Frequency of symbol occurrences in simple non-primitive 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 study 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 we assume is a simple non-ergodic model defined by the product of two primitive rational formal series, which form two distinct ergodic components. We obtain asymptotic evaluations for the mean and the variance of Y_n and its limit distribution. It turns out that there are two main cases: if one component is dominant and non-degenerate we get a Gaussian limit distribution; if the two components are equipotent and have different leading terms of the mean, we get a uniform limit distribution. Other particular limit distributions are obtained in the case of a degenerate dominant component and in the equipotent case when the leading terms of the expectation values are equal.

1 Introduction

The analysis of the frequency of pattern occurrences in a long string of symbols, usually called *text*, is a classical problem that is of interest in several research areas of computer science and molecular biology. In computer science for instance it has been studied in connection with the design of algorithms for approximate pattern-matching [13, 16] and the analysis of problems of code synchronization [10]. The problem is particularly relevant in molecular biology to study properties of DNA sequences and for gene recognition [20]. For instance, biological informations can be obtained from unexpected frequencies of special deviant motifs in a DNA text [7, 9, 15]. Moreover, the frequency problems in a probabilistic framework are studied in [12, 1, 17, 14]. In this context a set of one or more patterns is given and the text 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 the previous occurrence.

A more general approach, developed in the area of automata and formal languages, is recently proposed in [3, 4], where the pattern is reduced to a single

⋆ Appeared in revised form in *Proc. DLT 2003*, 7th International Conference on Developments in Language Theory, LNCS n.2710, 242–253, Springer, 2003.

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

symbol and the text is randomly generated according to a stochastic model defined by a rational formal series in two non-commutative variables. We recall that there are well-known linear time algorithms to generate a random word in such a model [6] which we call the *rational model* in this work. The frequency problem in this model is also related to the ambiguity of rational grammars and to the asymptotic form of the coefficients of rational and algebraic formal series studied for instance in [21, 18].

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. Under this hypothesis, asymptotic expressions for the mean and the variance of the statistics under investigation are obtained, together with their limit distributions expressed in the form of both central and local limit theorems [3, 4].

In this work we study the symbol frequency problem in the rational model in a simple non-ergodic case, that is when the rational series defining the stochastic source is the product of two primitive rational series. This case is rather representative of a more general situation where the matrix associated with the rational model has two primitive components. We obtain asymptotic evaluations for the mean and the variance of the number of symbol occurrences and its limit distribution. It turns out that there are two main cases. In the dominant case the main eigenvalue associated with one component is strictly greater than the main eigenvalue associated with the other. In the equipotent case these two eigenvalues are equal.

If one component is dominant and does not degenerate[1], the main terms of mean and variance are determined by such a component and we get a Gaussian limit distribution. We also determine the limit distribution when there exists a dominant degenerate component. Apparently, this has a large variety of possible forms depending even on the other (non-main) eigenvalues of the secondary component and including the geometric law in some simple cases.

If the two components are equipotent and have different leading terms of the mean, then the variance is of a quadratic order showing there is not a concentration phenomenon around the average value of our statistics. In this case we get a uniform limit distribution between the constants of the leading terms of the expected values associated with the two components.

However, in the equipotent case, if the leading terms of the two means are equal then the variance reduces to a linear order of growth and we have again a concentration phenomenon. In this case the limit distribution depends on the main terms of the variances associated with the two components: if they are equal we obtain a Gaussian limit distribution again; if they are different we

[1] i.e., considering the series of the dominant component, both symbols of the alphabet appear in some words with non-null coefficient.

obtain a limit distribution defined by a mixture of Gaussian random variables of mean 0 and variance uniformly distributed in a given interval.

The main contribution of these results is related to the non-ergodic hypothesis. To our knowledge, the pattern frequency problem in the Markovian model is usually studied in the literature under ergodic hypothesis and Gaussian limit distributions are generally obtained. On the contrary, here we get in many cases limit distributions quite different from the Gaussian one.

We think our analysis is significant also from a methodogical point of view: we adapt methods and ideas introduced to deal with the Markovian model to a more general stochastic model, the rational one, which seems to be the natural setting for these techniques.

Due to space constraints, in this paper all proofs are omitted. They can be found in [5] and rely on singularity analysis of the bivariate generating functions associated with the statistics under investigation.

The computations described in our examples are executed by using MATHEMATICA [22].

2 Preliminary notions

2.1 Perron–Frobenius theory

The Perron–Frobenius theory is a well-known subject widely studied in the literature (see for instance [19]). To recall its main results we first establish some notation. For every pair of matrices $T = [T_{ij}]$, $S = [S_{ij}]$, the expression $T > S$ means that $T_{ij} > S_{ij}$ for every pair of indices i, j. As usual, we consider any vector v as a column vector and denote by v' the corresponding row vector. We recall that a nonnegative matrix T is called *primitive* if there exists $m \in \mathbb{N}$ such that $T^m > 0$. The main properties of such matrices are given by the following theorem [19, Sect.1].

Theorem 1 (Perron–Frobenius) *Let T be a primitive nonnegative matrix. There exists an eigenvalue λ of T (called Perron–Frobenius eigenvalue of T) such that:*

1. *λ is real and positive;*
2. *with λ we can associate strictly positive left and right eigenvectors;*
3. *$|\nu| < \lambda$ for every eigenvalue $\nu \neq \lambda$;*
4. *if $0 \leq C \leq T$ and γ is an eigenvalue of C, then $|\gamma| \leq \lambda$; moreover $|\gamma| = \lambda$ implies $C = T$;*
5. *λ is a simple root of the characteristic polynomial of T.*

2.2 Moments and limit distribution of a discrete random variable

Let X be an integer valued random variable (r.v.), such that $\Pr\{X = k\} = p_k$ for every $k \in \mathbb{N}$. Consider its moment generating function $\Psi_X(z) = \sum_{k\in\mathbb{N}} p_k e^{zk}$; then 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

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

Φ_X is always well-defined for every $t \in \mathbb{R}$, it is periodic of period 2π and it completely characterizes the r.v. X. Moreover it represents the classical tool to prove convergence in distribution: a sequence of random variables $\{X_n\}_n$ converges to a r.v. X in distribution[2] 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 [8].

3 The rational stochastic model

Here we define the rational stochastic model. According to [2] a formal series in the non-commutative variables a, b, with coefficients in the semiring $\mathbb{R}_+$ of non-negative real numbers, is a function $r : \{a,b\}^* \longrightarrow \mathbb{R}_+$. For any word $w \in \{a,b\}^*$, we denote by (r,w) the value of r at w and a series r is usually represented as a sum in the form

$$r = \sum_{w \in \{a,b\}^*} (r,w)w$$

The set of all such series is denoted by $\mathbb{R}_+\langle\langle a,b\rangle\rangle$. It is well-known that $\mathbb{R}_+\langle\langle a,b\rangle\rangle$ forms a semiring with respect to the traditional operation of sum and Cauchy product.

Now, given $r \in \mathbb{R}_+\langle\langle a,b\rangle\rangle$, we can define a stochastic model as follows. Consider a positive $n \in \mathbb{N}$ such that $(r,w) \neq 0$ for some string $w \in \{a,b\}^*$ of length n. For every integer $0 \leq k \leq n$ set

$$\varphi_k^{(n)} = \sum_{|w|=n, |w|_a=k} (r,w)$$

and define the random variable (r.v.) Y_n such that

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

Roughly speaking, Y_n represents the number of occurrences of a in a word of length n randomly generated according to the stochastic model defined by r. This model is of particular interest in the case of rational series. We recall that a series $r \in \mathbb{R}_+\langle\langle a,b\rangle\rangle$ is said to be *rational* if for some integer $m > 0$ there exists a monoid morphism $\mu : \{a,b\}^* \longrightarrow \mathbb{R}_+^{m\times m}$, a pair of (column) vectors $\xi, \eta \in \mathbb{R}_+^m$ such that $(r,w) = \xi'\mu(w)\eta$ for every $w \in \{a,b\}^+$. The triple (ξ, μ, η) is called *linear representation* of r.

[2] I.e. $\lim_{n\to\infty} F_{X_n}(\tau) = F_X(\tau)$ for every point $\tau \in \mathbb{R}$ of continuity for F_X, where $F_{X_n}(\tau) = \Pr\{X_n \leq \tau\}$ and $F_X(\tau) = \Pr\{X \leq \tau\}$.

We say that $\{Y_n\}$ is defined in a *rational* stochastic model if the associated series r is rational. It turns out that classical probabilistic models as the Bernoulli or the Markov processes, frequently used to study the number of occurrences of regular patterns in random words [12, 14], are special cases of rational stochastic models [3].

4 The primitive case

The asymptotic behaviour of Y_n is studied in [3, 4] when r is rational and admits a primitive linear representation, i.e. a linear representation (ξ, μ, η) such that the matrix $M = A + B$ is primitive, where $A = \mu(a)$ and $B = \mu(b)$. Under this hypothesis, let λ be the Perron–Frobenius eigenvalue of M; from Theorem 1, 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, for some $c > 0$ and $0 \leq \varepsilon < 1$, $|C(n)_{ij}| \leq c\varepsilon^n$ (for any i, j and all n large enough) and v' and u are strictly positive left and right eigenvectors of M/λ corresponding to the eigenvalue 1, normed so that $v'u = 1$. Moreover, the matrix $C = \sum_{n=0}^{\infty} C(n)$ is well-defined and $v'C = Cu = 0$.

Using these properties, it is proved in [3] that the mean and the variance of Y_n satisfy the relations

$$\mathbb{E}(Y_n) = \beta n + \frac{\delta}{\alpha} + \mathrm{O}\left(\varepsilon^n\right), \qquad \mathbb{V}ar(Y_n) = \gamma n + \mathrm{O}(1) \tag{2}$$

where α, β, γ and δ are constants defined by

$$\beta = \frac{v'Au}{\lambda}\,, \qquad \gamma = \beta - \beta^2 + 2\frac{v'ACAu}{\lambda^2}$$

$$\alpha = (\xi'u)(v'\eta)\,, \qquad \delta = \left(\xi' C \frac{A}{\lambda} u\right)(v'\eta) + (\xi'u)\left(v'\frac{A}{\lambda} C\eta\right)\,.$$

Notice that $B = 0$ implies $\beta = 1$ and $\gamma = \delta = 0$, while $A = 0$ implies $\beta = \gamma = \delta = 0$; on the other side, if $A \neq 0 \neq B$ then $\beta > 0$ and it turns out that also $\gamma > 0$.

Relations (2) is proved from (1) observing that $\Psi_{Y_n}(z) = \frac{h_n(z)}{h_n(0)}$, where

$$h_n(z) = \sum_{k=0}^{n} \varphi_k^{(n)} e^{zk} = \xi'(Ae^z + B)^n \eta\,,$$

and studying the asymptotic behaviour of $h_n(0)$, $h_n'(0)$ and $h_n''(0)$. This analysis is essentially based on Theorem 1 and on a sort of simple differential calculus for matrices.

Finally, the characteristic function $\Phi_{Y_n}(t) = \frac{h_n(it)}{h_n(0)}$ is used in [3] to prove that, if M is primitive and $A \neq 0 \neq B$, then the distribution of Y_n approximates a normal distribution, i.e. for every $x \in \mathbb{R}$

$$\lim_{n \longrightarrow +\infty} \Pr\left\{ \frac{Y_n - \beta n}{\sqrt{\gamma n}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \, .$$

5 The product model

Given two primitive linear representations (ξ_1, μ_1, η_1) and (ξ_2, μ_2, η_2) over the alphabet $\{a, b\}$, let r be the formal series defined by

$$(r, w) = \sum_{w=xy} [\xi_1' \mu_1(x) \eta_1] \cdot [\xi_2' \mu_2(y) \eta_2]$$

for every $w \in \{a, b\}^*$. It turns out that r admits a linear representation (ξ, μ, η) given by

$$\xi = \begin{pmatrix} \xi_1 \\ 0 \end{pmatrix} \ , \quad \mu(x) = \begin{pmatrix} \mu_1(x) & \eta_1 \xi_2' \mu_2(x) \\ 0 & \mu_2(x) \end{pmatrix} \ , \quad \eta = \begin{pmatrix} \eta_1 \xi_2' \eta_2 \\ \eta_2 \end{pmatrix} \tag{3}$$

Using the notation introduced in the previous section, from now on we refer the terms M, A, B and $h_n(z)$ to the product series r. To avoid trivial cases, throughout this work we assume $A \neq 0 \neq B$. We also use the obvious extension of appending indices 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 $i = 1, 2$, the values $Y_n^{(i)}$, M_i, λ_i, A_i, B_i, $h_n^{(i)}(z)$, β_i, γ_i are well-defined and associated with the linear representation (ξ_i, μ_i, η_i).

From the decomposition (3) it is easy to see that $h_n(z)$ is given by

$$h_n(z) = \sum_{i=0}^{n} \xi_1' (A_1 e^z + B_1)^i \eta_1 \xi_2' (A_2 e^z + B_2)^{n-i} \eta_2 = \sum_{i=0}^{n} h_i^{(1)}(z) h_{n-i}^{(2)}(z) \tag{4}$$

which is the convolution of $h_n^{(1)}(z)$ and $h_n^{(2)}(z)$. Since (ξ_1, μ_1, η_1) and (ξ_2, μ_2, η_2) are primitive, we can consider the Perron-Frobenius eigenvalues λ_1, λ_2 of M_1 and M_2, respectively. The properties of Y_n now depend on whether these two values 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_i, μ_i, η_i) is the *dominant* component if $\lambda_1 \neq \lambda_2$ and $\lambda_i = \max\{\lambda_1, \lambda_2\}$. On the contrary, if $\lambda_1 = \lambda_2$, both components give a contribution to the asymptotic behaviour of Y_n and hence we say they are *equipotent*.

6 Main results

In this section we summarize the main results concerning the product model. 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.

6.1 Dominant case

Using (4) and the results of the primitive case, one can determine asymptotic expressions for $h_n(0)$ and its derivatives, which yield the following theorem.

Theorem 2 *Assume $\lambda_1 > \lambda_2$. Then the following statements hold:*
i) if $A_1 \neq 0 \neq B_1$, then $\mathbb{E}(Y_n) = \beta_1 n + O(1)$ and $\mathbb{V}ar(Y_n) = \gamma_1 n + O(1)$;
ii) if $A_1 \neq 0$ and $B_1 = 0$, then $\mathbb{E}(Y_n) = n + O(1)$ and $\mathbb{V}ar(Y_n) = c_1 + O(\varepsilon^n)$;
iii) if $A_1 = 0$ and $B_1 \neq 0$, then $\mathbb{E}(Y_n) = c_2 + O(\varepsilon^n)$ and $\mathbb{V}ar(Y_n) = c_3 + O(\varepsilon^n)$;
where $\beta_1 > 0$, $\gamma_1 > 0$ and c_i and ε are constants such that $c_i > 0$ and $|\varepsilon| < 1$.

As far as the limit distribution of $\{Y_n\}$ is concerned, if the dominant component does not degenerate (i.e. $A_1 \neq 0 \neq B_1$) the analysis is similar to the primitive case and gives rise to a Gaussian limit distribution [3]. On the contrary, if the dominant component degenerates, the limit distribution may assume different forms, depending on the second component. In both cases the proof is based on the analysis of the characteristic function of Y_n.

Theorem 3 *Let $\lambda_1 > \lambda_2$. If $A_1 \neq 0 \neq B_1$ then $\frac{Y_n - \beta_1 n}{\sqrt{\gamma_1 n}}$ converges in distribution to the normal random variable of mean 0 and variance 1.*

If either $A_1 = 0$ or $B_1 = 0$ then $\gamma_1 = 0$ and the previous theorem does not hold.

Theorem 4 *Let $\lambda_1 > \lambda_2$. If $A_1 = 0$, then the random variables Y_n converges in distribution to the random variable Z of characteristic function*

$$\Phi_Z(t) = \frac{\xi_2'(\lambda_1 I - A_2 e^{it} - B_2)^{-1}\eta_2}{\xi_2'(\lambda_1 I - M_2)^{-1}\eta_2} \tag{5}$$

If $B_1 = 0$, then the random variables $n - Y_n$ converges in distribution to the random variable W of characteristic function

$$\Phi_W(t) = \frac{\xi_2'(\lambda_1 I - A_2 - B_2 e^{it})^{-1}\eta_2}{\xi_2'(\lambda_1 I - M_2)^{-1}\eta_2}.$$

Some comments on the random variables Z and W are now necessary. First observe that, when the matrices M_2, A_2 and B_2 have size 1, Z and W are geometric random variables. Indeed, in this case $M_2 = A_2 + B_2 = \lambda_2 < \lambda_1$ and we get

$$\Phi_Z(t) = \frac{1 - \frac{A_2}{\lambda_1 - B_2}}{1 - \frac{A_2}{\lambda_1 - B_2} e^{it}} \qquad \text{and} \qquad \Phi_W(t) = \frac{1 - \frac{B_2}{\lambda_1 - A_2}}{1 - \frac{B_2}{\lambda_1 - A_2} e^{it}}$$

which are the characteristic functions of geometric random variables of parameter $\frac{A_2}{\lambda_1 - B_2}$ and $\frac{B_2}{\lambda_1 - A_2}$ respectively. However, the range of possible behaviours of these random variables is much larger than what these examples show. To see this fact consider the function $\Phi_Z(t)$ in (5); it can be expressed in the form

$$\Phi_Z(t) = \sum_{j=0}^{\infty} \frac{\xi_2' \left(M_2/\lambda_2\right)^j \eta_2 \cdot (\lambda_2/\lambda_1)^j}{\sum_{i=0}^{\infty} \xi_2' \left(M_2/\lambda_2\right)^i \eta_2 \cdot (\lambda_2/\lambda_1)^i} \; \Phi_{Y_j^{(2)}}(t)$$

and hence it describes the random variable $Y_N^{(2)}$, where N is the random variable defined by the law

$$\Pr\{N = j\} = \frac{\xi_2' \left(M_2/\lambda_2\right)^j \eta_2 \cdot \left(\lambda_2/\lambda_1\right)^j}{\sum_{i=0}^{\infty} \xi_2' \left(M_2/\lambda_2\right)^i \eta_2 \cdot \left(\lambda_2/\lambda_1\right)^i} \,. \tag{6}$$

If $B_2 = 0$ then by (5) Z reduces to N, and an example of the rich range of its possible forms is shown by considering the case where $(A_1 = 0 = B_2)$ $\lambda_1 = 1.009$, $\lambda_2 = 1$, and the second component is represented by a generic (2×2) - matrix whose eigenvalues are 1 and μ such that $-1 < \mu < 1$. In this case, since the two main eigenvalues have similar values, the behaviour of $\Pr\{N = j\}$ for small j depends on the second component and in particular on its smallest eigenvalue μ. In Figure 1 we plot the probability law of N defined in (6) for $j = 0, 1, \ldots, 200$ in three cases: $\mu = -0.89$, $\mu = 0.00001$ and $\mu = 0.89$; the first picture compares the curves in the cases $\mu = -0.89$ and $\mu = 0.00001$, while the second picture compares the curves when $\mu = 0.00001$ and $\mu = 0.89$. Note that in the second case, when μ is almost null, we find a distribution similar to a (lengthy) geometric law while, for $\mu = -0.89$ and $\mu = 0.89$, we get a quite different behaviour which approximates the previous one for large values of j.

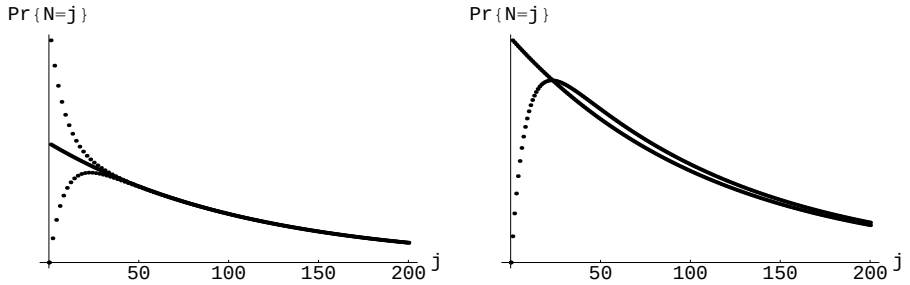

Fig. 1. Probability law of the random variable N defined in (6), for $j = 0, 1, \ldots, 200$. In the first picture we compare the case $\mu = 0.00001$ and $\mu = -0.89$. In the second one we compare the case $\mu = 0.00001$ and $\mu = +0.89$.

6.2 Equipotent components

In this section we consider the random variable Y_n assuming $\lambda_1 = \lambda_2$. Under this hypothesis two main subcases arise, depending on whether β_1 and β_2 are equal. First, we present the following theorem concerning the mean and the variance of Y_n, which can be obtained from equation (4) by singularity analysis.

Theorem 5 *If $\lambda_1 = \lambda_2$, then the mean and the variance of the random variable Y_n are given by*

$$\mathbb{E}(Y_n) = \frac{\beta_1 + \beta_2}{2} n + O(1)$$

$$\mathbb{V}ar(Y_n) = \begin{cases} \dfrac{(\beta_1 - \beta_2)^2}{12} n^2 + O(n) & \text{if } \beta_1 \neq \beta_2 \\ \dfrac{\gamma_1 + \gamma_2}{2} n + O(1) & \text{if } \beta_1 = \beta_2 \end{cases}$$

In the case $\beta_1 \neq \beta_2$ it is clear from the previous theorem that the variance is of a quadratic order. Hence, by using the characteristic function of Y_n/n one can prove the following result.

Theorem 6 *If $\lambda_1 = \lambda_2$ and $\beta_1 \neq \beta_2$ then Y_n/n converges in law to a random variable having uniform distribution in the interval* $[\min\{\beta_1, \beta_2\}, \max\{\beta_1, \beta_2\}]$.

If $\beta_1 = \beta_2$ then, since $A \neq 0 \neq B$, we have $A_i \neq 0 \neq B_i$ for $i = 1$ or $i = 2$. This implies $\gamma_i \neq 0$ and hence the variance is linear in n (see Theorem 5). In this case we get a concentration phenomenon of Y_n around its mean and we get two different limit distributions according to whether $\gamma_1 = \gamma_2$ or not. In the following, β and γ are defined by

$$\beta = \beta_1 = \beta_2, \qquad \gamma = \frac{\gamma_1 + \gamma_2}{2}.$$

First, we consider the case where the main terms of the variances are equal.

Theorem 7 *If $\lambda_1 = \lambda_2$, $\beta_1 = \beta_2$ and $\gamma_1 = \gamma_2$ then $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ converges in distribution to the normal random variable of mean* 0 *and variance* 1.

At last, we consider the case where the main terms of the variances are not equal.

Theorem 8 *If $\lambda_1 = \lambda_2$, $\beta_1 = \beta_2$ and $\gamma_1 \neq \gamma_2$ then $\frac{Y_n - \beta n}{\sqrt{\gamma n}}$ converges in distribution to the random variable of characteristic function*

$$\Phi(t) = \frac{2\left(e^{-\frac{\gamma_2}{2\gamma}t^2} - e^{-\frac{\gamma_1}{2\gamma}t^2}\right)\gamma}{(\gamma_1 - \gamma_2)t^2} \tag{7}$$

One can prove that the probability density corresponding to the characteristic function (7) is a mixture of Gaussian densities of mean 0, with variances uniformly distributed in the interval with extremes $\frac{\gamma}{\gamma_1}$ and $\frac{\gamma}{\gamma_2}$. Indeed, it is easy to see that

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

In Figure 2 we illustrate the form of the limit distributions obtained in this section (i.e. when $\lambda_1 = \lambda_2$ and $\beta_1 = \beta_2$). We represent the density of the random

variable having characteristic function (7), for different values of the ratio $p = \gamma_2/\gamma_1$. When p approaches 1, the curve tends to a Gaussian density according to Theorem 7; if γ_2 is much greater than γ_1, then we find a density with a cuspid in the origin corresponding to Theorem 8.

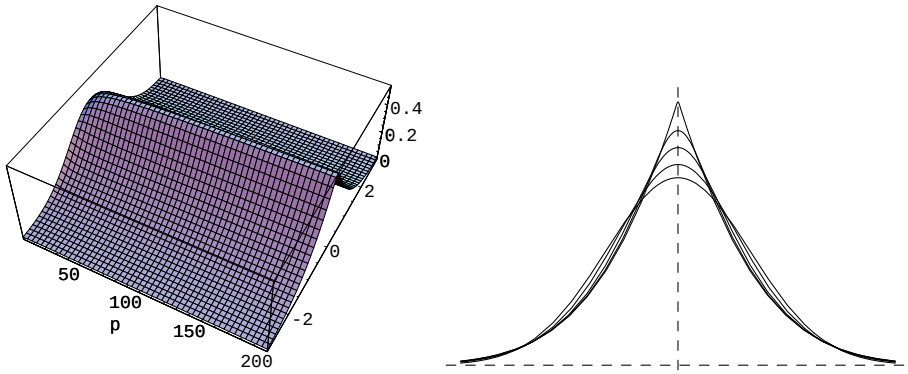

Fig. 2. The first picture represents the density of the random variable having characteristic function (7), according to the parameter $p = \gamma_2/\gamma_1$. The second picture represents some sections obtained for $p = 1.0001, 5, 15, 50, 20000$.

7 Examples

In this section we present an example which compares the limit distributions obtained in the non-degenerate dominant case and in the equipotent case, with different leading terms of the mean associated with each component.

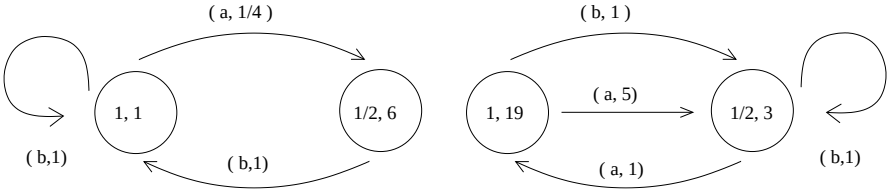

Fig. 3. Two weighted finite automata over the alphabet $\{a, b\}$, defining two primitive linear representations (ξ_i, μ_i, η_i), $i = 1, 2$. The matrices $A_i = \mu_i(a)$ and $B_i = \mu_i(b)$ are defined by the labels associated with transitions in the pictures. The values of the components of the arrays ξ_i and η_i are included in the corresponding states.

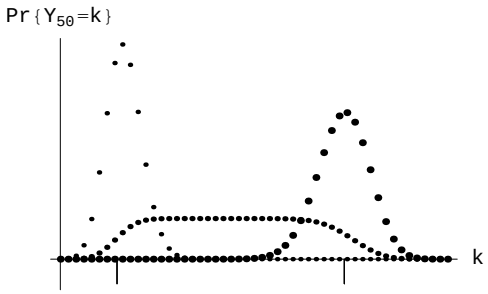

Fig. 4. Probability functions of Y_{50}, corresponding to a formal series derived from the automata of Figure 3 with weighted expanded by a constant factor, in the case where (λ_1, λ_2) are equal to (2,1), (1,2) and (1,1). The vertical bars have abscissas $50\beta_1$ and $50\beta_2$, respectively.

This example is based on the automata represented in Figure 3. They define two ergodic components with matrices M_i, A_i, B_i, $i = 1, 2$. The arrays ξ_i and η_i are given by the values included in the states.

Multiplying the matrices A_i and B_i (for $i = 1, 2$) by suitable factors, it is possible to build a family of primitive linear representations where we may have $\lambda_1 = \lambda_2$ or $\lambda_1 \neq \lambda_2$. In all cases, it turns out that $\beta_1 = 0.146447$ and $\beta_2 = 0.733333$ (and hence $\beta_1 \neq \beta_2$).

Figure 4 illustrates the probability function of the random variable Y_{50} in three different cases. If $\lambda_1 = 2$ and $\lambda_2 = 1$ we find a normal density of mean asymptotic to $50\,\beta_1$. If $\lambda_1 = 1$ and $\lambda_2 = 2$ we have a normal density of mean asymptotic to $50\beta_2$. Both situations correspond to Theorem 3. If $\lambda_1 = \lambda_2 = 1$, we recognize the convergence to the uniform distribution in the interval $[50\,\beta_1, 50\,\beta_2]$ according to Theorem 6.

References

1. E. A. Bender and F. Kochman. The distribution of subword counts is usually normal. *European Journal of Combinatorics*, 14:265–275, 1993.
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. *Rapporto Interno n. 274-02*, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, February 2002 (to appear in TCS).
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 (available at http://homes.dsi.unimi.it/∼goldwurm/home.html).
5. D. de Falco, M. Goldwurm, and V. Lonati. Frequency of symbol occurrences in simple non-primitive stochastic models. *Rapporto Interno n. 287-03*, Dipartimento di

Scienze dell'Informazione, Università degli Studi di Milano, February 2003 (available at http://homes.dsi.unimi.it/∼goldwurm/home.html).
6. A. Denise. Génération aléatoire uniforme de mots de langages rationnels. *Theoretical Computer Science*, 159:43–63, 1996.
7. J. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acid Res*, 10:5303–5318, 1982.
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. L. J. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM J. Appl. Math.*, 35:401–418, 1978.
11. L. J. Guibas and A. M. Odlyzko. Periods in strings. *Journal of Combinatorial Theory. Series A*, 30:19–43, 1981.
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. P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts *Proc. MFCS 91*, Lecture Notes in Computer Science, vol. n.520, Springer, 240–248, 1991.
14. P. Nicodeme, B. Salvy, and P. Flajolet. Motif statistics. In *Proceedings of the 7th ESA*, J. Nešeťril 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 the approximate pattern occurrence in a text. *Proc. Sequence '97*, Positano, 1997.
17. M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.
18. C. Reutenauer. *Propriétés arithmétiques et topologiques de séries rationnelles en variables non commutatives*, These Sc. Maths, Doctorat troisieme cycle, Université Paris VI, 1977.
19. E. Seneta. *Non-negative matrices and Markov chains*, Springer–Verlag, New York Heidelberg Berlin, 1981.
20. M. Waterman. *Introduction to computational biology*, Chapman & Hall, New York, 1995.
21. K. Wich. Sublinear ambiguity. In *Proceedings of the 25th MFCS*, M. Nielsen and B. Rovan editors. Lecture Notes in Computer Science, vol. n.1893, Springer, 2000, 690–698.
22. S. Wolfram. *The Mathematica book* Fourth Edition, Wolfram Media - Cambridge University Press, 1999.