Local Limit Distributions in Pattern Statistics: Beyond the Markovian Models*

Alberto Bertoni[1], Christian Choffrut[2], Massimiliano Goldwurm[1], and Violetta Lonati[1]

[1] Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano
Via Comelico 39/41, 20135 Milano – Italy
{bertoni,goldwurm,lonati}@dsi.unimi.it

[2] L.I.A.F.A. (Laboratoire d'Informatique Algorithmique, Fondements et Applications),
Université Paris VII, 2 Place Jussieu, 75221 Paris – France
Christian.Choffrut@liafa.jussieu.fr

Abstract. Motivated by problems of pattern statistics, we study the limit distribution of the random variable counting the number of occurrences of the symbol a in a word of length n chosen at random in $\{a, b\}^*$, according to a probability distribution defined via a finite automaton equipped with positive real weights. We determine the local limit distribution of such a quantity under the hypothesis that the transition matrix naturally associated with the finite automaton is primitive. Our probabilistic model extends the Markovian models traditionally used in the literature on pattern statistics.
This result is obtained by introducing a notion of symbol-periodicity for irreducible matrices whose entries are polynomials in one variable over an arbitrary positive semiring. This notion and the related results we prove are of interest in their own right, since they extend classical properties of the Perron–Frobenius Theory for non-negative real matrices.

Keywords: Automata and Formal Languages, Pattern statistics, Local Limit Theorems, Perron–Frobenius Theory.

1 Introduction

A typical problem in pattern statistics studies the frequency of occurrences of given strings in a random text, where the set of strings (patterns) is fixed in advance and the text is a word of length n randomly generated according to a probabilistic model (for instance, a Markovian model). In this context, relevant goals of research concern the asymptotic evaluations (as n grows) of the mean value and the variance of the number of occurrences of patterns in the text, as well as its limit distribution. This kind of problems are widely studied in the literature and they are of interest for the large variety of applications in different areas of computer science, probability theory and molecular biology (see for instance [8,12,11,14]). Many results show a normal limit distribution of the number of pattern occurrences in the sense of the central or local limit theorem [1]; here we recall

* This work has been supported by the Project M.I.U.R. COFIN 2003-2005 "Formal languages and automata: methods, models and applications".

V. Diekert and M. Habib (Eds.): STACS 2004, LNCS 2996, pp. 117–128, 2004.
© Springer-Verlag Berlin Heidelberg 2004

that the "local" result is usually stronger since it concerns the probability of single point values, while the "central" limit refers to the cumulative distribution function. In [10] limit distributions are obtained for the number of (positions of) occurrences of words from a regular language in a random string of length n generated in a Bernoulli or a Markovian model. These results are extended in [3] to the so-called *rational stochastic model*, where the pattern is reduced to a single symbol and the random text is a word over a two-letter alphabet, generated according to a probability distribution defined via a weighted finite automaton or, equivalently, via a rational formal series. 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 [10]. In the same paper [3], a normal local limit theorem is obtained for a proper subclass of primitive models. In this paper, we present a complete solution for primitive models, i.e. when the matrix associated with the rational formal series (counting the transitions between states) is primitive.

We now turn to a brief description of this paper. In Section 3, we introduce a notion of x-periodicity for irreducible matrices whose entries are polynomials in the variable x over an arbitrary positive semiring. Intuitively, considering the matrix as a labeled graph, its x-period is the GCD of the differences between the number of occurrences of x in (labels of) cycles of the same length. This notion and the related properties we prove are of interest in their own right, since they extend the classical notion of periodicity of non-negative matrices, studied in the Perron–Frobenius Theory for irreducible and primitive matrices [13]. In particular, these results are useful to study the eigenvalues of matrices of the form $Ax + B$, where A and B are matrices with coefficients in $\mathbb{R}_+$ and $x \in \mathbb{C}$ with $|x| = 1$ (see Theorem 2).

In Section 4 we prove our main result, concerning the local limit distribution of the random variable Y_n representing the number of occurrences of the symbol a in a word of length n chosen at random in $\{a, b\}^*$, according to any primitive rational model. Such a model can be described by means of a primitive matrix of the form $Ax+B$, where A and B are non-negative real matrices. If $Ax+B$ has x-period d, then we prove the existence of positive real constants α, β and non-negative real constants $C_0, C_1, \ldots, C_{d-1}$ with $\sum C_i = 1$ such that, as n tends to ∞, the relation

$$\mathrm{P}\{Y_n = k\} = \frac{d\, C_{\langle k \rangle_d}}{\sqrt{2\pi\alpha n}} \cdot e^{-\frac{(k-\beta n)^2}{2\alpha n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

holds uniformly for each $k = 0, 1, \ldots, n$ (here $\langle k \rangle_d = k - \lfloor k/d \rfloor$). If, in particular, $d = 1$ we get a normal local limit distribution, as already stated in [3].

2 Preliminaries

In this section we recall some basic notions and properties concerning rational formal series [2] and matrices over positive semirings [13].

2.1 Rational Formal Series and Weighted Automata

Let $\mathcal{S}$ be a positive semiring [9], that is a semiring such that $x+y=0$ implies $x=y=0$ and $x\cdot y=0$ implies $x=0$ or $y=0$. Examples are given by $\mathbb{N}$, $\mathbb{R}_+$ or the Boolean algebra $\mathbb{B}$. Given a finite alphabet Σ, we denote by Σ^* the set of all finite strings over Σ and by 1 the empty word. Moreover, for each $w\in\Sigma^*$, we denote by $|w|$ its length and by $|w|_b$ the number of occurrences of the symbol $b\in\Sigma$ in w.

We recall that a *formal series* over Σ with coefficients in $\mathcal{S}$ is a function $r:\Sigma^*\longrightarrow\mathcal{S}$. Usually, the value of r at w is denoted by (r,w) and we write $r=\sum_{w\in\Sigma^*}(r,w)\cdot w$. Moreover, r is called *rational* if there exists a *linear representation*, that is a triple (ξ,μ,η) where, for some integer $m>0$, ξ and η are (column) vectors in $\mathcal{S}^m$ and $\mu:\Sigma^*\longrightarrow\mathcal{S}^{m\times m}$ is a monoid morphism, such that $(r,w)=\xi^T\mu(w)\,\eta$ holds for each $w\in\Sigma^*$. We say that m is the *size* of the representation. Observe that considering such a triple (ξ,μ,η) is equivalent to defining a (weighted) non–deterministic automaton, where the state set is given by $\{1,2,\ldots,m\}$ and the transitions, the initial and the final states are assigned weights in $\mathcal{S}$ by μ, ξ and η respectively.

It is convenient to represent the morphism μ by its *state diagram*, see Figure 1, which is a labeled directed graph where the vertices are given by the set $\{1,2,\ldots,m\}$ and where there exists an edge with label $b\in\Sigma$ from vertex p to vertex q if $\mu(b)_{pq}\neq 0$. A *path* of *length* n is a sequence of labeled edges of the form

$$\ell = q_0 \xrightarrow{b_1} q_1 \xrightarrow{b_2} q_2 \ldots q_{n-1} \xrightarrow{b_n} q_n \ ;$$

in particular, if $q_n=q_0$ we say that ℓ is a q_0-cycle. Moreover we say that $w=b_1b_2\ldots b_n$ is the label of ℓ and we denote by $|\ell|_b=|w|_b$ the number of occurrences of b in ℓ.

Since we are interested in the occurrences of a particular symbol $a\in\Sigma$, we may set $A=\mu(a)$, $B=\sum_{b\neq a}\mu(b)$ and consider the *a-counting matrix* $M(x)=Ax+B$, which can be interpreted as a matrix whose entries are polynomials in $\mathcal{S}[x]$ of degree lower than 2. Moreover, observe that for every $n\in\mathbb{N}$ we can write

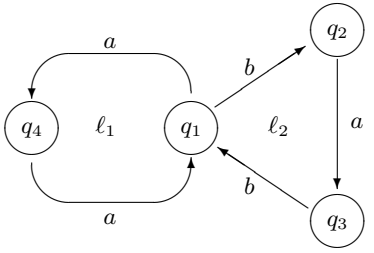
$$\xi^T M(x)^n \eta = \sum_{|w|=n} (r,w)\cdot x^{|w|_a} \ . \tag{1}$$

Therefore $M(x)^n$ is related to the paths of length n of the associated state diagram, in the sense that the pq-entry of $M(x)^n$ is the sum of monomials of the form sx^k where $k=|\ell|_a$ for some path ℓ of length n from p to q in the state diagram.

2.2 Matrix Periodicity

We now recall the classical notion of periodicity of matrices over positive semirings. Given a finite set Q and a positive semiring $\mathcal{S}$, consider a matrix $M:Q\times Q\to\mathcal{S}$. We say that M is *positive* whenever $M_{pq}\neq 0$ holds for all $p,q\in Q$, in which case we write $M>0$.

To avoid the use of brackets, from now on, we use the expression ${M^n}_{pq}$ to denote the pq-entry of the matrix M^n. For every index q, we call *period* of q the greatest common divisor (GCD) of the positive integers h such that ${M^h}_{qq}\neq 0$, with the convention that



$$M = \begin{pmatrix} 0 & 1 & 0 & x \\ 0 & 0 & x & 0 \\ 1 & 0 & 0 & 0 \\ x & 0 & 0 & 0 \end{pmatrix}$$

Fig. 1. Example of state diagram and a-counting matrix

$\mathrm{GCD}(\varnothing) = +\infty$. Moreover, we recall that a matrix M is said to be *irreducible* if for every pair of indices p, q, there exists a positive integer $h = h(p, q)$ such that ${M^h}_{pq} \neq 0$; in this case, it turns out that all indices have the same period, which is finite and is called the *period* of M. Finally, the matrix is called *primitive* if there exists a positive integer h such that $M^h > 0$, which implies $M^n > 0$ for every $n \geq h$. It is well-known that M is primitive if and only if M is irreducible and has period 1.

When $\mathcal{S}$ is the semiring of positive real numbers an important result is given by the following theorem (see [13]).

Theorem 1 (Perron–Frobenius). *Let M be a primitive matrix with entries in $\mathbb{R}_+$. Then, M admits exactly one eigenvalue λ of maximum modulus (called the Perron–Frobenius eigenvalue of M), which is a simple root of the characteristic polynomial of M. Moreover, λ is real and positive and there exist strictly positive left and right eigenvectors u and v associated with λ such that $v^T u = 1$.*

A consequence of this theorem is that, for any primitive matrix M with entries in $\mathbb{R}_+$, the relation $M^n \sim \lambda^n \cdot uv^T$ holds as n tends to $+\infty$, where λ, u and v are defined as above. A further application is given by the following proposition [13, Exercise 1.9], to be used in the next sections.

Proposition 1. *Let C be a complex matrix, set $|C| = (|C_{pq}|)$ and let γ be one of the eigenvalues of C. If M is a primitive matrix over $\mathbb{R}_+$ such that $|C_{pq}| \leq M_{pq}$ for every p, q and if λ is its Perron–Frobenius eigenvalue, then $|\gamma| \leq \lambda$. Moreover, if $|\gamma| = \lambda$, then necessarily $|C| = M$.*

3 The Symbol-Periodicity of Matrices

In this section we introduce the notion of x-periodicity for matrices in the semiring $\mathcal{S}[x]$ of polynomials in the variable x with coefficients in $\mathcal{S}$ and focus more specifically on the case of irreducible matrices.

3.1 The Notion of x-Periodicity

Given a polynomial $F = \sum_k f_k x^k \in \mathcal{S}[x]$, we define the *$x$-period* of F as the integer $d(F) = \mathrm{GCD}\{|h - k| \mid f_h \neq 0 \neq f_k\}$, where we assume $\mathrm{GCD}(\{0\}) = \mathrm{GCD}(\varnothing) = +\infty$. Observe that $d(F) = +\infty$ if and only if $F = 0$ or F is a monomial.

Now consider a finite set Q and a matrix $M : Q \times Q \to \mathcal{S}[x]$. For any index $q \in Q$ and for each integer n we set $d(q,n) = d({M^n}_{qq})$ and we define the x-period of q as the integer $d(q) = \text{GCD}\ \{d(q,n) \mid n \geq 0\}$, assuming that any non-zero element in $\mathbb{N} \cup \{+\infty\}$ divides $+\infty$. Notice that if M is the a-counting matrix of some linear representation, this definition implies that for every index q and for every pair of q-cycles $\mathcal{C}_1$ and $\mathcal{C}_2$ of equal length, $|\mathcal{C}_1|_a - |\mathcal{C}_2|_a$ is a multiple of $d(q)$.

Proposition 2. *If M is an irreducible matrix over $\mathcal{S}[x]$, then all indices have the same x-period.*

Proof. Consider an arbitrary pair of indices p, q. By symmetry, it suffices to prove that $d(p)$ divides $d(q)$, and this again can be proven by showing that $d(p)$ divides $d(q,n)$ for all $n \in \mathbb{N}$. As M is irreducible, there exist two integers s, t such that ${M^s}_{pq} \neq 0 \neq {M^t}_{qp}$. Then the polynomial ${M^{s+t}}_{pp} = \sum_r {M^s}_{pr} {M^t}_{rp} \neq 0$ and for some $k \in \mathbb{N}$ there exists a monomial in ${M^{s+t}}_{pp}$ with exponent k. Therefore, for every exponent h in ${M^n}_{qq}$, the integer $h + k$ appears as an exponent in ${M^{n+s+t}}_{pp}$. This proves that $d(p, n+s+t)$ divides $d(q,n)$ and since $d(p)$ divides $d(p, n+s+t)$, this establishes the result. □

Definition 1. *The* x-period *of an irreducible matrix over $\mathcal{S}[x]$ is the common x-period of its indices.*

Example 1. We compute the x-period of the matrix M over $\mathbb{B}[x]$ corresponding to the state diagram represented in Figure 1. Consider for instance state q_1 and let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two arbitrary q_1-cycles having the same length. Clearly they can be decomposed by using the simple q_1-cycles of the automaton, namely $\ell_1 = q_1 \xrightarrow{a} q_4 \xrightarrow{a} q_1$, $\ell_2 = q_1 \xrightarrow{b} q_2 \xrightarrow{a} q_3 \xrightarrow{b} q_1$. Hence, except for their order, $\mathcal{C}_1$ and $\mathcal{C}_2$ only differ in the number of cycles ℓ_1 and ℓ_2 they contain: for $k = 1, 2$, let $s_k \in \mathbb{Z}$ be the difference between the number of ℓ_k contained in $\mathcal{C}_1$ and the number of ℓ_k contained in $\mathcal{C}_2$. Then, necessarily, $s_1|\ell_1| + s_2|\ell_2| = 0$, that is $2s_1 + 3s_2 = 0$. This implies that $s_1 = 3n$ and $s_2 = -2n$ for some $n \in \mathbb{Z}$. Hence

$$|\mathcal{C}_1|_a - |\mathcal{C}_2|_a = 3n|\ell_1|_a - 2n|\ell_2|_a = 6n - 2n = 4n$$

This proves that 4 is a divisor of the x-period of M. Moreover, both the q_1-cycles ${\ell_1}^3$ and ${\ell_2}^2$ have length equal to 6 and the numbers of occurrences of a differ exactly by 4. Hence, in this case, the x-period of M is exactly 4. □

In the particular case where the entries of the matrix are all linear in x, the matrix decomposes $M = Ax + B$, where A and B are matrices over $\mathcal{S}$; this clearly happens when M is the a-counting matrix of some linear representation. If further M is primitive, the following proposition holds.

Proposition 3. *Let A and B be matrices over $\mathcal{S}$ and set $M = Ax + B$. If M is primitive and $A \neq 0 \neq B$, then the x-period of M is finite.*

Proof. Let q be an arbitrary index and consider the finite family of pairs $\{(n_j, k_j)\}_{j\in J}$ such that $0 \leq k_j \leq n_j \leq m$ where m is the size of M and k_j appears as an exponent in ${M^{n_j}}_{qq}$. Notice that since M is irreducible J is not empty. Since every cycle can be decomposed into elementary cycles all of which of length at most equal to m, the result is proved once we show that $d(q) = +\infty$ implies either $k_j = 0$ for all $j \in J$ or $k_j = n_j$ for all $j \in J$: in the first case we get $A = 0$ while in the second case we have $B = 0$.

Because of equality $M^{\prod_j n_j} = (M^{n_i})^{\prod_{j\neq i} n_j}$, the polynomial ${M^{\prod_j n_j}}_{qq}$ contains the exponent $k_i \prod_{j\neq i} n_j$ for each $i \in J$. Now, suppose by contradiction that $d(q)$ is not finite. This means that all exponents in ${M^{\prod_j n_j}}_{qq}$ are equal to a unique integer h such that $h = k_i \prod_{j\neq i} n_j$ for all $i \in J$. Hence, h must be a multiple of the least common multiple of all products $\prod_{j\neq i} n_j$. Now we have $\mathrm{LCM}\{\prod_{j\neq i} n_j \mid i \in J\} \cdot \mathrm{GCD}\{n_j \mid j \in J\} = \prod_j n_j$ and by the primitivity hypothesis $\mathrm{GCD}\{n_j \mid j \in J\} = 1$ holds. Therefore h is a multiple of $\prod_j n_j$. Thus the conditions $k_j \leq n_j$ leave the only possibilities $k_j = 0$ for all $j \in J$ or $k_j = n_j$ for all $j \in J$. □

Observe that the previous theorem cannot be extended to the case when M is irreducible or when M is a matrix over $\mathcal{S}[x]$ that cannot be written as $Ax + B$ for some matrices A and B over $\mathcal{S}$.

Example 2. The matrix M with entries $M_{11} = M_{22} = 0$, $M_{12} = x$ and $M_{21} = 1$ is irreducible but it is not primitive since it has period 2. It is easy to see that the non-null entries of all its powers are monomials, thus M has infinite x-period. □

Example 3. Consider again Figure 1 and set $M_{2,3} = x^3$. Then we obtain a primitive matrix over $\mathbb{B}[x]$ that cannot be written as $Ax+B$ and it does not have finite x-period. □

3.2 Properties of x-Periodic Matrices

Given a positive integer d, consider the cyclic group $C_d = \{1, g, g^2, \ldots, g^{d-1}\}$ of order d and the semiring $\mathcal{B}_d = \langle \mathcal{P}(C_d), +, \cdot \rangle$ (which is also called $\mathbb{B}$-algebra of the cyclic group) where $\mathcal{P}(C_d)$ denotes the family of all subsets of C_d and for every pair of subsets A, B of C_d we set $A+B = A \cup B$ and $A \cdot B = \{a \cdot b \mid a \in A, b \in B\}$; hence $\varnothing$ is the unit of the sum and $\{1\}$ is the unit of the product. Now, given a positive semiring $\mathcal{S}$, consider the map $\varphi_d : \mathcal{S}[x] \to \mathcal{B}_d$ which associates any polynomial $F = \sum_k f_k x^k \in \mathcal{S}[x]$ with the set $\{g^k \mid f_k \neq 0\} \in \mathcal{B}_d$. Note that since the semiring $\mathcal{S}$ is positive φ_d is a semiring morphism. Intuitively, φ_d associates F with the set of its exponents modulo the integer d. Of course φ_d extends to the semiring of $Q \times Q$-matrices over $\mathcal{S}[x]$ by setting $\varphi_d(T)_{pq} = \varphi_d(T_{pq})$, for every matrix $T : Q \times Q \to \mathcal{S}[x]$ and all $p, q \in Q$. Observe that, since φ_d is a morphism, ${\varphi_d(T)^n}_{pq} = \varphi_d(T^n)_{pq} = \varphi_d({T^n}_{pq})$.

Now, let $M : Q \times Q \to \mathcal{S}[x]$ be an irreducible matrix with finite x-period d. Simply by the definition of d and φ_d, we have that for each $n \in \mathbb{N}$ all non-empty entries $\varphi_d(M^n)_{pp}$ have cardinality 1. The following results also concern the powers of $\varphi_d(M)$.

Proposition 4. *Let M be an irreducible matrix over $\mathcal{S}[x]$ with finite x-period d. Then, for each integer n and each pair of indices p and q, the cardinality of the subset ${\varphi_d(M)^n}_{pq}$ of C_d is not greater than 1; moreover, if $\varphi_d(M)_{qq} \neq \varnothing$, then ${\varphi_d(M)^n}_{qq} = (\varphi_d(M)_{qq})^n$.*

Proof. Let n be an arbitrary integer and p, q an arbitrary pair of indices. By the remarks above we may assume $p \neq q$ and ${M^n}_{pq} \neq 0$. M being irreducible, there exists an integer t such that ${M^t}_{qp} \neq 0$. Note that if B is a non–empty subset of C_d then $|A \cdot B| \geq |A|$ holds for each $A \subseteq C_d$ and ${\varphi_d(M)^{n+t}}_{pp} \supseteq {\varphi_d(M)^n}_{pq} \cdot {\varphi_d(M)^t}_{qp}$. Therefore, since $|{\varphi_d(M)^{n+t}}_{pp}| \leq 1$, we have also $|{\varphi_d(M)^n}_{pq}| \leq 1$. The second statement is proved in a similar way reasoning by induction on n. □

Proposition 5. *Let M be an irreducible matrix over $\mathcal{S}[x]$ with finite x-period d. Then, for each integer n, all non-empty diagonal elements of $\varphi(M)^n$ are equal.*

Proof. Let n be an arbitrary integer and let p, q be an arbitrary pair of indices such that ${M^n}_{pp} \neq 0 \neq {M^n}_{qq}$. By the previous proposition, there exist h, k such that ${\varphi(M)^n}_{pp} = \{g^h\}$ and ${\varphi(M)^n}_{qq} = \{g^k\}$. If t is defined as in the previous proof then the two elements ${\varphi(M)^t}_{qp} \cdot \{g^h\}$ and $\{g^k\} \cdot {\varphi(M)^t}_{qp}$ belong to ${\varphi(M)^{t+n}}_{qp}$; since this subset contains only one element they must be equal and this completes the proof. □

Proposition 6. *Let M be a primitive matrix over $\mathcal{S}[x]$ with finite x-period d. There exists an integer $0 \leq \gamma < d$ such that for each integer n and each index q, if ${M^n}_{qq} \neq 0$, then ${\varphi_d(M)^n}_{qq} = \{g^{\gamma n}\}$.*

Proof. Since M is primitive, there exists an integer t such that ${M^n}_{pq} \neq \varnothing$ for every $n \geq t$ and for every pair of indices p and q. In particular, since $dt + 1 > t$, we have $|{\varphi_d(M^{dt+1}}_{qq})| = 1$ for each q and hence there exists $0 \leq \gamma < d$ such that ${\varphi_d(M)^{dt+1}}_{qq} = \{g^\gamma\}$. Observe that γ does not depend on q, by Proposition 5. Therefore, by Proposition 4, we have

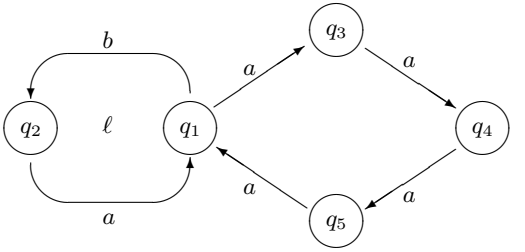
$$\{g^{\gamma n}\} = {\varphi_d(M)^{dtn+n}}_{qq} \supseteq {\varphi_d(M)^{dtn}}_{qq} \cdot {\varphi_d(M)^n}_{qq} = \{1\} \cdot {\varphi_d(M)^n}_{qq}$$

which proves the result. □

If M is the a-counting matrix of a linear representation, then the previous propositions can be interpreted by considering its state diagram. For any pair of states p, q, all paths of the same length starting in p and ending in q have the same number of occurrences of a modulo d. Moreover, if $\mathcal{C}_k$ is a q_k-cycle for $k = 1, 2$ and $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same length, then they also have the same number of occurrences of a modulo d. Finally, if M is primitive, for each cycle ℓ we have $|\ell|_a = \gamma|\ell|$ modulo d for some integer γ.

We conclude this section with an example showing that Proposition 6 cannot be extended to the case when M is irreducible but not primitive.

Example 4. Consider the a-counting matrix M associated with the state diagram of Figure 2. Then M is irreducible with x-period 2, but it is not primitive since also its period equals 2. Consider the path $\ell = q_1 \xrightarrow{b} q_2 \xrightarrow{a} q_1$. We have $|\ell| = 2$ and $|\ell|_a = 1$, hence for any γ, $\gamma|\ell|$ cannot be equal to $|\ell|_a$ modulo 2. Thus, Proposition 6 does not hold in this case.



$$M = \begin{pmatrix} 0 & 1 & x & 0 & 0 \\ x & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & 0 \\ 0 & 0 & 0 & 0 & x \\ x & 0 & 0 & 0 & 0 \end{pmatrix}$$

Fig. 2. State diagram and matrix of Example 4

3.3 Eigenvalues of x-Periodic Matrices

In this section we consider the semiring $\mathbb{R}_+$ of non–negative real numbers and we study the eigenvalues of primitive matrices $M(x)$ over $\mathbb{R}_+[x]$ when x assumes the complex values z such that $|z| = 1$. The next theorem shows how the eigenvalues of $M(z)$ are related to the x-period of the matrix.

Theorem 2. *Let $M(x)$ be a primitive matrix over $\mathbb{R}_+[x]$ with finite x-period d, set $M = M(1)$ and let λ be the Perron-Frobenius eigenvalue of M. Then, for all $z \in \mathbb{C}$ with $|z| = 1$, the following conditions are equivalent:*
1. $M(z)$ and M have the same set of moduli of eigenvalues;
2. If $\lambda(z)$ is an eigenvalue of maximum modulus of $M(z)$, then $|\lambda(z)| = \lambda$;
3. z is a d-th root of unity in $\mathbb{C}$.

Proof (outline). Clearly condition 1) implies condition 2). To prove that condition 2) implies condition 3) we reason by contradiction, that is we assume that z is not a d-th root of unity. It is possible to prove that in this case there exists an integer n such that $|M(z)^n| < M^n$. Therefore we can apply Proposition 1 and prove that λ^n is greater than the modulus of any eigenvalue of $M(z)^n$. In particular we have $\lambda^n > |\lambda(z)|^n$ which contradicts the hypothesis.

Finally we show that condition 3) implies condition 1). The case $d = 1$ is trivial; thus suppose $d > 1$ and assume that z is a $d-$th root of unity. It suffices to prove that if ν is an eigenvalue of M, then νz^γ is an eigenvalue of $M(z)$ with the same multiplicity, where γ is the constant introduced in Proposition 6. To this end, set $\hat{T} = I\nu z^\gamma - M(z)$ and $T = I\nu - M$. We now verify that $\mathrm{Det}\hat{T} = z^{\gamma m}\,\mathrm{Det}\,T$ holds, where m is the size of M. To prove this equality, recall that $\mathrm{Det}\,\hat{T} = \sum_\rho (-1)^{\sigma(\rho)} \hat{T}_{1\rho(1)} \cdots \hat{T}_{m\rho(m)}$. By Proposition 6, since z is a d-th root of 1 in $\mathbb{C}$, we have $\hat{T}_{qq} = (\nu - M_{qq})z^\gamma = z^\gamma T_{qq}$ for each state q and $\hat{T}_{q_0 q_1} \cdots \hat{T}_{q_{s-1} q_0} = z^{\gamma s}\ T_{q_0 q_1} \cdots T_{q_{s-1} q_0}$ for each simple cycle $(q_0, q_1, \ldots, q_{s-1}, q_0)$ of length $s > 1$. Therefore, for each permutation ρ, we get $\hat{T}_{1\rho(1)} \cdots \hat{T}_{m\rho(m)} = z^{\gamma m} \cdot T_{1\rho(1)} \cdots T_{m\rho(m)}$ which concludes the proof. □

Example 5. Let us consider again the primitive matrix of Figure 1. We recall that here $d = 4$; moreover it is easy to see that $\gamma = 3$. Indeed, for each $k = 1, 2$, we have that $|\ell_k| - 3|\ell_k|_a$ is equal to 0 modulo 4. Now consider the characteristic polynomial of the matrix $M(x)$, given by $\chi_x(y) = y^4 - y^2x^2 - yx$ and let ν be a root of χ_1. This implies

that $\chi_1(\nu) = \nu^4 - \nu^2 - \nu = 0$ and hence $-i\nu$ is a root of the polynomial χ_i, $-\nu$ is a root of the polynomial χ_{-1} and $i\nu$ is a root of the polynomial χ_{-i}. This is consistent with Theorem 2, since 1, i, -1 and $-i$ are the four roots of unity. □

4 Local Limit Properties for Pattern Statistics

In this section we turn again our attention to pattern statistics and study the symbol frequency problem in the rational stochastic model under primitivity hypothesis; our goal is to determine the local limit distribution of the corresponding random variable.

Formally, given a rational formal series $r : \{a, b\}^* \to \mathbb{R}_+$, let (ξ, μ, η) be a linear representation of r of size m. Set $A = \mu(a)$, $B = \mu(b)$ and, to avoid trivial cases, assume $A \neq 0 \neq B$. We also set $M(x) = Ax + B$ and $M = M(1)$. Then, consider the probability space of all words of length n in $\{a, b\}^*$ equipped with the probability function given by

$$\mathrm{P}\{w\} = \frac{(r, w)}{\xi^T M^n \eta} = \frac{\xi^T \mu(w) \eta}{\xi^T M^n \eta}$$

for every $w \in \{a, b\}^n$. Now, consider the random variable $Y_n : \{a, b\}^n \to \{0, 1, \ldots, n\}$ such that $Y_n(w) = |w|_a$ for every $w \in \{a, b\}^n$. For sake of brevity, we say that $\{Y_n\}_n$ *counts the occurrences of* a *in the model defined by* r. We study the asymptotic behaviour of Y_n under the hypothesis that the matrix $M(x)$ is primitive, obtaining a local limit distribution that strongly depends on the x-period of $M(x)$.

In the following analysis we consider triples (ξ, μ, η) where both ξ and η have just one non-null entry which is equal to 1: indeed, it turns out that the general case can be reduced to this kind of representation. To show this fact first observe that, for n large enough, all entries of the matrix M^n are strictly positive and thus for every integer $0 \leq k \leq n$ we have

$$\mathrm{P}\{Y_n = k\} = \sum_{|w|=n,\ |w|_a=k} \frac{\xi^T \mu(w) \eta}{\xi^T M^n \eta} = \sum_{p,q=1}^{m} \left(\frac{\xi_p {M^n}_{pq} \eta_q}{\xi^T M^n \eta} \sum_{|w|=n,\ |w|_a=k} \frac{\mu(w)_{pq}}{{M^n}_{pq}} \right). \quad (2)$$

Since $M(x)$ is primitive, by the Perron–Frobenius Theorem, M admits exactly one eigenvalue λ of maximum modulus, which is real and positive. Furthermore, we can associate to λ strictly positive left and right eigenvectors u and v such that $v^T u = 1$ and we know that, as n tends to infinity we have $M^n \sim \lambda^n \cdot uv^T$. Hence

$$\frac{\xi_p {M^n}_{pq} \eta_q}{\xi^T M^n \eta} \sim \frac{\xi_p u_p v_q \eta_q}{(\xi^T u)(v^T \eta)}.$$

Now, let Y_n^{pq} be the random variable associated with the linear representation (e_p, μ, e_q), where e_i denotes the characteristic vector of entry i. Thus, equation (2) can be reformulated as

$$\mathrm{P}\{Y_n = k\} = \sum_{p,q=1}^{m} C_{pq} \cdot \mathrm{P}\{Y_n^{pq} = k\} \quad (3)$$

where C_{pq} are non-negative constants such that $\sum C_{pq} = 1$.

Theorem 3. *Let $r : \{a,b\}^* \to \mathbb{R}_+$ be a rational formal series with a linear representation of the form (e_p, μ, e_q) such that $\mu(a) \neq 0 \neq \mu(b)$. Assume that its a-counting matrix $M(x)$ is primitive and let d be its x-period. Also let $\{Y_n^{pq}\}_n$ count the occurrences of a in the model defined by r. Then, there exist two real constants $0 < \alpha, \beta \leq 1$ and an integer $0 \leq \rho \leq d-1$, all of them depending on $M = M(1)$, such that as n tends to $+\infty$ the relation*

$$\mathrm{P}\{Y_n^{pq} = k\} = \begin{cases} \dfrac{d}{\sqrt{2\pi\alpha n}} \cdot e^{-\frac{(k-\beta n)^2}{2\alpha n}} + o\left(\dfrac{1}{\sqrt{n}}\right) & \text{if } \ k \equiv \rho \bmod d \\ 0 & \text{otherwise} \end{cases}$$

holds uniformly for all integers $0 \leq k \leq n$.

Proof. For the sake of simplicity, for any integer $0 \leq k \leq n$ set

$$p_n(k) = \mathrm{P}\{Y_n^{pq} = k\} = \sum_{|w|=n,\ |w|_a=k} \frac{\mu(w)_{pq}}{M^n{}_{pq}} \ .$$

Observe that by Proposition 4 there exists an integer $0 \leq \rho < d$ such that the number of occurrences of a in paths of the same length starting in p and ending in q are equal to ρ modulo d. Hence, $p_n(k) = 0$ for each $k \not\equiv \rho \bmod d$. Now, consider the smallest integer N such that $Nd \geq n+1$ and apply the N-Discrete Fourier Transform to the array

$$\big(\, p_n(\rho),\ p_n(\rho+d)\,, \ldots\,,\ p_n(\rho+(N-1)d)\big) \in \mathbb{C}^N$$

where the last coefficient is null if $n < \rho + (N-1)d$. We get the following values $f_n(s)$, for integers s such that $-\lceil N/2 \rceil < s \leq \lfloor N/2 \rfloor$:

$$f_n(s) = \sum_{j=0}^{N-1} p_n(\rho + jd) \cdot e^{\frac{2\pi i}{N} sj}$$

Observe that these coefficients are related to the characteristic function $F_n(\theta)$ of the random variable Y_n^{pq}, i.e.

$$F_n(\theta) = \sum_k p_n(k) e^{i\theta k} = \sum_{|w|=n} \frac{\mu(w)_{pq}}{M^n{}_{pq}}\ e^{i\theta |w|_a} \tag{4}$$

Indeed, for any $-\lceil N/2 \rceil < s \leq \lfloor N/2 \rfloor$, we have

$$f_n(s) = e^{-\frac{2\pi i}{Nd} s\rho} \cdot F_n\left(\frac{2\pi s}{Nd}\right)$$

Hence to obtain the values $p_n(\rho + jd)\,, j \in \{0, 1, \ldots\,, N-1\}$, it is sufficient to compute the N-th Inverse Transform of the $f_n(s)$'s. So, we have

$$p_n(\rho+jd) = \frac{1}{N} \sum_{s=-\lceil \frac{N}{2} \rceil + 1}^{\lfloor \frac{N}{2} \rfloor} f_n(s) \cdot e^{-\frac{2\pi i}{N} sj} = \frac{1}{N} \sum_{s=-\lceil \frac{N}{2} \rceil + 1}^{\lfloor \frac{N}{2} \rfloor} F_n\left(\frac{2\pi s}{Nd}\right)\ e^{-\frac{2\pi i}{Nd} s(\rho+jd)} \tag{5}$$

To evaluate $p_n(k)$ we now need asymptotic expressions of the function $F_n(\theta)$ in the interval $\theta \in \left(-\frac{\pi}{d}, \frac{\pi}{d}\right]$. To this aim observe that relations (4) and (1) imply

$$F_n(\theta) = \frac{M(e^{i\theta})^n{}_{pq}}{M^n{}_{pq}}$$

Now, by Theorem 2, we know that the eigenvalues of $M(e^{i\theta})$ are in modulus smaller than λ, for each $\theta \in [\frac{-\pi}{d}, \frac{\pi}{d}]$ different from 0. This property allows us to argue as in [3, Theorem 5]. As a consequence, for each $\theta \in [\frac{-\pi}{d}, \frac{\pi}{d}]$ we can approximate the function $F_n(\theta)$ with the function $\hat{F}_n(\theta) = \exp\,(-\frac{\alpha}{2} n\theta^2 + i\beta n\theta)$, where α and β are positive constants depending on M. Thus, we get

$$\left| F_n(\theta) - \exp\,(-\frac{\alpha}{2} n\theta^2 + i\beta n\theta) \right| = \Delta_n(\theta)$$

where, as n tends to $+\infty$,

$$\Delta_n(\theta) = \begin{cases} \mathrm{O}\,\left(\frac{1}{n^{\varepsilon}}\right) & \text{if } |\theta| \in [0, \frac{2\pi}{(n+1)^{\varepsilon}}] \\ \mathrm{O}\,\left(e^{-\alpha\pi^2 n^{1-2\varepsilon}}\right) & \text{if } |\theta| \in [\frac{2\pi}{(n+1)^{\varepsilon}}, \theta_0] \\ \mathrm{O}\,(\tau^n) & \text{if } |\theta| \in [\theta_0, \frac{\pi}{d}] \end{cases}$$

for some $0 < \theta_0 < \frac{\pi}{d}$, $0 < \tau < 1$, and for every $\frac{1}{3} < \varepsilon < \frac{1}{2}$.

Therefore, to find the approximation for $p_n(k)$ it is sufficient to replace into (5) the values $F_n\left(\frac{2\pi s}{Nd}\right)$ by their approximations $\hat{F}_n\left(\frac{2\pi s}{Nd}\right)$, so getting the following values for each $k \equiv \rho \bmod d$.

$$\hat{p}_n(k) = \frac{1}{N} \sum_{s=-\lceil \frac{N}{2} \rceil + 1}^{\lfloor \frac{N}{2} \rfloor} \hat{F}_n\left(\frac{2\pi s}{Nd}\right) \cdot e^{-\frac{2\pi i}{Nd} ks} \tag{6}$$

Indeed, one can verify that $\left| p_n(k) - \hat{p}_n(k) \right| = \mathrm{O}\,\left(\frac{1}{n^{2\varepsilon}}\right)$ for any $1/3 < \varepsilon < 1/2$ and every $k \equiv \rho \bmod d$. Finally, the sum in (6) can be computed by using the definition of Riemann integral and by means of standard mathematical tools: we find the following approximation which holds as n tends to $+\infty$, uniformly for all $k \equiv \rho \bmod d$:

$$\begin{aligned} \hat{p}_n(k) &\approx \int_{-\frac{1}{2}}^{\frac{1}{2}} \hat{F}_n\left(\frac{2\pi}{d}\, t\right) \cdot e^{-i\frac{2k\pi}{d} t}\, dt = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-\frac{\alpha}{2} n(\frac{2\pi}{d} t)^2} \cdot e^{i\beta n \frac{2\pi}{d} t} \cdot e^{-i\frac{2k\pi}{d} t} dt \\ &\approx \int_{-\infty}^{+\infty} e^{-\frac{2\alpha\pi^2 n}{d^2} t^2} \cdot e^{-i2\pi\frac{k-\beta n}{d} t} dt = \frac{d}{\sqrt{2\pi\alpha n}} \cdot e^{-\frac{(k-\beta n)^2}{2\alpha n}} \end{aligned}$$

□

Applying the theorem above to equation (3), we obtain the following result.

Theorem 4. *Let $r : \{a, b\}^* \to \mathbb{R}_+$ be a rational formal series with a linear representation (ξ, μ, η) such that $\mu(a) \neq 0 \neq \mu(b)$. Assume that its a-counting matrix $M(x)$ is primitive and let d be its x-period. Also let $\{Y_n\}_n$ count the occurrences of a in the model defined by r. Then, there exist two constants $0 < \alpha, \beta \leq 1$ depending on $M = M(1)$ and d constants $C_0, C_1, \ldots, C_{d-1}$ depending on M, ξ and η such that $C_i \geq 0$ for every $i = 0, \ldots, d-1$, $\sum_i C_i = 1$, and as n tends to $+\infty$ the relation*

$$P\,\{Y_n = k\} = \frac{d\, C_{\langle k \rangle_d}}{\sqrt{2\pi\alpha n}} \cdot e^{-\frac{(k-\beta n)^2}{2\alpha n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

holds uniformly for all integers $0 \leq k \leq n$ (here $\langle k \rangle_d = k - \lfloor k/d \rfloor$).

Summarizing the previous results, Theorem 3 states that if the weighted automaton associated with the linear representation has just one initial and one final state, then the local limit distribution has a sort of periodic behaviour: it reduces to 0 everywhere in the domain of possible values except for an integer linear progression of period d, where it approaches a normal density function expanded by a factor d. We also observe that the main terms of mean value and variance of Y_n^{pq} are given by βn and αn, respectively, which do not depend on the initial and final states.

In the general case, when the automaton has more than one initial or final state, by Theorem 4 the required limit distribution is given by a superposition of behaviours of the previous type, all of which have the same main terms of mean value and variance. In the case $d = 1$ the limit probability function of Y_n reduces exactly to a Gaussian density function as already proved in [3]. Such a limit density is the same obtained in the classical DeMoivre–Laplace Local Limit Theorem (see for instance [7, Sec. 12]).

References

1. E. A. Bender. Central and local limit theorems applied to asymptotic enumeration. *Journal of Combinatorial Theory*, 15:91–111, 1973.
2. J. Berstel and C. Reutenauer. *Rational series and their languages*, Springer-Verlag, New York - Heidelberg - Berlin, 1988.
3. A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati. On the number of occurrences of a symbol in words of regular languages. *Theoretical Computer Science*, 302(1-3):431–456, 2003.
4. D. de Falco, M. Goldwurm, and V. Lonati. Frequency of symbol occurrences in simple non-primitive stochastic models. *Proceedings 7th D.L.T. Conference*, Z. Esig and Z. Fülop editors, Lecture Notes in Computer Science, vol. n. 2710, Springer, 2003, 242–253.
5. 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.
6. P. Flajolet and R. Sedgewick. Analytic combinatorics: functional equations, rational and algebraic functions. *Rapport de recherche* n. 4103, INRIA Rocquencourt, January 2001.
7. B.V. Gnedenko. *The theory of probability* (translated by G. Yankovsky). Mir Publishers - Moscow, 1976.
8. 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.
9. W. Kuich and A. Salomaa. *Semirings, automata, languages*. Springer–Verlag, New York Heidelberg Berlin Tokyo, 1986.
10. P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–617, 2002.
11. 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.
12. M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22(4):621–649, 1998.
13. E. Seneta. *Non-negative matrices and Markov chains*, Springer–Verlag, New York Heidelberg Berlin, 1981.
14. M. Waterman. *Introduction to computational biology*, Chapman & Hall, New York, 1995.