A Local Limit Property for Pattern Statistics in Bicomponent Stochastic Models

Massimiliano Goldwurm, Jianyi Lin, Marco Vignati

A local limit property for pattern statistics in bicomponent stochastic models

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

(1) Dipartimento di Matematica, Università degli Studi di Milano,
Milano – Italy
(2) Department of Mathematics, Khalifa University,
Abu Dhabi - United Arab Emirates

Abstract. We present a non-Gaussian local limit theorem for the number of occurrences of a given symbol in a word of length n generated at random. The stochastic model for the random generation is defined by a rational formal series with non-negative real coefficients. The result yields a local limit towards a uniform density function and holds under the assumption that the formal series defining the model is recognized by a weighted finite state automaton with two primitive components having equal dominant eigenvalue.

1 Introduction

A classical counting problem concerning formal languages is the evaluation of the number of occurrences of a given symbol a in a word w of length n generated at random by a suitable stochastic source. Denoting by Y_n such a random variable, traditional goals of interest are the asymptotic evaluation of mean value and variance of Y_n as well as its limit distribution and local limit evaluations for its probability function. Clearly these properties depend on the stochastic model used for generating the random text w. Classical models are the Bernoulli and Markovian ones [13,11]. Here we consider the rational stochastic models, which are defined by rational formal series in non-commutative variables with coefficients in $\mathbb{R}_+$ [3]. In these models the computation of a random word w can be done easily in linear time [8] once we know a $\mathbb{R}_+$-weighted finite state automaton that recognizes the series. Such a probabilistic source is rather general since it includes as special cases the traditional Bernoulli and Markovian models and also encompasses the random generation of words of length n in any regular language under uniform distribution.

The problem above is of interest for several reasons. First, it has been studied in connection with the analysis of pattern statistics and in particular those occurring in computational biology [12,13,11,3]. It turns out that evaluating the frequency of patterns from a regular expression in a random text generated by a Markovian model can be reduced to determining the frequency of a single symbol in a word over a binary alphabet generated by a rational stochastic model [11,3].

Moreover, it is well-known that the average number of occurrences of symbols in words of regular and context-free languages plays a relevant role in the analysis of the descriptional complexity of languages and computational models [5,6]. Clearly the limit distributions of these quantities (also in the local form) yield a more complete information and in particular they allow to evaluate their dispersion around the average values.

Our problem is also related to the asymptotic estimate of the coefficients of rational formal series in commutative variables. In particular the local limit properties of Y_n enabled to show that the maximum coefficients of monomials of size n for some of those series is of the order $\Theta(n^{k-3/2}\lambda^n)$ for $\lambda \geq 1$ and any positive $k \in \mathbb{N}$ [4, Corollary 16]. Similar consequences hold for the study of the degree of ambiguity of rational trace languages (subset of free partially commutative monoids) [3].

The asymptotic behaviour of Y_n assuming the rational stochastic model defined by a series r depends on the finite state automata that recognize r. The main results known in the literature concern the case when such an automaton has a primitive transition matrix. In this case asymptotic expressions for the mean value $E(Y_n)$ and the variance $var(Y_n)$ are known. In particular, there is a real value β, $0 < \beta < 1$, such that $E(Y_n) = \beta n + O(1)$ and a similar result holds for $var(Y_n)$. Under the same hypothesis it is also proved that Y_n has a Gaussian limit distribution (i.e. it satisfies a central limit theorem) and it admits local limit properties intuitively stating that its probability function approximates a normal density [3,4].

The properties of Y_n have been studied also when the transition matrix consists of two primitive components. A variety of results on the asymptotic behaviour of Y_n are obtained in this case [7], but none of them concerns local limit properties. Here, we extend the previous results by showing a non-Gaussian local limit theorem that holds assuming that the two components have equal dominant eigenvalue while the main constants of the average value, β_1 and β_2 (associated with the first and the second component, respectively) are different. Under these hypotheses, it is known that Y_n/n converges in distribution to a random variable U uniformly distributed over the interval $[\min\{\beta_1, \beta_2\}, \max\{\beta_1, \beta_2\}]$ [7]. In the present work, assuming a further natural aperiodicity condition on the transition matrix, we prove that, as n grows to $+\infty$ and for any integer expression $k = k(n)$ such that k/n converges to a value x different from β_1 and β_2, we have $n\Pr(Y_n = k) = f_U(x) + o(1)$, where f_U is the density function of the uniform random variable U defined above.

The proof of our result is based on the analysis of the characteristic function of Y_n and it is obtained by adapting to our settings the so-called Saddle Point Method, traditionally used for proving Gaussian local limit properties [9].

The material we present is organized as follows. In Section 2 we recall the rational stochastic model and other preliminary notions. In Section 3 we revisit the properties of Y_n when the transition matrix of the automaton is primitive (Gaussian case). In Section 4 we introduce the bicomponent model and prove

the main result comparing it with the convergence in distribution given in [7]. In the last section we discuss possible extensions and future work.

2 Preliminary notions

In this section we give some preliminary notions and define our problem.

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 $\{w \in \{a, b\}^* : |w| = n\}$. A *formal series* in the non-commutative variables a, b with coefficients in the set $\mathbb{R}_+$ of non-negative real numbers is a function $r : \{a, b\}^* \to \mathbb{R}_+$, usually represented in the form $r = \sum_{w \in \{a,b\}^*} (r, w)w$, where each coefficient (r, w) is the value of r at w. The set $\mathbb{R}_+\langle\langle a, b\rangle\rangle$ of all such formal series forms a semiring with respect to the operations of sum and Cauchy product. A series $r \in \mathbb{R}_+\langle\langle a, b\rangle\rangle$ is called *rational* if for some integer $m > 0$ there is a monoid morphism $\mu : \{a, b\}^* \to \mathbb{R}_+^{m\times m}$ 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.

Now, consider a rational formal series $r \in \mathbb{R}_+\langle\langle a, b\rangle\rangle$ with linear representation (ξ, A, B, η) and let μ be the morphism generated by A and B. Assume that the set $\{w \in \{a, b\}^n : (r, w) > 0\}$ is not empty for every positive integer n. Then we can consider the probability measure Pr over the set $\{a, b\}^n$ given by

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

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 $L \cap \{a, b\}^n$. Moreover, it is easy to see that any Bernoullian or Markovian source, for the random generation of words in $\{a, b\}^*$, produces strings in $\{a, b\}^n$ with probability Pr for some rational series $r \in \mathbb{R}_+\langle\langle a, b\rangle\rangle$. We also recall that there are linear time algorithms that on input n generate a random word in $\{a, b\}^n$ according with probability Pr [8].

Then we can define the random variable (r.v.) Y_n representing the number of occurrences of the symbol a in a word w chosen at random in $\{a, b\}^n$ with probability $\Pr(w)$. In this work we are interested in the asymptotic properties of $\{Y_n\}$. It is clear that, for every $k \in \{0, 1, \ldots, n\}$,

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

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

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

where $[x^k]q(x)$ denotes the coefficient of x^k in a polynomial $q \in \mathbb{R}[x]$. For sake of brevity we say that Y_n is *defined* by the linear representation (ξ, A, B, η). Then the distribution of each Y_n can be characterized by function $h_n(z)$ given by

$$h_n(z) = \xi'(Ae^z + B)^n \eta$$

Indeed, setting $M = A + B$, the moment generating function of Y_n is given by

$$F_n(z) = \sum_{k=0}^{n} p_n(k) e^{zk} = \frac{\xi'(Ae^z + B)^n \eta}{\xi' M^n \eta} = \frac{h_n(z)}{h_n(0)} \qquad \forall\ z \in \mathbb{C}$$

This leads to determine mean value and variance of Y_n as

$$E(Y_n) = F_n'(0) = \frac{h_n'(0)}{h_n(0)}, \quad Var(Y_n) = \frac{h_n''(0)}{h_n(0)} - \left(\frac{h_n'(0)}{h_n(0)}\right)^2 \tag{2}$$

Analogously, the characteristic function of Y_n is given by

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

It turns out that the limit distribution of Y_n depends on the properties of the matrix $M = A + B$. A relevant case occurs when M is primitive (i.e. $\exists k \in \mathbb{N} : M^k > 0$). In this case it is known that Y_n has a Gaussian limit distribution [1,3] and satisfies a local limit theorem (in the sense of De Moivre - Laplace Theorem [10]) we recall in the next section.

3 Primitive case

In this section we consider the case when $M = A + B$ is a primitive matrix and recall some properties proved in [3,4,11] that are useful in the sequel.

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

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

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 $Ae^z + B$.

Now, consider the bivariate matrix-valued function $H(x, y)$ given by

$$H(x, y) = \sum_{n=0}^{+\infty} (Ax + B)^n\ y^n \ = \ (I - (Ax + B)y)^{-1}$$

Clearly, $\xi' H(e^z, y)\eta$ is the generating function of $\{h_n(z)\}_n$, i.e.

$$\xi' H(e^z, y)\eta = \sum_{n=0}^{+\infty} h_n(z) y^n = \frac{\xi' \mathrm{Adj}\,(I - (Ae^z + B)y)\,\eta}{\mathrm{Det}\,(I - (Ae^z + B)y)}$$

Thus, for every z near 0, the singularities of $\xi' H(e^z, y)\eta$ are the values μ^{-1} for all (non-null) eigenvalues μ of $Ae^z + B$ and hence $u(z)^{-1}$ is its (unique) singularity of minimum modulus. Then, by the properties of $u(z)$ one can get the following

Proposition 1. *[3] If M is primitive then there are two positive constants c, ρ and a function $r(z)$ analytic and non-null at $z = 0$, such that for every $|z| \leq c$*

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

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

Mean value and variance of Y_n can be estimated from equations (2). In turns out that the constants $\beta = u'(0)/\lambda$ and $\gamma = \frac{u''(0)}{\lambda} - \left(\frac{u'(0)}{\lambda}\right)^2$ are positive and satisfy the equalities $E(Y_n) = \beta n + O(1)$ and $var(Y_n) = \gamma n + O(1)$ [3]. Explicit expressions of β and γ are also obtained in [3] that depend on the matrices A, M, and in particular on λ and the corresponding left and right eigenvectors.

Other properties concern the function $y(t) = u(it)/\lambda$ used in Section 4, defined for real t in a neighbourhood of 0. By Proposition 1, for any t near 0, $y(t)^n$ is the leading term of the characteristic function $\Psi_n(t)$. Moreover, for some $c > 0$ and every $|t| \leq c$, the following relations hold [3] ([1]):

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

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 that

$$y(t)^n = e^{-\frac{\gamma}{2}t^2 n + i\beta t n}(1 + O(t^3)n) \qquad \text{for } |t| \leq n^{-q} \tag{5}$$

The previous properties can be used to prove a local limit theorem for $\{Y_n\}$ when M is primitive [3]. The result holds under a further assumption (introduced to avoid periodicity phenomena) stating that for every $0 < t < 2\pi$

$$|\mu| < \lambda \quad \text{for every eigenvalue } \mu \text{ of } Ae^{it} + B \tag{6}$$

Such a property is studied in detail in [4] and is often verified. For instance it holds true whenever there are two indices i, j such that $A_{ij} > 0$ and $B_{ij} > 0$, or $A_{ii} > 0$ and $B_{jj} > 0$. Intuitively, it corresponds to an aperiodicity property of the oriented graph defined by matrices A and B concerning the number of occurrences of the label a in cycles of equal length.

The local limit theorem in the primitive case can be stated as follows.

Theorem 2. *Let $\{Y_n\}$ be defined by a linear representation (ξ, A, B, η) such that the matrix $M = A + B$ is primitive and assume that property (6) holds for every $0 < t < 2\pi$. Moreover, let β and γ be defined as above. Then, as n tends to $+\infty$, the following equation holds uniformly for every $k = 0, 1, \ldots, n$,*

$$Pr\{Y_n = k\} \ = \ \frac{e^{-\frac{(k-\beta n)^2}{2\gamma n}}}{\sqrt{2\pi\gamma n}} \ + \ o\left(\frac{1}{\sqrt{n}}\right) \tag{7}$$

[1] Here, for every interval $I \subseteq \mathbb{R}$ and functions $f, g : I \to \mathbb{C}$, by "$g(t) = O(f(t))$ for $t \in I$" we mean "$|g(t)| \leq b|f(t)|$ for all $t \in I$", for some constant $b > 0$.

4 Bicomponent models

In this section we study the behaviour of $\{Y_n\}_{n\in\mathbb{N}}$ in the bicomponent model. We first recall some notions and properties introduced in [7] for this model: in particular we need a sort of analogous of Proposition 1 in this case.

Here $\{Y_n\}_{n\in\mathbb{N}}$ is defined by a linear representation (ξ, A, B, η) of size m, such that the matrix $M = A+B$ consists of two primitive components. More precisely, there are two linear representations $(\xi_1, A_1, B_1, \eta_1)$, $(\xi_2, A_2, B_2, \eta_2)$, of size m_1 and m_2, respectively, with $m = m_1 + m_2$, such that for some $A_0, B_0 \in \mathbb{R}_+^{m_1\times m_2}$

$$\xi' = (\xi_1', \xi_2'), \quad A = \begin{pmatrix} A_1 & A_0 \\ 0 & A_2 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & B_0 \\ 0 & B_2 \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \tag{8}$$

Moreover we assume the following conditions:

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

B) $\xi_1 \neq 0 \neq \eta_2$ and $A_0 + B_0 \neq 0$;

C) $A_1 \neq 0 \neq B_1$ and $A_2 \neq 0 \neq B_2$.

Since the two components are primitive the properties presented in the previous section hold for each of them. In particular, for $j = 1, 2$, we can define $H^{(j)}(x, y)$, $h_n^{(j)}(z)$, $u_j(z)$, $y_j(t)$, β_j, and γ_j, respectively, as the values $H(x, y)$, $h_n(z)$, $u(z)$, $y(t)$, β, γ referred to component j. Note that condition C) guarantees that $0 < \beta_j < 1$ and $0 < \gamma_j$ for every $j = 1, 2$, while condition B) implies that both components contribute to probability values of Y_n.

In such a bicomponent model the limit distribution of $\{Y_n\}$ mainly depends on whether $\lambda_1 \neq \lambda_2$ or $\lambda_1 = \lambda_2$. If $\lambda_1 > \lambda_2$ then $\frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}}$ converges in distribution to a standard normal r.v. (the case $\lambda_1 < \lambda_2$ is symmetric) [7]. If $\lambda_1 = \lambda_2$ and $\beta_1 \neq \beta_2$ then Y_n/n converges in distribution to a random variable U uniformly distributed over the interval $[b_1, b_2]$, where $b_1 = \min\{\beta_1, \beta_2\}$ and $b_2 = \max\{\beta_1, \beta_2\}$ [7, Theorem 15]. This means that, for every $x \in \mathbb{R}$,

$$\lim_{n\to+\infty} \Pr(Y_n/n \leq x) = \Pr(U \leq x) \tag{9}$$

However this relation does not give information about the probability that Y_n takes a specific value $k \in \mathbb{N}$ (possibly depending on n). Here we want to show that adding a further condition on the model such a probability can be estimated at least for reasonable expressions $k = k(n)$. To this end, we still consider the case $\lambda_1 = \lambda_2$ and $\beta_1 \neq \beta_2$ and assume a further hypothesis analogous to condition (6): for every $0 < t < 2\pi$

$$|\mu| < \lambda \text{ for all eigenvalues } \mu \text{ of the matrices } A_1 e^{it} + B_1 \text{ and } A_2 e^{it} + B_2 \tag{10}$$

where we set $\lambda = \lambda_1 = \lambda_2$.

In this case, following [7], the matrix-valued function $H(x,y)$ is given by

$$H(x,y)=\sum_{n=0}^{+\infty}(Ax+B)^n y^n=\begin{bmatrix} H^{(1)}(x,y) & G(x,y) \\ 0 & H^{(2)}(x,y)\end{bmatrix}\ , \quad \text{where}$$
$$H^{(1)}(x,y)=\frac{\mathrm{Adj}\,(I-(A_1x+B_1)y)}{\mathrm{Det}\,(I-(A_1x+B_1)y)},\ H^{(2)}(x,y)=\frac{\mathrm{Adj}\,(I-(A_2x+B_2)y)}{\mathrm{Det}\,(I-(A_2x+B_2)y)} \tag{11}$$
$$\text{and}\ \ G(x,y)=H^{(1)}(x,y)\ (A_0x+B_0)y\ H^{(2)}(x,y)\ .$$

Thus, the generating function of $\{h_n(z)\}_n$ is now given by

$$\sum_{n=0}^{\infty}h_n(z)y^n=\xi' H(e^z,y)\eta=\xi_1' H^{(1)}(e^z,y)\eta_1+\xi_1' G(e^z,y)\eta_2+\xi_2' H^{(2)}(e^z,y)\eta_2$$

An analysis of the singularities of $\xi' H(e^z,y)\eta$ is presented in [7, Sec.7.2] where the following property is proved.

Proposition 3. *For some constant $c>0$ and every $z\in\mathbb{C}$ such that $|z|\le c$, we have*

$$h_n(z)=s(z)\sum_{j=0}^{n-1}u_1(z)^j u_2(z)^{n-1-j}+O(u_1(z)^n)+O(u_2(z)^n)$$

where $s(z)$ is a function analytic and non-null for $|z|\le c$.

Since $u_1(0)=\lambda=u_2(0)$ the previous proposition implies

$$h_n(0)=s(0)n\lambda^{n-1}+O(\lambda^n)\qquad (s(0)\neq 0) \tag{12}$$

4.1 Analysis of the characteristic function

Here we study the characteristic function $\Psi_n(t)=\frac{h_n(it)}{h_n(0)}$, for $-\pi\le t\le\pi$. We split this interval in three sets:

$$|t|\le n^{-q},\quad n^{-q}<|t|<c,\quad c\le|t|\le\pi$$

where c and q are positive constants and $\frac{1}{3}<q<\frac{1}{2}$. We observe that such a splitting is typical of the "Saddle Point Method", and it is often used to derive local limit properties in the Gaussian case [9].

Proposition 4. *For every $0<c<\pi$ there exists $0<\varepsilon<1$ such that*

$$|\Psi_n(t)|=O(\varepsilon^n)\qquad \textit{for all}\ \ c\le|t|\le\pi.$$

Proof. From equations (11) it is clear that, for every $z\in\mathbb{C}$, the singularities of the generating function $\xi' H(e^z,y)\eta$ are the inverses of the eigenvalues of the matrices $(A_1e^z+B_1)$ and $(A_2e^z+B_2)$. Then, by condition (10), for every $0<c<\pi$, all singularities of $\xi' H(e^{it},y)\eta$, for any $c\le|t|\le\pi$, are in modulus greater than a

value τ^{-1} such that $0 < \tau < \lambda$, and hence $|h_n(it)| = O(\tau^n)$. Thus, by equality (12), for some $0 < \varepsilon < 1$ we have

$$|\Psi_n(t)| = \left|\frac{h_n(it)}{h_n(0)}\right| = \frac{O(\tau^n)}{\Theta(n\lambda^n)} = O(\varepsilon^n) \qquad \text{for any } \; c \le |t| \le \pi$$

□

Now, let us study $\Psi_n(t)$ for t in a neighbourhood of 0. We recall that in such a set both functions $y_1(t) = u_1(it)/\lambda$ and $y_2(t) = u_2(it)/\lambda$ satisfy equations (4). Then, for some $c > 0$ and every $|t| \le c$, we have

$$y_1(t) = 1 + i\beta_1 t + O(t^2)\ , \qquad y_2(t) = 1 + i\beta_2 t + O(t^2) \tag{13}$$

$$|y_1(t)| \le e^{-\frac{\gamma_1}{4}t^2}\ , \qquad |y_2(t)| \le e^{-\frac{\gamma_2}{4}t^2} \tag{14}$$

Moreover, by Proposition 3 we immediately get, for $|t| \le c$, with $t \neq 0$,

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

Thus from equalities (12), (13), (14), we have

$$\Psi_n(t) = \frac{h_n(it)}{h_n(0)} = (1 + O(t))\left(\frac{y_1(t)^n - y_2(t)^n}{it\,(\beta_1 - \beta_2)\,n}\right) + \sum_{j=1,2} O\left(\frac{e^{-\frac{\gamma_j}{4}t^2 n}}{n}\right) \tag{15}$$

This leads to evaluate $\Psi_n(t)$ in the second set, i.e. for $n^{-q} < |t| < c$.

Proposition 5. *Let $0 < q < 1/2$. Then there are two positive constants a, c such that, for every real t satisfying $n^{-q} < |t| < c$,*

$$|\Psi_n(t)| = O\left(e^{-an^{1-2q}}\right)$$

Proof. From equation (15), taking $a = \min\{\gamma_1, \gamma_2\}/4$, we obtain for some $c > 0$

$$|\Psi_n(t)| \le \frac{|y_1(t)|^n + |y_2(t)|^n}{\Theta(n^{1-q})} + O\left(e^{-at^2 n}/n\right) \qquad \text{for all } n^{-q} < |t| < c$$

and by (14) we get $|\Psi_n(t)| = O\left(n^{q-1}e^{-an^{1-2q}}\right)$ proving the result. □

Now, let us evaluate $\Psi_n(t)$ in the first set, that is for $|t| \le n^{-q}$ where $1/3 < q < 1/2$. First note that, by simple computations, the following relations can be proved:

$$\int_{|t|\le n^{-q}} O\left(e^{-\frac{\gamma_j}{4}t^2 n}/n\right) dt = O(n^{-1-q}) = o(n^{-4/3}) \qquad \text{for } j = 1, 2\ ,$$

$$\int_{|t|\le n^{-q}} O(t)\frac{y_1(t)^n - y_2(t)^n}{it\,(\beta_1 - \beta_2)\,n} dt = \int_{|t|\le n^{-q}} O(1/n) dt = o(n^{-4/3})$$

Therefore, by equation (15), for every $k \in \{0, 1, \ldots, n\}$ we get

$$\int_{|t|\le n^{-q}} \Psi_n(t) e^{-ikt} dt = \int_{|t|\le n^{-q}} \left(\frac{y_1(t)^n - y_2(t)^n}{it\,(\beta_1 - \beta_2)\,n} \right) e^{-ikt} dt \;+\; o(n^{-4/3}) \quad (16)$$

Also observe that both $y_1(t)$ and $y_2(t)$ satisfy equation (5), whence

$$y_j(t)^n = e^{-\frac{\gamma_j}{2} t^2 n + i\beta_j tn}(1 + O(t^3)n) \quad \text{for all } |t| \le n^{-q}, \quad j = 1, 2$$

Thus, replacing these values in (16), after some computations (similar to the previous ones) we obtain the following

Proposition 6. *For every $k \in \{0, 1, \ldots, n\}$ and every $1/3 < q < 1/2$ we have*

$$\int_{|t|\le n^{-q}} \Psi_n(t) e^{-ikt} dt = \int_{|t|\le n^{-q}} \left(\frac{e^{-\frac{\gamma_1}{2} t^2 n + i\beta_1 tn} - e^{-\frac{\gamma_2}{2} t^2 n + i\beta_2 tn}}{it\,(\beta_1 - \beta_2)\,n} \right) e^{-ikt} dt + o(1/n)$$

4.2 Main result

Without loss of generality assume $\beta_1 < \beta_2$, and denote by $f_U(x)$ the density function of a uniform r.v. U in the interval $[\beta_1, \beta_2]$, that is

$$f_U(x) = \frac{1}{\beta_2 - \beta_1} \chi_{[\beta_1, \beta_2]}(x) \qquad \forall x \in \mathbb{R}$$

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

For our purpose we need the following property.

Lemma 7. *For $k, m \in \mathbb{N}$, $k < m$, let $g : [2k\pi, 2m\pi] \to \mathbb{R}_+$ be a monotone function, and let $I_{k,m} = \int_{2k\pi}^{2m\pi} g(x) \sin(x) dx$. Then:*

a) if g is non-increasing we have $0 \le I_{k,m} \le 2[g(2k\pi) - g(2m\pi)]$;

b) if g is non-decreasing we have $2[g(2k\pi) - g(2m\pi)] \le I_{k,m} \le 0$.

In both cases $|I_{k,m}| \le 2|g(2k\pi) - g(2m\pi)|$.

Proof. If g is non-increasing, for each integer $k \le j < m$ the following relations hold

$$I_{j,j+1} = \int_{2j\pi}^{(2j+1)\pi} g(x) \sin(x) dx - \int_{(2j+1)\pi}^{2(j+1)\pi} g(x) |\sin(x)| dx \le 2[g(2j\pi) - g(2(j+1)\pi)]$$

while $0 \le I_{j,j+1}$ is obvious. Thus a) follows by summing the expressions above for $j = k, \ldots, m-1$.

Part b) follows by applying a) to the function $h(x) = g(2m\pi) - g(x)$. □

Now, we can state our main result.

Theorem 8. *Let (ξ, A, B, η) be a linear representation of the form (8) satisfying conditions A), B), C) above; also assume $\lambda_1 = \lambda_2$, $\beta_1 \ne \beta_2$ together with the aperiodicity condition (10). Then, the r.v. Y_n satisfies the relation*

$$\lim_{n \to +\infty} \; n \; Pr(Y_n = k) \;=\; f_U(x) \quad (17)$$

for every integer $k = k(n)$, provided that $k/n \to x$ for a constant x such that $\beta_1 \ne x \ne \beta_2$ (as $n \to +\infty$).

Proof. It is known [10] that the probability $p_n(k) = \Pr\{Y_n = k\}$, for every $k \in \{0, 1, \ldots, n\}$, can be obtained from $\Psi_n(t)$ by the inversion formula

$$p_n(k) \;=\; \frac{1}{2\pi}\int_{-\pi}^{\pi} \Psi_n(t) e^{-itk} dt$$

To evaluate the integral above let us split the interval $[-\pi, \pi]$ into the three sets

$$[-n^{-q}, n^{-q}]\,, \qquad \{t \in \mathbb{R} : n^{-q} < |t| < c\}\,, \qquad \{t \in \mathbb{R} : c \le |t| \le \pi\}$$

with c as in Proposition 5 and some $1/3 < q < 1/2$. Then, by Propositions 4, 5, 6, we obtain

$$p_n(k) = \frac{1}{2\pi}\int_{|t|\le n^{-q}} \left(\frac{e^{-\frac{\gamma_2}{2}t^2 n + i\beta_2 tn} - e^{-\frac{\gamma_1}{2}t^2 n + i\beta_1 tn}}{it\,(\beta_2 - \beta_1)\,n} \right) e^{-ikt} dt + o(1/n) \quad (18)$$

Now, set $v = k/n$ and note that for $n \to +\infty$, v approaches a value different from β_1 and β_2. Thus, defining

$$\Delta_n(v) = \int_{|t|\le n^{-q}} \frac{e^{i(\beta_2 - v)tn - \frac{\gamma_2}{2}t^2 n} - e^{i(\beta_1 - v)tn - \frac{\gamma_1}{2}t^2 n}}{i(\beta_2 - \beta_1)t}\, dt$$

we have to prove that

$$\Delta_n(v) = 2\pi f_U(v) + o(1) \quad (19)$$

Since $\beta_1 < \beta_2$ set $\delta = \beta_2 - \beta_1$. Then, $\Delta_n(v)$ is the integral of the difference of two functions of the form

$$A_n(t, v) = \frac{e^{i(\beta - v)tn - \frac{\gamma}{2}t^2 n} - 1}{i\delta t}$$

where β and γ take the values β_2, γ_2 and β_1, γ_1, respectively. Using the symmetries of real and imaginary part of A_n, by a change of variable we get

$$\int_{|t|\le n^{-q}} A_n(t, v) dt \;= \frac{2}{\delta}\int_0^{n^{-q}} \frac{e^{-\frac{\gamma}{2}t^2 n}\sin((\beta - v)tn)}{t}\, dt \;\;=$$

$$= \;\frac{2}{\delta}\int_0^{(\beta - v)n^{1-q}} \frac{\sin(u)}{u}\, du - \frac{2}{\delta}\int_0^{(\beta - v)n^{1-q}} \left(1 - e^{-\frac{\gamma u^2}{2(\beta - v)^2 n}}\right) \frac{\sin(u)}{u}\, du \quad (20)$$

As $\int_0^{+\infty} \frac{\sin(u)}{u} du = \pi/2$, for $n \to +\infty$ the first term converges to $\frac{\pi}{\delta}\,\mathrm{sgn}(\beta - v)$. Now we show that the second term of (20) tends to 0 as $n \to +\infty$. This term is equal to

$$\frac{2}{\delta}\int_0^{(\beta - v)n^{1-q}} B_n(u)\sin(u) du \quad (21)$$

where $B_n(u) = u^{-1}\left(1 - e^{-\frac{\gamma u^2}{2(\beta - v)^2 n}}\right)$. To evaluate (21) we use Lemma 7. Note that $B_n(u) > 0$ for all $u > 0$, and $\lim_{u\to 0} B_n(u) = 0 = \lim_{u\to+\infty} B_n(u)$. Moreover in the set $(0, +\infty)$ its derivative is null only at the point $u_n = \alpha|\beta - v|\sqrt{n/\gamma}$,

for a constant $\alpha \in (1,2)$ independent of n and v. Thus, for n large enough, u_n belongs to the interval $(0, |\beta - v| n^{1-q})$, $B_n(u)$ is increasing in the set $(0, u_n)$ and decreasing in $(u_n, +\infty)$, while its maximum value is

$$B_n(u_n) = \frac{1 - e^{-\frac{\alpha^2}{2}}}{\alpha |\beta - v|} \sqrt{\frac{\gamma}{n}} = \Theta(n^{-1/2})$$

Defining $k_n = \lfloor \frac{u_n}{2\pi} \rfloor$ and $K = \lfloor \frac{|\beta - v| n^{1-q}}{2\pi} \rfloor$, we can apply Lemma 7 to the intervals $[0, 2k_n]$ and $[2k_n + 2, 2K]$, to get

$$\begin{aligned}
\left| \int_0^{|\beta - v| n^{1-q}} B_n(u) \sin u du \right| &\leq 2B_n(2k_n\pi) + \left| \int_{2k_n\pi}^{2(k_n+1)\pi} B_n(u) \sin u du \right| + \\
&\quad + 2[B_n(2(k_n+1)\pi) - B_n(2K\pi)] + \int_{2K\pi}^{|\beta - v| n^{1-q}} B_n(u) \sin u du \\
&\leq 2B_n(2k_n\pi) + 2B_n(u_n) + 2[B_n(2(k_n+1)\pi) - B_n(2K\pi)] + 2B_n(2K\pi) \\
&\leq 6B_n(u_n) \;=\; \frac{c\,\sqrt{\gamma}}{|\beta - v|\sqrt{n}}
\end{aligned}$$

where c is a positive constant independent of v and n.

This implies that, for any v approaching a constant different from β_1 and β_2, the second term of (20) is $O(n^{-1/2})$. Therefore, we get

$$\begin{aligned}
\Delta_n(v) \;&=\; \frac{2}{\delta} \left[\int_0^{(\beta_2 - v) n^{1-q}} \frac{\sin u}{u} du - \int_0^{(\beta_1 - v) n^{1-q}} \frac{\sin u}{u} du \right] + \mathrm{O}(n^{-1/2}) \\
&=\; \frac{\pi}{\delta} \left[\mathrm{sgn}(\beta_2 - v) - \mathrm{sgn}(\beta_1 - v) \right] + o(1) \;=\; 2\pi f_U(v) + o(1)
\end{aligned}$$

which proves equation (19) and the proof is complete. □

A typical consequence of this result is that $n\Pr(Y_n = \lfloor xn \rfloor)$ converges to $f_U(x)$ for every real x different from β_1 and β_2. Intuitively equalities of the form (17) are considered more precise than convergence in distribution since they estimate the probability that the n-th random variable of the sequence takes a specific value rather than lying on an interval.

On the other hand we observe that (without condition (10)) the convergence in distribution (9) does not imply our equality (17). In particular if there are periodicity phenomena in the occurrences of letter a it may happen that (9) holds while (17) does not. For instance if the overall series r of linear representation (ξ, A, B, η) has non-zero coefficients (r, w) only for words w with even $|w|_a$, then $\Pr(Y_n = k) = 0$ for all odd integers k, and hence (17) cannot hold while (9) may still be valid. This observation also shows that condition (10) prevents such periodicity phenomena in the stochastic model.

5 Conclusions

In this work we have presented a non-Gaussian local limit property for the number of occurrences of a symbol in words generated at random according with a rational stochastic model defined by a linear representation with two primitive components. Our result concerns the case when the two components have the same dominant eigenvalue but different main constants of the respective mean value (β_1 and β_2). We expect that in case of different dominant eigenvalues a Gaussian local limit property holds, where the main terms of mean value and variance correspond to the dominant component. On the contrary, we conjecture that results similar to ours (that is of a non-Gaussian type) hold for other rational stochastic models, defined by assuming different hypotheses on the key parameters associated to mean value and variance of the statistic of interest (e.g. $\beta_1 = \beta_2$), or assuming more than two primitive components with equal dominant eigenvalues.

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 descriptional complexity through analytic combinatorics. *Theory Comput. Systems*, 528:85–100, 2014.
6. S. Broda, A. Machiavelo, N. Moreira, and R. Reis. On the average complexity of strong star normal form. In G. Pighizzini and C. Campeanu (eds), Proc. 19th DCFS, LNCS vol. 10316, 77-88, 2017.
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. A. Denise. Génération aléatoire uniforme de mots de langages rationnels. *Theoret. Comput. Sci.*, 159:43–63, 1996.
9. P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge Univ. Press, 2009.
10. B.V. Gnedenko. *Theory of probability*. Gordon and Breach Science Publ., 1997.
11. P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. *Theoret. Comput. Sci.*, 287(2): 593–617, 2002.
12. B. Prum, F. Rudolphe, E. Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequence. *J. Roy. Statist. Soc. Ser. B*, 57: 205–220, 1995.
13. M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.
14. A. Salomaa and M. Soittola. *Automata-Theoretic Aspects of Formal Power Series*. Springer–Verlag, 1978.
15. E. Seneta. *Non-negative matrices and Markov chains*. Springer–Verlag, New York Heidelberg Berlin, 1981.