always depends on $g_n(z)$ and hence on M_0. As a consequence, if $M_0 \neq 0$ the bicomponent model is well represented by the product model; on the other hand, if $M_0 = 0$ then $g_n(z)$ vanishes and we have the sum model; this last case is considered in detail in Section 6.

The properties of Y_n depend on whether the Perron-Frobenius eigenvalues λ_1, λ_2 of M_1 and M_2 are distinct or equal. In the first case the rational representation associated with the largest one determines the main characteristics of Y_n. We say that (ξ_j, μ_j, η_j) is the *dominant* component if $\lambda_1 \neq \lambda_2$ and $\lambda_j = \max\{\lambda_1, \lambda_2\}$. On the contrary, if $\lambda_1 = \lambda_2$ we say that the components are *equipotent* and they both give a contribution to the asymptotic behaviour of Y_n.

In the following sections we extend the notation introduced so far, by appending indeces 1 and 2 to the values associated with the linear representation (ξ_1, μ_1, η_1) and (ξ_2, μ_2, η_2), respectively. Thus, for each $j = 1, 2$, the values $Y_n^{(j)}$, u_j, v_j, C_j, α_j, β_j, γ_j, δ_j are well-defined and associated with the linear representation (ξ_j, μ_j, η_j).

6 Analysis of the sum model

In this section we study the behaviour of Y_n assuming $M_0 = 0$. This case corresponds to the case where the stochastic model is defined by the sum of two primitive formal series, having linear representation (ξ_1, μ_1, η_1) and (ξ_2, μ_2, η_2), respectively. Since here $M_0 = 0$, to avoid trivial cases, we also assume $\xi_2 \neq 0 \neq \eta_1$.

We recall that the main difference with respect to the general analysis is that here $g_n(z)$ disappears and hence

$$h_n(z) = h_n^{(1)}(z) + h_n^{(2)}(z) \ .$$

Thus, if $\lambda_1 > \lambda_2$ the leading term is $h_n^{(1)}(z)$ and hence $h_n(z)$ behaves almost as in the primitive case. On the other side, if $\lambda_1 = \lambda_2$, the bivariate generating function of $\{h_n(z)\}_n$ has a simple pole in the main sigularity, due to the contribution of both $h_n^{(1)}(z)$ and $h_n^{(2)}(z)$.

6.1 Dominant component in the sum model

In this section we study the behaviour of $\{Y_n\}$ assuming $\lambda_1 > \lambda_2$ (the case $\lambda_1 < \lambda_2$ is symmetric). We first determine asymptotic expressions for mean and variance of Y_n and then we study its limit distribution.

Proposition 6 *In the sum model, if $\lambda_1 > \lambda_2$ then the mean value and variance of Y_n satisfy the following relations:*

$$\mathbb{E}(Y_n) = \beta_1 n + \frac{\delta_1}{\alpha_1} + O(\varepsilon^n) \ , \qquad \mathbb{V}ar(Y_n) = \gamma_1 n + O(1) \ ,$$

where $0 < \varepsilon < 1$.

Proof. To find asymptotic expressions for $h_n(0) = h_n^{(1)}(0) + h_n^{(2)}(0)$ and its derivatives, since M_1 and M_2 are primitive, we apply Proposition 2 to both $h_n^{(1)}(z)$ and $h_n^{(2)}(z)$. Being $\lambda_1 > \lambda_2$, the main contribution is given by the first component, and hence $h_n(z)$ behaves almost as in the primitive case. Indeed, we get

$$\begin{aligned} h_n(0) &= \lambda_1^n \alpha_1 + O(\rho^n) \\ h'_n(0) &= n\lambda_1^n\, \alpha_1\beta_1 + \lambda_1^n\, \delta_1 + O(\rho^n) \\ h''_n(0) &= n^2\lambda_1^n\, \alpha_1\beta_1^2 + n\lambda_1^n(\beta_1\delta_1 + \alpha_1\gamma_1) + O(\lambda_1^n) \end{aligned}$$

where $|\rho| < \lambda_1$. Then, the result follows from (6). □

As far as the limit distribution is concerned, observe that if the main component does not degenerate (i.e. assume $A_1 \neq 0 \neq B_1$), then $\beta_1 > 0$ and $\gamma_1 > 0$. Moreover $h_n^{(1)}(z)$ satisfies Proposition 4, and hence by Theorem 1 we obtain the following result.

Theorem 7 *In the sum model, if $\lambda_1 > \lambda_2$ and $A_1 \neq 0 \neq B_1$ then the distribution of $\frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}}$ converges to the normal standard distribution.*

Now consider the degenerate cases $A_1 = 0$ or $B_1 = 0$ (note that they cannot occur at the same time, otherwise $M_1 = 0$ and the first component vanishes). If $B_1 = 0$ then $\beta_1 = 1$ and $\gamma_1 = \delta_1 = 0$, hence we get $\mathbb{E}(Y_n) = n + O(\varepsilon^n)$, $0 < \varepsilon < 1$. On the other side, if $A_1 = 0$ then $\beta_1 = \gamma_1 = \delta_1 = 0$ and hence we get $\mathbb{E}(Y_n) = O(\varepsilon^n)$. In both cases we have $\gamma_1 = 0$ and a direct computation proves $\mathbb{V}ar(Y_n) = O(\varepsilon^n)$, showing that Y_n almost surely reduces to a single value (n or 0, respectively). Indeed, by Chebyshev's inequality, if $B_1 = 0$ we have for every $c > 0$

$$\Pr\{|Y_n - n| > c\} \leq \frac{\mathbb{V}ar(Y_n)}{c^2} = \mathrm{O}(\varepsilon^n)$$

and hence, $Y_n - n = o(1)$ in probability. A similar result can be obtained in the case $A_1 = 0$.

Theorem 8 *In the sum model, assume $\lambda_1 > \lambda_2$. If $B_1 = 0$ (resp. $A_1 = 0$) then $n - Y_n$ (resp. Y_n) tends to 0 in probability.*

6.2 Equipotent components in the sum model

Here we study the behaviour of Y_n assuming $\lambda_1 = \lambda_2$. Under this hypothesis two main subcases arise. The first one occurs when the constants β_1 and β_2 characterizing the mean value of $Y_n^{(1)}$ and $Y_n^{(2)}$ are different. In this case the variance of Y_n is of the order $\Theta(n^2)$ and Y_n itself approximates a random variable which may only assume two values. On the contrary, when $\beta_1 = \beta_2$ the order of growth of the variance reduces to $\Theta(n)$ and hence the asymptotic behaviour of Y_n is again concentrated around its expected value and the limit distribution is a mixture of gaussians.

As before we first study the asymptotic behaviour of the moments of Y_n and then we turn our attention to the limit distributions. For sake of brevity, let $\lambda_1 = \lambda_2 = \lambda$.

Proposition 9 *In the sum model, assume $\lambda_1 = \lambda_2$. If $\beta_1 \neq \beta_2$ then*

$$\mathbb{E}(Y_n) = n \cdot \frac{\alpha_1\beta_1 + \alpha_2\beta_2}{\alpha_1 + \alpha_2} + O(1)\ , \quad \mathbb{V}ar(Y_n) = n^2 \cdot \frac{\alpha_1\alpha_2\, (\beta_1 - \beta_2)^2}{(\alpha_1 + \alpha_2)^2} + O(n)\ .$$

If $\beta_1 = \beta_2 = \beta$ then

$$\mathbb{E}(Y_n) = n \cdot \beta + O(1)\ , \quad \mathbb{V}ar(Y_n) = n \cdot \frac{\alpha_1\gamma_1 + \alpha_2\gamma_2}{\alpha_1 + \alpha_2} + O(1)\ .$$

Proof. To find asymptotic expressions for $h_n(0) = h_n^{(1)}(0) + h_n^{(2)}(0)$ and its derivatives, since M_1 and M_2 are primitive, we apply Proposition 2 to both $h_n^{(1)}(z)$ and $h_n^{(2)}(z)$. Being $\lambda_1 = \lambda_2 = \lambda$, the contributions of both components are relevant, and hence we get

$$\begin{aligned} h_n(0) &= \lambda^n(\alpha_1 + \alpha_2) + O(\rho^n) \\ h'_n(0) &= n\lambda^n(\alpha_1\beta_1 + \alpha_2\beta_2) + \lambda^n(\delta_1 + \delta_2) + O(\rho^n) \\ h''_n(0) &= n^2\lambda^n(\alpha_1\beta_1^2 + \alpha_2\beta_2^2) + n\lambda^n(\beta_1\delta_1 + \beta_2\delta_2 + \alpha_1\gamma_1 + \alpha_2\gamma_2) + O(\lambda^n) \end{aligned}$$

where $|\rho| < \lambda$. Hence from (6) we get the following results, which prove the statement:

$$\begin{aligned} \mathbb{E}(Y_n) &= n \cdot \frac{\alpha_1\beta_1 + \alpha_2\beta_2}{\alpha_1 + \alpha_2} + \frac{\delta_1 + \delta_2}{\alpha_1 + \alpha_2} + O(\varepsilon^n) , \\ \mathbb{V}ar(Y_n) &= n^2 \cdot \frac{\alpha_1\alpha_2 \, (\beta_1 - \beta_2)^2}{(\alpha_1 + \alpha_2)^2} + n \cdot \left(\frac{\alpha_1\gamma_1 + \alpha_2\gamma_2}{\alpha_1 + \alpha_2} + \frac{2(\beta_1 - \beta_2)(\alpha_2\delta_1 - \alpha_1\delta_2)}{(\alpha_1 + \alpha_2)^2} \right) + O(1) . \end{aligned}$$

□

Now, let us study the limit distribution. Let U_n be the Bernoullian random variable $U_n : \Omega_n \to \{0, 1\}$ such that for each $\ell \in \Omega_n$

$$U_n(\ell) = \begin{cases} 1 & \text{if } \ell \text{ is entirely contained in the first component,} \\ 0 & \text{if } \ell \text{ is entirely contained in the second component.} \end{cases}$$

It is easy to show that

$$\Pr\{U_n = x\} = \begin{cases} \frac{\xi_1' M_1^n \eta_1}{\xi' M^n \eta} & \text{if } x = 1 , \\ \\ \frac{\xi_2' M_2^n \eta_2}{\xi' M^n \eta} & \text{if } x = 0 . \end{cases}$$

Furthermore, let $L_n = \beta_1 U_n + \beta_2(1 - U_n)$ and observe that if $\beta_1 = \beta_2$, then $L_n = \beta_1 = \beta_2$.

Proposition 10 *In the sum model, if $\lambda_1 = \lambda_2$ then the random variable $\frac{Y_n}{n} - L_n$ converges to 0 in probability.*

Proof. We first evaluate the variance of $Y_n - nL_n$. Clearly Y_n and L_n are not independent, but we can express their dependence by writing $Y_n = U_n Y_n^{(1)} + (1 - U_n) Y_n^{(2)}$ and hence

$$Y_n - nL_n = U_n \cdot (Y_n^{(1)} - n\beta_1) + (1 - U_n) \cdot (Y_n^{(2)} - n\beta_2)$$

where U_n is independent of $Y_n^{(i)}$, for each $i = 1, 2$.

Moreover, by the previous proposition $\mathbb{E}(Y_n - nL_n) = \mathrm{O}(1)$ and so

$$\begin{aligned} \mathbb{V}ar(Y_n - nL_n) &= \mathbb{E}((Y_n - nL_n)^2) + \mathrm{O}(1) = \sum_{i=0,1} \mathbb{E}((Y_n - nL_n)^2 \mid U_n = i) \cdot \Pr\{U_n = i\} + \mathrm{O}(1) \\ &= \sum_{j=1,2} \mathbb{E}((Y_n^{(j)} - n\beta_j)^2) \cdot \frac{\alpha_j}{\alpha_1 + \alpha_2} + \mathrm{O}(1) = n \cdot \frac{\alpha_1\gamma_1 + \alpha_2\gamma_2}{\alpha_1 + \alpha_2} + \mathrm{O}(1) . \end{aligned}$$

Thus, by Chebyshev's inequality, for every $c > 0$ one gets

$$\Pr\left\{ \left| \frac{Y_n}{n} - L_n \right| \geq c \right\} = \mathrm{O}\left(\frac{1}{n} \right) .$$

□

Since convergence in probability implies convergence in law we obtain the following

Corollary 11 *In the sum model, if $\lambda_1 = \lambda_2$ then the distribution of Y_n/n converges to the distribution having probability mass $\frac{\alpha_1}{\alpha_1+\alpha_2}$ at β_1 and probability mass $\frac{\alpha_2}{\alpha_1+\alpha_2}$ at β_2.*

The above results intuitively state that $Y_n \sim nL_n$, where L_n may only assume two values. Thus, a natural question concernes the limit distribution of $Y_n - nL_n$. To deal with this problem assume $\gamma_1 \neq 0 \neq \gamma_2$ and consider the random variable $\mathcal{V}$ constructed by considering a Bernoullian r.v. U of parameter $p = \alpha_1/(\alpha_1 + \alpha_2)$, two normal r.v.'s N_1, N_2 of mean 0 and variance γ_1 and γ_2, respectively, and setting

$$\mathcal{V} = U \cdot N_1 + (1 - U) \cdot N_2 \tag{13}$$

where we assume U, N_1, N_2 independent of one another. Note that, if $\gamma_1 = \gamma_2$ then $\mathcal{V}$ is a normal random variable of mean 0 and variance $\gamma_1 = \gamma_2$. The characteristic function of $\mathcal{V}$ is given by

$$\mathbb{E}(e^{it\mathcal{V}}) = \frac{\alpha_1}{\alpha_1 + \alpha_2} e^{-\frac{\gamma_1}{2}t^2} + \frac{\alpha_2}{\alpha_1 + \alpha_2} e^{-\frac{\gamma_2}{2}t^2} .$$

Proposition 12 *In the sum model, if* $\lambda_1 = \lambda_2$ *and* $\gamma_1 \neq 0 \neq \gamma_2$ *then the distribution of* $\frac{Y_n - nL_n}{\sqrt{n}}$ *converges to the mixture, with weights* $\frac{\alpha_1}{\alpha_1+\alpha_2}$ *and* $\frac{\alpha_2}{\alpha_1+\alpha_2}$*, of two normal distributions with mean zero and variance* γ_1 *and* γ_2 *respectively. In particular, if* $\gamma_1 = \gamma_2 = \gamma$ *then* $\frac{Y_n - nL_n}{\sqrt{n\gamma}}$ *converges in law to the standard normal random variable.*

Proof. Let us define the r.v. $\mathcal{V}_n = \frac{Y_n - nL_n}{\sqrt{n}}$. Its characteristic function is given by

$$\begin{aligned}\mathbb{E}(e^{it\mathcal{V}_n}) &= \sum_{i=0,1} \mathbb{E}(e^{it\mathcal{V}_n} \mid U_n = i) \cdot \Pr\{U_n = i\} = \sum_{j=1,2} \mathbb{E}\left(e^{it\frac{Y_n^{(j)} - n\beta_j}{\sqrt{n}}}\right) \cdot \left(\frac{\alpha_j}{\alpha_1 + \alpha_2} + \mathrm{O}(\varepsilon^n)\right) \\ &= \frac{\alpha_1}{\alpha_1 + \alpha_2} e^{-\frac{\gamma_1}{2}t^2} + \frac{\alpha_2}{\alpha_1 + \alpha_2} e^{-\frac{\gamma_2}{2}t^2} + \mathrm{O}\left(n^{-1/2}\right) .\end{aligned}$$

□

The previous results hold even if $\beta_1 = \beta_2 = \beta$; notice that in that case L_n reduces to the constant β and $\gamma_1 \neq 0 \neq \gamma_2$ otherwise either $A = 0$ or $B = 0$. Hence we obtain the following

Corollary 13 *In the sum model, assume* $\lambda_1 = \lambda_2$ *and* $\beta_1 = \beta_2 = \beta$ *. Then the distribution of* $\frac{Y_n - n\beta}{\sqrt{n}}$ *converges to the mixture, with weights* $\frac{\alpha_1}{\alpha_1+\alpha_2}$ *and* $\frac{\alpha_2}{\alpha_1+\alpha_2}$*, of two normal distributions with mean zero and variance* γ_1 *and* γ_2 *respectively. In particular, if* $\gamma_1 = \gamma_2 = \gamma$ *then* $\frac{Y_n - n\beta}{\sqrt{n\gamma}}$ *converges in law to the standard normal random variable.*

7 Analysis of the general model

In this section, we consider the bicomponent model in the general case when $M_0 \neq 0$. The results we present here extend those obtained in [5] for the product model, which now becomes a particular case. We prove that the limit distributions for the dominant non-degenerate case and for the equipotent case are the same as in the product model. Hence in these cases they do not depend on the matrix M_0. On the contrary, in the dominant degenerate case, the limit distribution is specific for each model, depends on the matrix M_0 and also on the dominated component, even via its eigenvalues of lower modulus.

In the proofs, the main difference with respect to the previous sections is that now $g_n(z)$ is not null and $h_n(z)$ always depends on its contribution. Due to space constraints, all proofs of this section are omitted and can be found in [6]. We simply observe that most of them are based on a sort of singularity analysis for matrix functions that can be developed in the same way as for traditional analytic functions.

We consider separately the case $\lambda_1 > \lambda_2$ (the case $\lambda_1 < \lambda_2$ is symmetric) and the case $\lambda_1 = \lambda_2$. In both cases, we first determine asymptotic expressions for mean and variance of Y_n and then we study its limit distribution.

7.1 Dominant component in the general model

Assuming $\lambda_1 > \lambda_2$, the analysis of Y_n depends on whether the dominant component degenerates (i.e. $A_1 = 0$ or $B_1 = 0$). If this is not the case, the results are the same as in the sum model. This is due to the fact that now $g_n(z)$ gives a contribution to $h_n(z)$ of the same order as $h_n^{(1)}$ and this allows us to argue as in Section 6.1. On the other hand, if the dominant component degenerates, a different reasoning is needed, where a key role is played by the matrix Q defined by

$$Q = (\lambda_1 I - M_2)^{-1} \tag{14}$$

The following proposition gives asymptotic expressions for mean value and variance.

Proposition 14 *Assume* $M_0 \neq 0$ *and* $\lambda_1 > \lambda_2$*. Then the mean value and variance of* Y_n *satisfy the following relations:*

1. *If* $A_1 \neq 0 \neq B_1$ *then* $\mathbb{E}(Y_n) = \beta_1 n + O(1)$ *and* $\mathbb{V}ar(Y_n) = \gamma_1 n + O(1)$*, where* $\beta_1 > 0$ *and* $\gamma_1 > 0$*;*
2. *If* $B_1 = 0$ *then* $\mathbb{E}(Y_n) = n - \frac{v_1'(B_0 + M_0 Q B_2) Q \eta_2}{v_1'(\eta_1 + M_0 Q \eta_2)} + O(\varepsilon^n)$ *and* $\mathbb{V}ar(Y_n) = c + O(1)$*;*

3. *If* $A_1 = 0$ *then* $\mathbb{E}(Y_n) = \frac{v_1'(A_0+M_0QA_2)Q\eta_2}{v_1'(\eta_1+M_0Q\eta_2)} + O(\varepsilon^n)$ *and* $\mathbb{V}ar(Y_n) = c + O(1)$*;*

where $0 < \varepsilon < 1$ *and* $c > 0$*.*

The same classification holds for the limit distributions.

Theorem 15 *Assume* $M_0 \neq 0$ *and* $\lambda_1 > \lambda_2$*. Then the following statements hold:*

1. *If* $A_1 \neq 0 \neq B_1$ *then the distribution of* $\frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}}$ *converges to the standard normal distribution;*

2. *If* $B_1 = 0$ *then the distribution of* $n - Y_n$ *converges to the distribution having characteristic function*

$$\Phi_1(t) = \frac{v_1'\eta_1 + v_1'(A_0 + B_0e^{it})(\lambda_1 I - A_2 - B_2e^{it})^{-1}\eta_2}{v_1'(\eta_1 + M_0Q\eta_2)} \ ;$$

3. *If* $A_1 = 0$ *then the distribution of* Y_n *converges to the distribution having characteristic function*

$$\Phi_2(t) = \frac{v_1'\eta_1 + v_1'(A_0e^{it} + B_0)(\lambda_1 I - A_2e^{it} - B_2)^{-1}\eta_2}{v_1'(\eta_1 + M_0Q\eta_2)} \ . \tag{15}$$

The random variables Z_1 and Z_2 of characteristic functions Φ_1 and Φ_2 respectively may assume a large variety of possible forms. The simplest cases occur when the matrices M_1 and M_2 have size 1×1 and hence $M_1 = \lambda_1$, $M_2 = \lambda_2$ and both A_2 and B_2 are constants. In this case $Z_1 = W(X + G)$, where X and W are Bernoullian r.v. of parameter p_x and p_w, respectively given by

$$p_x = B_0/M_0 \qquad \text{and} \qquad p_w = \frac{M_0(\lambda_1 - \lambda_2)^{-1}\eta_2}{\eta_1 + M_0(\lambda_1 - \lambda_2)^{-1}\eta_2} \ ,$$

while G is a geometric r.v. of parameter $B_2/(\lambda_1 - A_2)$. More complicated forms for Z_1 and Z_2 occur when the matrices M_1 and M_2 have more than one entry. Some examples of their behaviour can be found in [5] in the case of the product model.

7.2 Equipotent component in the general model

Now, we consider the behaviour of Y_n assuming $\lambda_1 = \lambda_2$. Under this hypothesis two main subcases arise. The first one occurs when the constants β_1 and β_2 are different. In this case the variance of Y_n is of the order $\Theta(n^2)$ and Y_n itself converges in distribution to a uniform random variable (note that this distribution is different from the one obtained in the sum model). On the contrary, when $\beta_1 = \beta_2$ the order of growth of the variance reduces to $\Theta(n)$ and hence the asymptotic behaviour of Y_n is again concentrated around its expected value, as for the sum.

The following proposition gives asymptotic expressions for mean value and variance.

Proposition 16 *Assume* $M_0 \neq 0$ *and* $\lambda_1 = \lambda_2 = \lambda$*. Then the following statements hold:*

1. *If* $\beta_1 \neq \beta_2$*, then* $\mathbb{E}(Y_n) = \frac{\beta_1+\beta_2}{2}\, n + O(1)$ *and* $\mathbb{V}ar(Y_n) = \frac{(\beta_1-\beta_2)^2}{12}\, n^2 + O(n)$*;*

2. *If* $\beta_1 = \beta_2 = \beta$*, then* $\mathbb{E}(Y_n) = \beta\, n + O(1)$ *and* $\mathbb{V}ar(Y_n) = \frac{\gamma_1+\gamma_2}{2}\, n + O(1)$ *, where* $\gamma_i > 0$ *for some* $i \in \{1, 2\}$*.*

As far as the limit distribution is concerned, we obtain three different cases, summarized by the following

Theorem 17 *Assume* $M_0 \neq 0$*,* $\lambda_1 = \lambda_2 = \lambda$ *and set* $\gamma = \frac{\gamma_1+\gamma_2}{2}$*. Then the following statements hold:*

1. *If* $\beta_1 \neq \beta_2$*, then the distribution of* Y_n/n *converges to the uniform distribution in the interval* $[b_1, b_2]$*, where* $b_1 = \min\{\beta_1, \beta_2\}$ *and* $b_2 = \max\{\beta_1, \beta_2\}$*;*

2. *If* $\beta_1 = \beta_2$ *and* $\gamma_1 \neq \gamma_2$ *then the distribution of* $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ *converges to the distribution having characteristic function*

$$\Phi(t) = \frac{e^{-\frac{\gamma_2}{2\gamma}t^2} - e^{-\frac{\gamma_1}{2\gamma}t^2}}{(\frac{\gamma_1}{2\gamma} - \frac{\gamma_2}{2\gamma})t^2} \tag{16}$$

3. *If* $\beta_1 = \beta_2$ *and* $\gamma_1 = \gamma_2$ *then the distribution of* $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ *converges to the standard normal distribution.*

By direct inspection, one can see that the characteristic function (16) describes a mixture of Gaussian distribution of mean 0, with variances uniformly distributed in the interval with extremes $\frac{\gamma}{\gamma_1}$ and $\frac{\gamma}{\gamma_2}$. Indeed:

$$\Phi(t) = \frac{1}{\left(\frac{\gamma_2}{\gamma} - \frac{\gamma_1}{\gamma}\right)} \int_{\frac{\gamma_1}{\gamma}}^{\frac{\gamma_2}{\gamma}} e^{-\frac{1}{2}vt^2} dv$$

References

[1] E. A. Bender. Central and local limit theorems applied to asymptotic enumeration. *Journal of Combinatorial Theory*, 15:91–111, 1973.

[2] J. Berstel and C. Reutenauer. *Rational series and their languages*, Springer-Verlag, New York - Heidelberg - Berlin, 1988.

[3] A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati. On the number of occurrences of a symbol in words of regular languages. *Theoretical Computer Science*, 302(1-3):431–456, 2003.

[4] A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati. The symbol-periodicity of irreducible finite automata. *Rapporto Interno n. 277-02*, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, April 2002.

[5] D. de Falco, M. Goldwurm, and V. Lonati. Frequency of symbol occurrences in simple non-primitive stochastic models. To appear in Proceedings 7th D.L.T. Conference, Lecture Notes in Computer Science, Springer, 2003; extended version in *Rapporto Interno n. 287-03*, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, February 2003.

[6] D. de Falco, M. Goldwurm, and V. Lonati. Pattern statistics in bicomponent stochastic models (extended version). *Rapporto Interno*, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, June 2003 (available at http://homes.dsi.unimi.it/∼goldwurm/home.html).

[7] A. Denise. Génération aléatoire uniforme de mots de langages rationnels. *Theoretical Computer Science*, 159:43–63, 1996.

[8] P. Flajolet and R. Sedgewick. The average case analysis of algorithms: multivariate asymptotics and limit distributions. *Rapport de recherche* n. 3162, INRIA Rocquencourt, May 1997.

[9] M. S. Gelfand. Prediction of function in DNA sequence analysis. *J. Comput. Biol.*, 2:87–117, 1995.

[10] B.V. Gnedenko. *The theory of probability* (translated by G. Yankovsky). Mir Publishers - Moscow, 1976.

[11] L. J. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM J. Appl. Math.*, 35:401–418, 1978.

[12] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A*, 30(2):183–208, 1981.

[13] H.K. Hwang. *Théorèmes limites pour les structures combinatoires et les fonctions arithmétiques.* Ph.D. Dissertation, École polytechnique, Palaiseau, France, 1994.

[14] P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. In *Proceedings of the 7th ESA*, J. Nešetril editor. Lecture Notes in Computer Science, vol. n.1643, Springer, 1999, 194–211.

[15] B. Prum, F. Rudolphe and E. Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequence. *J. Roy. Statist. Soc. Ser. B*, 57: 205–220, 1995.

[16] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.

[17] E. Seneta. *Non-negative matrices and Markov chains*, Springer–Verlag, New York Heidelberg Berlin, 1981.

[18] M. Waterman. *Introduction to computational biology*, Chapman & Hall, New York, 1995.