

Analysis of symbol statistics in bicomponent rational models [★]

M. Goldwurm⁽¹⁾, J. Lin⁽²⁾, M. Vignati⁽¹⁾

(1) Dipartimento di Matematica, Università degli Studi di Milano, Italy

(2) Department of Mathematics, Khalifa University,
Abu Dhabi - United Arab Emirates

Abstract. We study the local limit distribution of sequences of random variables representing the number of occurrences of a symbol in words of length n in a regular language, generated at random according to a rational stochastic model. We present an analysis of the main local limits when the finite state automaton defining the stochastic model consists of two primitive components. Our results include an evaluation of the convergence rate, which in the various cases is of an order slightly slower than $O(n^{-1/2})$.

1 Introduction

This work continues the analysis developed in [3,7,10] on the limit distribution of the number of symbol occurrences in words of given length, chosen at random in regular languages. More precisely, we consider sequences of random variables $\{Y_n\}$, where each Y_n is the number of occurrences of a symbol a in a word w of length n , generated at random in a *rational stochastic model*. Such a model can be formally defined by a finite state automaton with real positive weights on transitions. In this setting the probability of generating a word w is proportional to the weight the automaton associates with w ; thus, the language recognized by the automaton is the family of all words having non-null probability to be generated. This model is quite general, it includes as special cases the traditional Bernoullian and Markovian sources [14,13] and contains the random generation of words of length n in any regular language under uniform distribution.

The properties of $\{Y_n\}$ are of particular interest for the analysis of regular patterns occurring in words generated by Markovian models [14,13,3] and for the asymptotic estimate of the coefficients of rational series in commutative variables [3,4]. They are also related to the study of the descriptive complexity of languages and computational models [5] and to the analysis of the values of additive functions defined on regular languages [11]. Clearly, the asymptotic behaviour of $\{Y_n\}$ depends on the properties of the finite state automaton \mathcal{A} defining the stochastic model. It is known that if \mathcal{A} has a primitive transition matrix then

[★] Accepted in revised form at DLT 2019, 23rd International Conference on Developments in Language Theory, August 5-9, 2019 (Warsaw, Poland), to appear in Lecture Notes in Computer Science, Springer.

Y_n has a Gaussian limit distribution [13,3] and, under a suitable aperiodicity condition, it also satisfies a local limit theorem [3]. The limit distribution of Y_n in the global sense is known also when the transition matrix of \mathcal{A} consists of two primitive components [7] and a first (non-Gaussian) local limit theorem in a particular bicomponent case is presented in [10].

Here we improve these results presenting an analysis of the local limits of $\{Y_n\}$ when the transition matrix of \mathcal{A} consists of two primitive components equipped with some transition from the first to the second component. At the cost of adding suitable aperiodicity conditions, we prove that the main convergences in distribution obtained in [7] also hold true in the local sense. Moreover, we evaluate the rates of convergence to our limits both in the primitive case and in all bicomponent cases (a tight convergence rate is a natural goal in these contexts [12]). Our results are obtained by applying the Saddle Point Method [8, Chapter VII] and, as our limit densities often are not normal, proofs can be regarded as an application of this tool in non-Gaussian cases ¹.

In this context it is crucial to observe that a local limit theorem does not follow immediately from a traditional convergence in distribution (which occurs for instance in the usual central limit theorems), since single probabilities are differences of values of the corresponding distribution functions, and hence they may not be detected by a standard analysis of convergence in law. Usually, in order to prove a local limit theorem from a convergence in distribution, some additional regularity or aperiodicity conditions are necessary; standard counterexamples show that such conditions cannot be avoided [9,4].

The material we present is organized as follows. In Section 2 we define the problem, recalling the notions of convergence in distribution and local limit law. In Section 3 we revise the primitive case stating a local limit theorem for our statistics Y_n with a convergence rate of the order $O(n^{-1/2})$. In Section 4 we study the behaviour of Y_n in (communicating) bicomponent models: first we show a Gaussian local limit property when there is a dominant component, yielding a convergence rate analogous to the primitive model. Then, in Subsection 4.1, we consider the equipotent bicomponent case, occurring when the main eigenvalues of the two components coincide; in this case the results depend on the values of four constants: β_1, γ_1 and β_2, γ_2 , representing the leading terms of mean value and variance of our statistics associated to the first and the second component, respectively. When $\beta_1 \neq \beta_2$ we strengthen the result on local limit towards a uniform density obtained in [10] by showing a convergence rate “almost” of the order $O(n^{-1/2} \log^{3/2} n)$. If $\beta_1 = \beta_2$ but $\gamma_1 \neq \gamma_2$, then the local limit density turns out to be a suitable mixture of Gaussian densities, with a convergence rate “almost” of the order $O(n^{-1/2} \log^2 n)$. When $\beta_1 = \beta_2$ and $\gamma_1 = \gamma_2$ we obtain again a Gaussian local limit with convergence rate $O(n^{-1/2})$. Finally, these results are summarized in the last section, where we discuss possible future investigations.

¹ However, due to space constraints, all proofs in the present work are omitted.

2 Problem setting

Given the binary alphabet $\{a, b\}$, for every word $w \in \{a, b\}^*$ we denote by $|w|$ the length of w and by $|w|_a$ the number of occurrences of a in w . For each $n \in \mathbb{N}$, we also represent by $\{a, b\}^n$ the set $\{w \in \{a, b\}^* : |w| = n\}$. Here a *formal series* in the non-commutative variables a, b is a function $r : \{a, b\}^* \rightarrow \mathbb{R}_+$, where $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$, and for every $w \in \{a, b\}^*$ we denote by (r, w) the value of r at w . Such a series r is called *rational* if for some integer $m > 0$ there is a monoid morphism $\mu : \{a, b\}^* \rightarrow \mathbb{R}_+^{m \times m}$ and two arrays $\xi, \eta \in \mathbb{R}_+^m$, such that $(r, w) = \xi' \mu(w) \eta$, for every $w \in \{a, b\}^*$. In this case, as the morphism μ is generated by matrices $A = \mu(a)$ and $B = \mu(b)$, we say that the 4-tuple (ξ, A, B, η) is a *linear representation* of r of size m . Clearly, such a 4-tuple can be considered as a finite state automaton over the alphabet $\{a, b\}$, with transitions (as well as initial and final states) weighted by positive real values. Throughout this work we assume that the set $\{w \in \{a, b\}^n : (r, w) > 0\}$ is not empty for every $n \in \mathbb{N}_+$ (so that $\xi \neq 0 \neq \eta$), and that A and B are not null matrices, i.e. $A \neq [0] \neq B$. Then we can consider the probability measure \Pr over the set $\{a, b\}^n$ given by

$$\Pr(w) = \frac{(r, w)}{\sum_{x \in \{a, b\}^n} (r, x)} = \frac{\xi' \mu(w) \eta}{\xi' (A + B)^n \eta} \quad \forall w \in \{a, b\}^n$$

Note that, if r is the characteristic series of a language $L \subseteq \{a, b\}^*$ then \Pr is the uniform probability function over the set $L \cap \{a, b\}^n$. Thus we can define the random variable (r.v. for short) $Y_n = |w|_a$, where w is chosen at random in $\{a, b\}^n$ with probability $\Pr(w)$. As $A \neq [0] \neq B$, Y_n is not a degenerate r.v.. It is clear that, for every $k \in \{0, 1, \dots, n\}$,

$$p_n(k) := \Pr(Y_n = k) = \frac{\sum_{|w|=n, |w|_a=k} (r, w)}{\sum_{w \in \{a, b\}^n} (r, w)}$$

Since r is rational also the previous probability can be expressed by using its linear representation. It turns out that

$$p_n(k) = \frac{[x^k] \xi' (Ax + B)^n \eta}{\xi' (A + B)^n \eta} \quad \forall k \in \{0, 1, \dots, n\} \quad (1)$$

For sake of brevity we say that Y_n is *defined* by the linear representation (ξ, A, B, η) . The distribution of Y_n can be represented by the map $h_n(z)$ and the characteristic function $\Psi_n(t)$, given respectively by

$$h_n(z) = \xi' (Ae^z + B)^n \eta \quad \forall z \in \mathbb{C} \quad (2)$$

$$\Psi_n(t) = \sum_{k=0}^n p_n(k) e^{itk} = \frac{\xi' (Ae^{it} + B)^n \eta}{\xi' (A + B)^n \eta} = \frac{h_n(it)}{h_n(0)} \quad \forall t \in \mathbb{R} \quad (3)$$

In particular mean value and variance of Y_n are determined by

$$\mathbb{E}(Y_n) = \frac{h'_n(0)}{h_n(0)}, \quad \text{Var}(Y_n) = \frac{h''_n(0)}{h_n(0)} - \left(\frac{h'_n(0)}{h_n(0)} \right)^2 \quad (4)$$

Our general goal is to study the limit distribution of $\{Y_n\}$ as n grows to $+\infty$ and in particular its possible local limit law.

We recall that a sequence of r.v.'s $\{X_n\}$ *converges in distribution* (or in law) to a random variable X of distribution function F if $\lim_{n \rightarrow +\infty} \Pr(X_n \leq x) = F(x)$, for every $x \in \mathbb{R}$ of continuity for F . The central limit theorems yield classical examples of convergence in distribution to a Gaussian random variable.

Instead, the local limit laws establish the convergence of single probabilities to a density function (see for instance [9,8]). More precisely, consider a sequence of r.v.'s $\{X_n\}$ such that each X_n takes value in $\{0, 1, \dots, n\}$. We say that $\{X_n\}$ *satisfies a local limit law* of Gaussian type if there are two real sequences $\{a_n\}$, $\{s_n\}$, satisfying $a_n \sim E(X_n)$, $s_n^2 \sim \text{Var}(X_n)$ and $s_n > 0$ for all n , such that for some real $\epsilon_n \rightarrow 0$, the relation

$$\left| s_n \Pr(X_n = k) - \frac{e^{-\left(\frac{k-a_n}{s_n}\right)^2/2}}{\sqrt{2\pi}} \right| \leq \epsilon_n \quad (5)$$

holds uniformly for every $k \in \{0, 1, \dots, n\}$ and every $n \in \mathbb{N}$ large enough. Here, ϵ_n yields the *convergence rate* (or the speed) of the law. A well-known example of such a property is given by the de Moivre-Laplace local limit theorem, which concerns sequences of binomial r.v.'s [9].

Similar definitions can be given for other (non-Gaussian) types of local limit laws. In this case the Gaussian density $e^{-x^2/2}/\sqrt{2\pi}$ appearing in (5) is replaced by some density function $f(x)$; clearly, if $f(x)$ is not continuous at some points, the uniformity of k must be adapted to the specific case.

3 Primitive models

A relevant case occurs when $M = A + B$ is primitive, i.e. $M^k > 0$ for some $k \in \mathbb{N}$ [16]. In this case it is known that Y_n has a Gaussian limit distribution and satisfies a local limit property [13,3]. Here we improve this result, showing a convergence rate $O(n^{-1/2})$, and revisit some material appearing in [3,4] that is useful in the following sections.

Since M is primitive, by Perron-Frobenius Theorem, it admits a real eigenvalue $\lambda > 0$ greater than the modulus of any other eigenvalue. Thus, we can consider the function $u = u(z)$ implicitly defined by the equation

$$\text{Det}(Iu - Ae^z - B) = 0$$

such that $u(0) = \lambda$. It turns out that, in a neighbourhood of $z = 0$, $u(z)$ is analytic, is a simple root of the characteristic polynomial of $Ae^z + B$ and $|u(z)|$ is strictly greater than the modulus of all other eigenvalues of $Ae^z + B$. Moreover, a precise relationship between $u(z)$ and function $h(z)$, defined in (2), is proved in [3] stating that there are two positive constants c, ρ and a function $r(z)$ analytic and non-null at $z = 0$, such that

$$h_n(z) = r(z) u(z)^n + O(\rho^n) \quad \forall z \in \mathbb{C} : |z| \leq c \quad (6)$$

where $\rho < |u(z)|$ and in particular $\rho < \lambda$.

Mean value and variance of Y_n can be estimated from relations (6) and (4). It turns out [3] that the constants

$$\beta = \frac{u'(0)}{\lambda} \quad \text{and} \quad \gamma = \frac{u''(0)}{\lambda} - \left(\frac{u'(0)}{\lambda} \right)^2 \quad (7)$$

are strictly positive and satisfy the relations

$$\mathbb{E}(Y_n) = \beta n + O(1) \quad \text{and} \quad \text{Var}(Y_n) = \gamma n + O(1)$$

Other properties concern function $y(t) = u(it)/\lambda$, defined for real t in a neighbourhood of 0. In particular, there exists a constant $c > 0$, for which relation (6) holds true, satisfying the following relations [3]:

$$|y(t)| = 1 - \frac{\gamma}{2}t^2 + O(t^4), \quad \arg y(t) = \beta t + O(t^3), \quad |y(t)| \leq e^{-\frac{\gamma}{4}t^2} \quad \forall |t| \leq c \quad (8)$$

The behaviour of $y(t)$ can be estimated precisely when t tends to 0. For any q such that $1/3 < q < 1/2$ it can be proved [3] that

$$y(t)^n = e^{-\frac{\gamma}{2}t^2 n + i\beta t n} (1 + O(t^3)n) \quad \text{for } |t| \leq n^{-q} \quad (9)$$

The previous properties can be used to prove a local limit theorem for $\{Y_n\}$ when M is primitive, with a convergence rate $O(n^{-1/2})$. The result, stated in Theorem 1 below, holds under a further assumption, introduced to avoid periodicity phenomena. To state this condition properly, consider the transition graph of the finite state automaton defined by matrices A and B , i.e. the directed graph G with vertex set $\{1, 2, \dots, m\}$ such that, for every $i, j \in \{1, 2, \dots, m\}$, G has an edge from i to j labelled by a letter a (b , respectively) whenever $A_{ij} > 0$ ($B_{ij} > 0$, resp.). Also denote by d the GCD of the differences in the number of occurrences of a in the (labels of) cycles of equal length of G . Here and in the sequel we say that the pair (A, B) is *aperiodic* if $d = 1$. Such a property is often verified; for instance it holds true whenever $A_{ij} > 0$ and $B_{ij} > 0$ for two (possibly equal) indices i, j .

Theorem 1. *Let $\{Y_n\}$ be defined by a linear representation (ξ, A, B, η) such that $M = A + B$ is primitive, $A \neq [0] \neq B$ and (A, B) is aperiodic. Moreover, let β and γ be defined by equalities (7). Then, as n tends to $+\infty$, the relation*

$$\left| \sqrt{n} \Pr(Y_n = k) - \frac{e^{-\frac{(k-\beta n)^2}{2\gamma n}}}{\sqrt{2\pi\gamma}} \right| = O(n^{-1/2}) \quad (10)$$

holds true uniformly for every $k \in \{0, 1, \dots, n\}$.

4 Bicomponent models

In this section we study the behaviour of $\{Y_n\}_{n \in \mathbb{N}}$ defined by a linear representation (ξ, A, B, η) of size m , such that the matrix $M = A + B$ consists of two irreducible components. Formally, there are two linear representations, $(\xi_1, A_1, B_1, \eta_1)$ and $(\xi_2, A_2, B_2, \eta_2)$, of size m_1 and m_2 respectively, where $m = m_1 + m_2$, such that :

1) for some $A_0, B_0 \in \mathbb{R}_+^{m_1 \times m_2}$ we have

$$\xi' = (\xi'_1, \xi'_2), \quad A = \begin{pmatrix} A_1 & A_0 \\ 0 & A_2 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & B_0 \\ 0 & B_2 \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \quad (11)$$

2) $M_1 = A_1 + B_1$ and $M_2 = A_2 + B_2$ are irreducible matrices and we denote by λ_1 and λ_2 the corresponding Perron-Frobenius eigenvalues;

3) $\xi_1 \neq 0 \neq \eta_2$ and matrix $M_0 = A_0 + B_0$ is different from $[0]$.

Note that condition 2) is weaker than a primitivity hypothesis for M_1 and M_2 . Condition 3) assures that there is communication from the first to the second component and hence the main term of the probability function of Y_n also depends on the convolution of their behaviours.

Assuming these hypotheses the limit properties of $\{Y_n\}$ first depend on whether $\lambda_1 \neq \lambda_2$ or $\lambda_1 = \lambda_2$. In the first case there is a dominant component, corresponding to the maximum between λ_1 and λ_2 , which determines the asymptotic behaviour of $\{Y_n\}$. In the second case the two components are equipotent and they both contribute to the limit behaviour of $\{Y_n\}$. In both cases the corresponding characteristic function has some common properties.

For $j = 1, 2$, let us define $h_n^{(j)}(z)$, $u_j(z)$, $y_j(t)$, β_j , and γ_j , respectively, as the values $h_n(z)$, $u(z)$, $y(t)$, β , γ referred to component j . We also define $H(x, y)$ as the matrix-valued function given by

$$H(x, y) = \sum_{n=0}^{+\infty} (Ax + B)^n y^n = \begin{bmatrix} H^{(1)}(x, y) & G(x, y) \\ 0 & H^{(2)}(x, y) \end{bmatrix}, \quad \text{where} \\ H^{(1)}(x, y) = \frac{\text{Adj}(I - (A_1x + B_1)y)}{\text{Det}(I - (A_1x + B_1)y)}, \quad H^{(2)}(x, y) = \frac{\text{Adj}(I - (A_2x + B_2)y)}{\text{Det}(I - (A_2x + B_2)y)}, \quad (12) \\ G(x, y) = H^{(1)}(x, y) (A_0x + B_0)y H^{(2)}(x, y).$$

Thus, the generating function of $\{h_n(z)\}_n$ satisfies the following identities

$$\sum_{n=0}^{\infty} h_n(z) y^n = \xi' H(e^z, y) \eta = \xi'_1 H^{(1)}(e^z, y) \eta_1 + \xi'_1 G(e^z, y) \eta_2 + \xi'_2 H^{(2)}(e^z, y) \eta_2 \quad (13)$$

Hence, setting $g_n(z) = [y^n] \xi'_1 G(e^z, y) \eta_2$, we obtain

$$h_n(z) = h_n^{(1)}(z) + g_n(z) + h_n^{(2)}(z) \quad (14)$$

to be used in the analysis of the characteristic function $\Psi_n(it)$ given by (3).

The dominant case is similar to the primitive one. Assume that $\lambda_1 > \lambda_2$, M_1 is aperiodic (and hence primitive) and $A_1 \neq [0] \neq B_1$. For sake of brevity, we

say that $\{Y_n\}$ is defined in a *dominant bicomponent model* with $\lambda_1 > \lambda_2$. In this case we have $0 < \beta_1 < 1$, $0 < \gamma_1$, and it is known that $\frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}}$ converges in distribution to a normal r.v. of mean value 0 and variance 1 [7]. Moreover, one can prove the following result:

Theorem 2. *Let $\{Y_n\}$ be defined in a dominant bicomponent model with $\lambda_1 > \lambda_2$ and assume (A_1, B_1) aperiodic. Then, as n tends to $+\infty$, the relation*

$$\left| \sqrt{n} \Pr(Y_n = k) - \frac{e^{-\frac{(k - \beta_1 n)^2}{2\gamma_1 n}}}{\sqrt{2\pi\gamma_1}} \right| = O(n^{-1/2})$$

holds true uniformly for every $k \in \{0, 1, \dots, n\}$.

4.1 Equipotent case

Now, let us assume that $\lambda_1 = \lambda_2 = \lambda$, both matrices M_1 and M_2 are aperiodic (and hence primitive) and $A_j \neq [0] \neq B_j$ for $j = 1, 2$. Under these hypotheses we say that $\{Y_n\}$ is defined in an *equipotent bicomponent model*. In this case the limit distribution of $\{Y_n\}$ depends on the parameters $\beta_1, \beta_2, \gamma_1, \gamma_2$, defined as in (7), which now satisfy conditions $0 < \beta_j < 1$ and $0 < \gamma_j$, for both $j = 1, 2$. Before studying the different cases, we recall some properties presented in [7] that are useful in our context.

Observe that both $h_n^{(1)}(z)$ and $h_n^{(2)}(z)$ satisfy relation (6). Moreover, from relations (12) and an analysis of function $\xi_1^t G(e^z, y) \eta_2$, for some $c > 0$ it can be shown that

$$g_n(z) = s(z) \sum_{j=0}^{n-1} u_1(z)^j u_2(z)^{n-1-j} + O(\rho^n) \quad \forall z \in \mathbb{C} : |z| \leq c \quad (15)$$

where $s(z)$ is an analytic and non-null function for $|z| \leq c$, and $\rho < \max\{|u_1(z)|, |u_2(z)|\}$. Therefore, by equality (14) we obtain

$$h_n(z) = s(z) \sum_{j=0}^{n-1} u_1(z)^j u_2(z)^{n-1-j} + O(u_1(z)^n) + O(u_2(z)^n) \quad \forall z \in \mathbb{C} : |z| \leq c \quad (16)$$

This relation has two consequences. First, since $u_1(0) = \lambda = u_2(0)$, it implies

$$h_n(0) = s(0) n \lambda^{n-1} (1 + O(1/n)) \quad (s(0) \neq 0) \quad (17)$$

Second, if $u_1(z) \neq u_2(z)$ for some $z \in \mathbb{C}$ satisfying $0 < |z| \leq c$, then one gets

$$h_n(z) = s(z) \frac{u_1(z)^n - u_2(z)^n}{u_1(z) - u_2(z)} + O(u_1(z)^n) + O(u_2(z)^n) \quad (18)$$

Finally, in the equipotent bicomponent models the aperiodicity condition consists of requiring that both pairs (A_1, B_1) and (A_2, B_2) are aperiodic. Under this hypothesis, the following property holds true.

Proposition 3. *Let $\{Y_n\}$ be defined in an equipotent bicomponent model and let both pairs (A_1, B_1) and (A_2, B_2) be aperiodic. Then, for every $c \in (0, \pi)$ there exists $\varepsilon \in (0, 1)$ such that $|\Psi_n(t)| = O(\varepsilon^n)$ for all $t \in \mathbb{R}$ satisfying $c \leq |t| \leq \pi$.*

4.1.1 Local limit with different β 's. In this subsection we assume an equipotent bicomponent model with $\beta_1 \neq \beta_2$. In this case it is known that $\{Y_n/n\}$ converges in distribution to a uniform r.v. [7]. Here we state a local limit theorem with a speed of convergence of an order arbitrarily slower than $O(n^{-1/2}(\log n)^{3/2})$, thus improving a recent result presented in [10]. To this end, in view of Proposition 3, we study the characteristic function $\Psi_n(t)$ for $|t| \leq c$, where $c \in (0, \pi)$ is a constant satisfying relation (16). Recall that in such a set both functions $y_1(t) = u_1(it)/\lambda$ and $y_2(t) = u_2(it)/\lambda$ satisfy relations (8), and hence for every real t such that $|t| \leq c$, we have

$$y_1(t) = 1 + i\beta_1 t + O(t^2), \quad y_2(t) = 1 + i\beta_2 t + O(t^2) \quad (19)$$

$$|y_1(t)| \leq e^{-\frac{\gamma_1}{4}t^2}, \quad |y_2(t)| \leq e^{-\frac{\gamma_2}{4}t^2} \quad (20)$$

Moreover, since in the present case (18) holds true for z near 0, using the previous relations, for a suitable $c \in (0, \pi)$ and every $t \in \mathbb{R}$ such that $0 < |t| \leq c$, we obtain

$$\Psi_n(t) = \frac{h_n(it)}{h_n(0)} = \frac{1 + O(t)}{1 + O(1/n)} \left(\frac{y_1(t)^n - y_2(t)^n}{i(\beta_1 - \beta_2)tn} \right) + \sum_{j=1,2} O\left(\frac{y_j(t)^n}{n}\right) \quad (21)$$

Now, for such a constant c , let us split $[-c, c]$ into sets S_n and V_n , given by

$$S_n = \left\{ t \in \mathbb{R} : |t| \leq \sqrt{\frac{\log n}{n}} \tau_n^{1/3} \right\}, \quad V_n = \left\{ t \in \mathbb{R} : \sqrt{\frac{\log n}{n}} \tau_n^{1/3} < |t| \leq c \right\} \quad (22)$$

where $\{\tau_n\} \subset \mathbb{R}_+$ is any sequence such that $\tau_n \rightarrow +\infty$ and $\tau_n = o(\log \log n)$ (i.e. τ_n tends to $+\infty$ with an arbitrarily slow order of growth). The behaviour of $\Psi_n(t)$ in these sets is given by the following two propositions, where we assume an equipotent bicomponent model with $\beta_1 \neq \beta_2$.

Proposition 4. *For some $a > 0$ one has $|\Psi_n(t)| = o\left(n^{-a\tau_n^{2/3}}\right)$ for all $t \in V_n$.*

In order to evaluate $\Psi_n(t)$ for $t \in S_n$, let us define

$$K_n(t) = \frac{e^{-\frac{\gamma_1}{2}t^2n + i\beta_1tn} - e^{-\frac{\gamma_2}{2}t^2n + i\beta_2tn}}{i(\beta_1 - \beta_2)tn} \quad (23)$$

and consider relation (21). Since for $t \in S_n$ one has $nO(t^3) = o(1)$, relation (9) applies to both $y_1(t)$ and $y_2(t)$ yielding

$$y_j(t)^n = e^{-\frac{\gamma_j}{2}t^2n + i\beta_jtn} (1 + nO(t^3)) \quad \forall t \in S_n, \quad j = 1, 2$$

Replacing these values in (21), after some computation one gets

$$\Psi_n(t) = [1 + O(t) + nO(t^3) + O(1/n)] K_n(t) + O(1/n) \quad \forall t \in S_n \quad (24)$$

Proposition 5. *Defining S_n and $K_n(t)$ as in (22) and (23), we have*

$$\left| \int_{S_n} (\Psi_n(t) - K_n(t)) dt \right| = O \left(\left(\frac{\log n}{n} \right)^{3/2} \tau_n \right)$$

Now, we are able to state the local limit in the present case. Set $b_1 = \min\{\beta_1, \beta_2\}$, $b_2 = \max\{\beta_1, \beta_2\}$ and denote by $f_U(x)$ the density function of a uniform r.v. U in the interval $[b_1, b_2]$, that is

$$f_U(x) = \frac{1}{b_2 - b_1} \chi_{[b_1, b_2]}(x) \quad \forall x \in \mathbb{R}$$

where χ_I denotes the indicator function of the interval $I \subset \mathbb{R}$. Then we have

Theorem 6. *Let $\{Y_n\}_{n \in \mathbb{N}}$ be defined in an equipotent bicomponent model with $\beta_1 \neq \beta_2$ and assume aperiodic both pairs (A_1, B_1) and (A_2, B_2) . Then, for n tending to $+\infty$, the r.v. Y_n satisfies the relation*

$$|n \Pr(Y_n = k) - f_U(x)| = O \left(\frac{(\log n)^{3/2} \tau_n}{\sqrt{n}} \right) \quad (25)$$

for every real sequence $\{\tau_n\}$ satisfying $\tau_n \rightarrow +\infty$, $\tau_n = o(\log \log n)$ and for every integer $k = k(n)$, provided that $k/n \rightarrow x$ for a constant x such that $\beta_1 \neq x \neq \beta_2$.

As an example, consider the rational stochastic model defined by the weighted finite automaton of Figure 1, where each transition is labelled by an alphabet symbol and a weight, together with the arrays $\xi = (1, 0, 0, 0)$ and $\eta = (0, 0, 1, 1)$. Such an automaton recognizes the set of all words $w \in \{a, b, c\}^*$ of the form $w = xcy$, such that $x, y \in \{a, b\}^*$ and the strings aa and bb do not occur in x and y , respectively. Clearly this is a bicomponent model, with both pairs (A_1, B_1) and (A_2, B_2) aperiodic. Moreover $M_1 = M_2$, while $A_1 \neq A_2$. Hence the two components are equipotent and $\beta_1 \neq \beta_2$. This means that Y_n/n converges in distribution to a uniform r.v. of extremes β_1, β_2 , and Y_n satisfies Theorem 6. Note that simple changes may modifies the limit distribution: for instance, setting to 3 the weight of transition $2 \xrightarrow{b} 1$ makes dominant the first component, implying a Gaussian local limit law (Theorem 2).

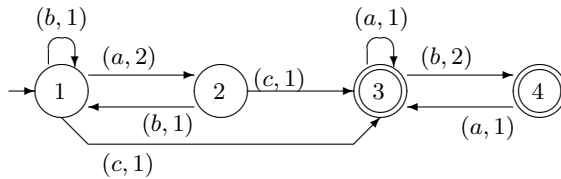


Fig. 1. Weighted finite automaton defining an equipotent bicomponent model ($\lambda_1 = \lambda_2 = 2$) with $1/3 = \beta_1 \neq \beta_2 = 2/3$.

4.1.2 Local limit with equal β 's and different γ 's. In this section we present a local limit theorem for $\{Y_n\}$ defined in an equipotent bicomponent model with $\beta_1 = \beta_2$ and $\gamma_1 \neq \gamma_2$. In this case, setting $\beta = \beta_1 = \beta_2$ and $\gamma = \frac{\gamma_1 + \gamma_2}{2}$, it is known [7] that $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ weakly converges to a random variable T whose distribution is a mixture of Gaussian laws of mean 0 and variance uniformly distributed over the interval of extremes $\frac{\gamma_1}{\gamma}$ and $\frac{\gamma_2}{\gamma}$.

Formally, the density function of T is given by

$$f_T(x) = \frac{\gamma}{\gamma_2 - \gamma_1} \int_{\frac{\gamma_1}{\gamma}}^{\frac{\gamma_2}{\gamma}} \frac{e^{-\frac{x^2}{2s}}}{\sqrt{2\pi s}} ds \quad \forall x \in \mathbb{R} \quad (26)$$

In passing, we observe that, for each $x \in \mathbb{R}$, $f_T(x)$ may be regarded as the mean value of the “heat kernel” $K(x, t) = (4\pi t)^{-1/2} e^{-\frac{x^2}{4t}}$ at point x in the time interval of extremes $\gamma_1/(2\gamma)$ and $\gamma_2/(2\gamma)$ [6].

Note that $E(T) = 0$ and $\text{Var}(T) = 1$, while its characteristic function is

$$\Phi_T(t) = \int_{-\infty}^{+\infty} f_T(x) e^{itx} dx = 2\gamma \frac{e^{-\frac{\gamma_1}{2\gamma} t^2} - e^{-\frac{\gamma_2}{2\gamma} t^2}}{(\gamma_2 - \gamma_1) t^2} \quad (27)$$

Then, $f_T(x)$ can be expressed in the form

$$f_T(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi_T(t) e^{-itx} dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} 2\gamma \frac{e^{-\frac{\gamma_1}{2\gamma} t^2} - e^{-\frac{\gamma_2}{2\gamma} t^2}}{(\gamma_2 - \gamma_1) t^2} e^{-itx} dt$$

Our goal is to present a local limit property for $\{Y_n\}$ (suitably scaled) toward the r.v. T , with a speed of convergence of an order arbitrarily slower than $O\left(\frac{\log^2 n}{\sqrt{n}}\right)$.

Also in this case we assume aperiodic both pairs (A_1, B_1) and (A_2, B_2) , which implies Proposition 3. As in the previous section, $c \in (0, \pi)$ is a constant for which relation (16) holds true; as a consequence, both functions $y_1(t)$ and $y_2(t)$ satisfy relations (8), which now can be refined in the following form:

$$y_j(t) = \frac{u_j(it)}{\lambda} = 1 + i\beta t - \frac{\gamma_j + \beta^2}{2} t^2 + O(t^3), \quad \forall x \in \mathbb{R} : |t| \leq c, j = 1, 2$$

Applying these values in (18), which is valid also in the present case for z near to 0, and using (17), for some $c \in (0, \pi)$ and every $t \in \mathbb{R}$ such that $0 < |t| \leq c$, we obtain

$$\Psi_n(t) = \frac{h_n(it)}{h_n(0)} = 2 \frac{1 + O(t)}{n + O(1)} \frac{y_1(t)^n - y_2(t)^n}{(\gamma_2 - \gamma_1) t^2 + O(t^3)} + \sum_{j=1,2} O\left(\frac{y_j(t)^n}{n}\right) \quad (28)$$

Now, for such a constant c , split the interval $[-c, c]$ into sets S_n and V_n given by

$$S_n = \left\{ t \in \mathbb{R} : |t| \leq \sqrt{\frac{\log n}{n}} \tau_n^{1/4} \right\}, \quad V_n = \left\{ t \in \mathbb{R} : \sqrt{\frac{\log n}{n}} \tau_n^{1/4} < |t| \leq c \right\} \quad (29)$$

where τ_n is defined as in (22). The behaviour of $\Psi_n(t)$ in these sets is described by the propositions below where we assume an equipotent bicomponent model with $\beta_1 = \beta_2 = \beta$ and $\gamma_1 \neq \gamma_2$.

Proposition 7. *For some $a > 0$ we have $|\Psi_n(t)| = o\left(n^{-a\tau_n^{1/2}}\right)$ for every $t \in V_n$.*

For sake of brevity, we define

$$K_n(t) = 2 \frac{e^{-\frac{\gamma_1}{2}t^2n} - e^{-\frac{\gamma_2}{2}t^2n}}{(\gamma_2 - \gamma_1)t^2n} e^{i\beta tn}, \quad \forall t \in \mathbb{R} \quad (30)$$

It is easy to see that $|K_n(t)| \leq 2 \sum_{j=1,2} \left(\frac{1 - e^{-\frac{\gamma_j}{2}t^2n}}{|\gamma_2 - \gamma_1|t^2n} \right)$ for every $t \in \mathbb{R}$. A simple study of these expressions shows that both addends take their maximum value at $t = 0$, where they have a removable singularity, and such values are independent of n . As a consequence we can state that $|K_n(t)| \leq \frac{\gamma_1 + \gamma_2}{|\gamma_2 - \gamma_1|}$, for every $n \in \mathbb{N}_+$ and every $t \in S_n$.

Proposition 8. *Defining S_n and $K_n(t)$ by (29) and (30), respectively, we have*

$$\int_{S_n} |\Psi_n(t) - K_n(t)| dt = O\left(\frac{(\log n)^2 \tau_n}{n}\right)$$

Now we can state the local limit theorem in the present case:

Theorem 9. *Let $\{Y_n\}_{n \in \mathbb{N}}$ be defined in an equipotent bicomponent model with $\beta_1 = \beta_2 = \beta$, $\gamma_1 \neq \gamma_2$, assume aperiodic pairs (A_1, B_1) and (A_2, B_2) and set $\gamma = (\gamma_1 + \gamma_2)/2$. Then, for n tending to $+\infty$, Y_n satisfies the relation*

$$\left| \sqrt{\gamma n} \Pr(Y_n = k) - f_T\left(\frac{k - \beta n}{\sqrt{\gamma n}}\right) \right| = O\left(\frac{(\log n)^2 \tau_n}{\sqrt{n}}\right) \quad (31)$$

uniformly for $k \in \{0, 1, \dots, n\}$, where f_T is defined in (26) and $\{\tau_n\} \subset \mathbb{R}_+$ is any sequence such that $\tau_n \rightarrow +\infty$ and $\tau_n = o(\log \log n)$.

4.1.3 Local limit with equal β 's and equal γ 's. In this section we study the local limit properties of $\{Y_n\}$ assuming an equipotent bicomponent model with $\beta_1 = \beta_2 = \beta$ and $\gamma_1 = \gamma_2 = \gamma$. In this case, it is known [7] that $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ converges in distribution to a Gaussian random variable of mean 0 and variance 1. Here we prove that a Gaussian local limit property holds true with a convergence rate of the order $O(n^{-1/2})$, assuming aperiodic both pairs (A_1, B_1) and (A_2, B_2) .

Again we assume $c \in (0, \pi)$ is a constant for which equality (16) holds true, so that both functions $y_1(t)$ and $y_2(t)$ satisfy relations (8) and (9), which we now restate in the following form for sake of clearness:

$$|y_j(t)| \leq e^{-\frac{\gamma}{4}t^2} \quad \forall t \in \mathbb{R} : |t| \leq c, \quad j = 1, 2 \quad (32)$$

$$y_j(t)^n = e^{-\frac{\gamma}{2}t^2n + i\beta tn(1 + nO(t^3))} \quad \forall t \in \mathbb{R} : |t| \leq n^{-q}, \quad j = 1, 2 \quad (33)$$

where q is an arbitrary value such that $1/3 < q < 1/2$.

The following propositions yield properties of the characteristic function $\Psi_n(t)$ respectively for $|t| \leq n^{-q}$ and $n^{-q} < |t| \leq c$.

Proposition 10. *For every $q \in (1/3, 1/2)$, we have*

$$|\Psi_n(t)| = O\left(e^{-\frac{\gamma}{4}n^{1-2q}}\right) \quad \forall t \in \mathbb{R} : n^{-q} < |t| \leq c$$

Proposition 11. *For every $q \in (1/3, 1/2)$, we have*

$$\int_{|t| \leq n^{-q}} \left| \Psi_n(t) - e^{-\frac{\gamma}{2}t^2n + i\beta tn} \right| dt = O(n^{-1})$$

Then, our last result follows:

Theorem 12. *Let $\{Y_n\}_{n \in \mathbb{N}}$ be defined in an equipotent bicomponent model with $\beta_1 = \beta_2 = \beta$ and $\gamma_1 = \gamma_2 = \gamma$, and assume aperiodic both pairs (A_1, B_1) and (A_2, B_2) . Then, for n tending to $+\infty$ the relation*

$$\left| \sqrt{n} \Pr(Y_n = k) - \frac{e^{-\frac{(k-\beta n)^2}{2\gamma n}}}{\sqrt{2\pi\gamma}} \right| = O(n^{-1/2})$$

holds true uniformly for every $k \in \{0, 1, \dots, n\}$.

5 Conclusions

The analysis of the symbol statistics Y_n 's presented in this work concerns the cases when the rational stochastic model consists of one or two primitive components. Our results are summarized in Table 1, which refers to the previous literature for already known properties.

	Primitive Models	Bicomponent Models			
		dominant	equipotent		
			$\beta_1 \neq \beta_2$	$\beta_1 = \beta_2$ $\gamma_1 \neq \gamma_2$	$\beta_1 = \beta_2$ $\gamma_1 = \gamma_2$
Local limit distribution	$N_{0,1}$ (see [3])	$N_{0,1}$	U_{β_1, β_2} (see [10])	T	$N_{0,1}$
Convergence rate	$O(n^{-1/2})$	$O(n^{-1/2})$	$O\left(\frac{\tau_n \log^{3/2} n}{\sqrt{n}}\right)$	$O\left(\frac{\tau_n \log^2 n}{\sqrt{n}}\right)$	$O(n^{-1/2})$

Table 1. Symbols $N_{0,1}$, U_{β_1, β_2} and T denote respectively a Gaussian, uniform and T -type local limit, T being defined in Section 4.1.2. Also, τ_n is defined in Theorem 6

Natural extensions of these results concern rational models with more than two primitive components having equal dominant eigenvalue and, possibly, the evaluation of neglected terms in the asymptotic expressions. Also in the case

of bicomponent models our analysis is not complete as it does not include the non-communicating cases ($M_0 = [0]$) nor the degenerate cases (when, for a dominant component $i \in \{1, 2\}$, either $A_i = 0$ or $B_i = 0$). In these cases rather different limit distributions are obtained [7, Section 8], due to the diverse type of generating functions appearing therein. Even if these situations are somehow particular, they are representative of typical regular languages, and hence they seem to be natural subjects for future investigations.

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, V. Lonati. On the number of occurrences of a symbol in words of regular languages. *Theoret. Comput. Sci.*, 302:431–456, 2003.
4. A. Bertoni, C. Choffrut, M. Goldwurm, V. Lonati. Local limit properties for pattern statistics and rational models. *Theory Comput. Systems*, 39:209–235, 2006.
5. S. Broda, A. Machiavelo, N. Moreira, and R. Reis. A hitchhiker’s guide to descriptive complexity through analytic combinatorics. *Theoret. Comput. Sci.*, 528:85–100, 2014.
6. J.R. Cannon. *The one-dimensional Heat Equation*. Encyclopedia of Mathematics and its Applications, vol. 23, Addison–Wesley Publishing Company, 1984.
7. D. de Falco, M. Goldwurm, V. Lonati. Frequency of symbol occurrences in bicomponent stochastic models. *Theoret. Comput. Sci.*, 327 (3):269–300, 2004.
8. P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge Univ. Press, 2009.
9. B.V. Gnedenko. *Theory of probability*. Gordon and Breach Science Publ., 1997.
10. M. Goldwurm, J. Lin, M. Vignati. A local limit property for pattern statistics in bicomponent stochastic models. Proc. 20th DCFS, LNCS vol. 10952, 114–125, 2018.
11. P. Grabner, M. Rigo. Distribution of additive functions with respect to numeration systems on regular languages. *Theory Comput. Systems*, 40:205–223, 2007.
12. H.-K. Hwang. On convergence rates in the Central Limit Theorem for combinatorial structures. *Europ. J. Combinatorics*, 19:329–343, 1998.
13. P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. *Theoret. Comput. Sci.*, 287(2): 593–617, 2002.
14. M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.
15. A. Salomaa and M. Soittola. *Automata-Theoretic Aspects of Formal Power Series*. Springer-Verlag, 1978.
16. E. Seneta. *Non-negative matrices and Markov chains*. Springer-Verlag, New York Heidelberg Berlin, 1981.