Theoretical Informatics and Applications Will be set by the publisher
Informatique Théorique et Applications

PROBABILISTIC MODELS FOR PATTERN STATISTICS *, **

Massimiliano Goldwurm[1] and Roberto Radicioni[1]

Abstract. In this work we study some probabilistic models for the random generation of words over a given alphabet used in the literature in connection with pattern statistics. Our goal is to compare models based on Markovian processes (where the occurrence of a symbol in a given position only depends on a finite number of previous occurrences) and the stochastic models that can generate a word of given length from a regular language under uniform distribution. We present some results that show the differences between these two stochastic models and their relationship with the rational probabilistic measures.

1991 Mathematics Subject Classification. 68Q45, 68Q10, 60J99.

Introduction

In this work we study some probability models for the random generation of words over a given alphabet. Our main purpose is to investigate the difference between the uniform random generation of words of given length in a regular language and the process of generating at random strings of the same length by means of a (homogeneous) Markovian source. Markovian processes on words are widely used in the literature in a large variety of contexts. Here we mainly refer to pattern statistics where they are often assumed as standard models to study the frequency of patterns in a text generated at random [11, 14, 16]. In these models the probability of occurrence of a symbol in a given position only depends on a

Keywords and phrases: pattern statistics, Markov chains, probabilistic automata, rational formal series.

* *Appeared in revised form in R.A.I.R.O. Theoretical Informatics and Applications, vol. 40, 207–225, July 2006.*

** *This work has been supported by Project MIUR PRIN 2005-2007 "Automata and Formal Languages: mathematical and applicative aspects".*

[1] Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39/41, 20135 Milano, Italy; e-mail: goldwurm@dsi.unimi.it & radicion@dsi.unimi.it.

fixed number of previous occurrences. However, as shown in [3], a word of given length n generated uniformly at random in a regular set cannot always be obtained as the result of a Markovian process. This leads to consider more general models, defined by means of weighted finite automata, called rational models in [3]. The corresponding algorithms of random generation easily derive from the well-known procedures for the uniform random generation of strings in regular languages [6,7].

In order to compare these processes, here we study three stochastic models for the random generation of words over a given alphabet, called Markovian, sequential, and rational model, respectively. A Markovian model is essentially defined by a deterministic finite automaton with transitions weighted by probabilities. These probabilistic sources can generate the usual Markovian sequences of arbitrary order, as considered for instance in [16].

The sequential models can be seen as a nondeterministic extension of the Markovian models. They represent a unary version of the so-called stochastic sequential machines [15] and define the rational probability measures on free monoids [9]. Other extensions of the Markovian models have already been considered in the literature ([5]). In Section 2 we prove an aymptotic property of the Markovian models concerning the order of growth of the probabilities associated with periodic words, which does not hold for sequential models. We think this represents a key difference between these two models.

Then we compare sequential and rational models. It is easy to see that any sequential model is also a rational model. However, we prove that this inclusion is strict by studing the (multivariate) generating function of probabilities of symbol occurrences. In particular, we show that for some simple rational models such a function is not holonomic. Note that in the sequential models the same functions are rational.

We also give furher properties of these models. In Section 3 we show that any rational model is equivalent to an absorbing sequential model conditioned to the event of terminating the process in a given transient state. Intuitively, this means that a rational model is related to the transient phase of a sequential model with a unique absorbing state. In Section 4 we study some standard statistics of rational and sequential models in the primitive case, such as (an analogous of) the stationary distribution and the time of first occurrence of a given symbol.

Finally, in the last section we present a result concerning the stochastic models for pattern statistics where the set of patterns is given by a regular language. We show that the frequency of any regular set of patterns in a word generated in a rational model is equal to the frequency of a single symbol in a binary text generated at random by another suitable rational model. A similar result holds for the sequential models and this extends an analogous property proved in [3].

1. Preliminary notions

In this section we recall some basic notions concerning nonnegative matrices, Markov chains and formal series [2, 12, 17, 18].

Let T be a nonnegative square matrix, i.e. $T \in \mathbb{R}_+^{m\times m}$ for some positive $m \in \mathbb{N}$, where $\mathbb{R}_+$ is the semiring of nonnegative real numbers. We recall that T is *primitive* if all entries of T^n are strictly positive, for some $n \in \mathbb{N}$. Moreover, T is called *irreducible* if for every pair of indexes p, q there exists $n \in \mathbb{N}$ such that $T^n_{pq} > 0$ (all over this work T^n_{pq} is the pq-entry of T^n).

The main properties of primitive matrices are given by the Perron–Frobenius theorem [18, Sect.1]. It states that every primitive nonnegative matrix T admits a real positive eigenvalue λ, called Perron–Frobenius eigenvalue of T, which is a simple root of the characteristic polynomial of T, such that $|\nu| < \lambda$ for every eigenvalue $\nu \neq \lambda$ and we can associate with λ strictly positive left and right eigenvectors. Moreover, one can prove that for every $n \in \mathbb{N}$

$$T^n = \lambda^n(vu' + C(n)) \ ,$$

where u' and v are strictly positive left and right eigenvectors [1] of T corresponding to the eigenvalue λ, normed so that $u'v = 1$ and $C(n)$ is a real matrix such that $C(n)_{ij} = \mathrm{O}(\varepsilon^n)$, for some $0 < \varepsilon < 1$ and every pair of indexes i, j. Moreover, for every $n \in \mathbb{N}$, $u'C(n) = \mathbf{0}'$ and $C(n)v = \mathbf{0}$, where $\mathbf{0}' = (0, 0, \cdots, 0)$.

The properties of nonnegative matrices are widely used to study the behaviour of Markov chains [10, 18]. We recall that a real vector $\pi' = (\pi_1, \pi_2, \ldots, \pi_m)$ is stochastic if $0 \leq \pi_i \leq 1$ for every i and $\sum_{i=1}^n \pi_i = 1$. A real matrix P is stochastic if all its rows are stochastic vectors. It is easy to see that the product of two stochastic matrices yields a stochastic matrix. Any stochastic matrix P has eigevalue 1, which admits right eigenvector $\mathbf{1}' = (1, 1, \ldots, 1)$, while $|\gamma| \leq 1$ for every eigevalue γ of P different from 1.

A Markov chain is a sequence of random variables $\{X(n)\}_{n\in N}$ taking on values in a set of states $Q = \{1, 2, \ldots, m\}$ such that there exists a stochastic matrix $P \in [0, 1]^{m\times m}$ whose entries satisfy the relations

$$P_{i_1 j} = \Pr(X(n) = j \mid X(n-1) = i_1, X(n-2) = i_2, \ldots, X(n-k) = i_k)$$

for every $n, k \in \mathbb{N}$, $n \geq k > 0$, and any tuple of states $j, i_1, \ldots, i_k \in Q$. Thus, their probability functions are defined by the matrix P together with the stochastic vector $\pi \in [0, 1]^m$ such that $\Pr(X(0) = i) = \pi_i$ for every $i \in Q$. We represent such a Markov chain by the pair (π, P). Note that $\Pr(X(n) = j) = (\pi' P^n)_j$, for each state j and every $n \in \mathbb{N}$.

To recall the usual classification of states, consider the directed graph G defined by P, where $\{1, 2, \ldots, m\}$ is the set of vertexes and any pair (i, j) is an edge if and only $P_{ij} \neq 0$. A *class* is defined as a subset of $\{1, 2, \ldots, m\}$ that forms a strongly connected component of G. The reduced graph of G is a directed acyclic graph G' whose nodes are the classes and the edges are the pairs of classes (C, D) such that $P_{ij} \neq 0$ for some $i \in C$ and $j \in D$. A class C is said to be *recurrent* if there is no edge in G' from C to a class $D \neq C$. A class is *transient* if it is not recurrent. A state is recurrent (respectively, transient) if it belongs to a recurrent

[1] In this work a vector v is represented as column vector while v' is its transposed (row) vector.

(resp. transient) class. Moreover a state i is *absorbing* if $\{i\}$ is a recurrent class. It is easy to see that the restriction P_C of the matrix P to the entries belonging to a recurrent class C is an irreducible stochastic matrix. On the other hand, one can prove (see for instance [10, Sec.2.5.5]) that if C is a transient class then, as n tends to 0, the entries of P_C^n go to 0 exponentially, i.e.

$$P_C^n = \mathrm{O}(\varepsilon^n) \ , \qquad \text{for some} \quad 0 < \varepsilon < 1 \ . \tag{1}$$

Moreover, it is clear that a state i is absorbing if and only if $P_{ii} = 1$.

We will say that a Markov chain (π, P) is primitive if P is a primitive matrix. In this case its Perron–Frobenius eigenvalue is 1 and we have

$$P^n = \mathbf{1}u' + \mathrm{O}(\varepsilon^n) \ , \tag{2}$$

where u is the left eigenvector of P corresponding to the eigenvalue 1 normed so that $u'\mathbf{1} = 1$ and $0 < \varepsilon < 1$. Observe that $\mathbf{1}u'$ is a stable matrix, i.e. all its rows equal u'; moreover, u is a stochastic vector, called the stationary vector of the chain, and it is the unique stochastic vector such that $u'P = u'$.

The overall behaviour of a Markov chain depends on the form of its reduced graph and on the behaviour of the chain in its reduced classes. For this reason primitive (or irreducible) Markov chains are particularly important. Typical quantities studied in the primitive models are the time of first entrance in a given state and the number of occurrences of a fixed state in the first n steps, and the results obtained in these cases can often be extended to more general models (see for instance the so-called *mixing* Markov chains [10]). A relevant parameter in these analysis is the so-called fundamental matrix Z, defined by $Z = [I - (P - \mathbf{1}u')]^{-1}$; it turns out that Z is related to the moments of the random variables representing the time of first entrance in a given state or the number of entrances in a given state during the first n steps [10, Sect. 4.3].

We now turn our attention to rational formal series. Throughout this work $A = \{a_1, a_2, \ldots, a_s\}$ is a finite alphabet and for every $x \in A^*$, $|x|$ is the length of x while $|x|_{a_i}$ is the number of occurrences of a_i in x. We also denote by A^n the set $\{x \in A^* \mid |x| = n\}$ for every $n \in \mathbb{N}$. A formal series over A with coefficients in $\mathbb{R}_+$ is a function $r : A^* \longrightarrow \mathbb{R}_+$, usually represented in the form $r = \sum_{x \in A^*} r(x) \cdot x$, where $r(x)$ denotes the value of r at $x \in A^*$. We denote by $\mathbb{R}_+\langle\langle A \rangle\rangle$ the family of all formal series over A with coefficients in $\mathbb{R}_+$. This set forms a semiring with respect to the traditional operations of sum and Cauchy product.

A classical tool to transform formal series into traditional generating functions is the canonical monoid morphism $\Phi : A^* \longrightarrow A^{\oplus}$, where $A^{\oplus}$ is the totally commutative monoid over the alphabet A whose elements can be represented in the form $\underline{a}^i = a_1^{i_1} a_2^{i_2} \cdots a_s^{i_s}$, $i \in \mathbb{N}^s$. For every $x \in A^*$ the value $\Phi(x)$ is given by $\Phi(x) = \underline{a}^i$ where $i_j = |x|_{a_j}$ for every $j = 1, 2, \ldots, s$. Clearly Φ extends to a semiring morphism from $\mathbb{R}_+\langle\langle A \rangle\rangle$ towards the traditional ring $\mathbb{R}[\![A]\!]$ of formal power series in the commutative variables $a_1, a_2, \ldots, a_s$ with real coefficients. Therefore Φ can be

considered as a function $\Phi : \mathbb{R}_+\langle\langle A\rangle\rangle \longrightarrow \mathbb{R}[\![A]\!]$ where, for each $r \in \mathbb{R}_+\langle\langle A\rangle\rangle$,

$$\Phi(r) = \sum_{i\in\mathbb{N}^s} f(i)\ \underline{a}^i, \qquad \text{and} \qquad f(i) = \sum_{|x|_{a_1}=i_1,\ldots,|x|_{a_s}=i_s} r(x)\ .$$

We recall that an element $g \in \mathbb{R}[\![A]\!]$ is rational if there exist two polynomials $p, q \in \mathbb{R}[A]$ such that $g = pq^{-1}$; moreover, g is holonomic if, for each $a \in A$, g is solution of a linear partial differential equation in a with polynomial coefficients [1]. All algebraic series are holonomic as well as several types of transcendental series [13, 19]. Further, we recall that the sequence of coefficients associated with a holonomic series is solution of a linear recurrence relation with polynomial coefficients [13].

A formal series $r \in \mathbb{R}_+\langle\langle A\rangle\rangle$ is called *rational* if it admits a *linear representation*, that is a triple $\langle\xi, \mu, \eta\rangle$ where, for some integer $m > 0$, ξ and η are (column) vectors in $\mathbb{R}_+^m$ and $\mu : A^* \longrightarrow \mathbb{R}_+^{m\times m}$ is a monoid morphism, such that $r(x) = \xi'\mu(x)\ \eta$ holds for each $x \in A^*$. We say that m is the *size* of the representation. Such a triple $\langle\xi, \mu, \eta\rangle$ can be interpreted as a weighted nondeterministic automaton, where the set of states is given by $\{1, 2, \ldots, m\}$ and the transitions, the initial and the final states are assigned weights in $\mathbb{R}_+$ by μ, ξ and η, respectively. For each $a \in A$ the matrix $\mu(a)$ represents the weights of all transitions of the automaton labeled by A. To avoid redundancy it is convenient to assume that $\langle\xi, \mu, \eta\rangle$ is trim (meaning that all indexes are used to define the series), i.e. for every index i there are two indexes p, q and two words $x, y \in A^*$ such that $\xi_p\mu(x)_{pi} \neq 0$ and $\mu(y)_{iq}\eta_q \neq 0$. We say that $\langle\xi, \mu, \eta\rangle$ is *primitive* if $T = \sum_{a\in A} \mu(a)$ is a primitive matrix. Clearly, if $r \in \mathbb{R}_+\langle\langle A\rangle\rangle$ is rational then $\Phi(r)$ is rational in $\mathbb{R}[\![A]\!]$.

2. Stochastic models on words

Inspired by the classical Bernoullian and Markovian processes analogous models have been proposed in the literature to study probability measures on free monoids [9,15]. Incidentally, we recall that a *probability measure* on A^* is a map $f : A^* \longrightarrow [0, 1]$ such that $f(\epsilon) = 1$ and $\sum_{a\in A} f(xa) = f(x)$ for every $x \in A^*$ [9].

For our purpose, we may consider a probabilistic model over A as a formalism to define a probability function on the set A^n for every integer $n > 0$; moreover, it is naturally equipped with an effective procedure to generate on input n a word in A^n with the prescribed probability. In this section we study three types of probabilistic models called, respectively, Markovian, sequential and rational. Our main purpose is to stress the differences among these models. For instance it will be clear that for Markovian and sequential models the associated probability function is a probability measure on a free monoid, while the same property is not always true rational models.

The simplest probabilistic model on words is the well-known Bernoullian model. A *Bernoullian* model $\mathcal{B}$ over A is defined by a function $p : A \to [0, 1]$ such that $\sum_{a\in A} p(a) = 1$. A word $x \in A^+$ is generated in this model by choosing each letter of x under the distribution defined by p independently of one another. Thus, the probability of $x = x_1x_2\cdots x_n$, where $x_i \in A$ for each i, is given by $\Pr_{\mathcal{B}}(x) =$

$p(x_1)p(x_2)\cdots p(x_n)$, which clearly defines a probability function over A^n for every integer $n > 0$.

2.1. Markovian models

A *Markovian* model over A is defined as a pair $\mathcal{M} = (\pi, M)$ where, for some integer $k > 0$, $\pi \in [0,1]^k$ is a stochastic vector and M is a function $M : A \to [0,1]^{k\times k}$ such that the matrix $T = \sum_{a\in A} M(a)$ is stochastic and for every $a \in A$, each row of $M(a)$ has at most one non-null entry.

The probability of a word $x = x_1x_2\cdots x_n$, where $x_i \in A$ for each $i = 1, 2, \ldots, n$, is given by

$$\Pr_{\mathcal{M}}(x) = \pi' M(x_1)M(x_2)\cdots M(x_n)\mathbf{1}.$$

Since both π and T are stochastic arrays, $\Pr_{\mathcal{M}}$ defines a probability function over A^n for each positive integer n. Note that (π, T) is a Markov chain over the set of states $Q = \{1, 2, \ldots, k\}$, which we may call the *underlying* Markov chain of $\mathcal{M}$.

Note that every Bernoullian model is a Markovian model. Moreover, the pair $\mathcal{M} = (\pi, M)$ defines a deterministic finite state automaton where transitions are weighted by probabilities: the set of states is Q, the transition function $\delta_{\mathcal{M}} : Q \times A \longrightarrow Q \cup \{\perp\}$ is defined so that for every $i \in Q$ and every $a \in A$, $\delta_{\mathcal{M}}(i, a) = j$ if $M(a)_{ij} \neq 0$, and the same value $M(a)_{ij}$ is the weight of the transition, while $\delta_{\mathcal{M}}(i, a) = \perp$ if $M(a)_{ij} = 0$. Clearly, $\delta_{\mathcal{M}}$ can be extended to all words in A^*. Moreover, the sum of weights of all transitions outgoing from any state equals 1. Since the automaton is deterministic, for every word $x = x_1x_2\cdots x_n$ and every $i_0 \in Q$ there exists at most one path labeled by x starting from i_0. The corresponding weight is given by the value $P_{i_0}(x) = \pi_{i_0} \cdot \overline{m}_{i_0}(x)$, where $\overline{m}_{i_0}(x) = M(x_1)_{i_0i_1}M(x_2)_{i_1i_2}\cdots M(x_n)_{i_{n-1}i_n}$ if there exist $i_1, \ldots, i_n \in Q$ such that $\delta(i_{j-1}, x_j) = i_j$ for each $j = 1, \ldots, n$, while $\overline{m}_{i_0}(x) = 0$ otherwise. As a consequence, the probability of x can be expressed by $\Pr_{\mathcal{M}}(x) = \sum_{i=1}^{k} P_i(x)$.

The following lemma gives an asymptotic property of the probabilities defined in Markovian models. Here we use the symbol Θ to represent the order of growth of sequences: for a pair of sequences $\{f_n\}$, $\{g_n\}$, both included in $\mathbb{R}_+$, the equality $f_n = \Theta(g_n)$ means that for some positive constants c_1, c_2, the relation $c_1g_n \leq f_n \leq c_2g_n$ holds for any n large enough.

Lemma 2.1. *Let $\mathcal{M} = (\pi, M)$ be a Markovian model of size k over the alphabet A and let $x \in A^+$. Then, there exists $0 \leq \beta \leq 1$ such that, as n tends to $+\infty$,*

$$\Pr_{\mathcal{M}}(x^n) \;=\; \Theta(\beta^n).$$

Proof. If $\Pr_{\mathcal{M}}\{x^k\} = 0$, then the property holds true with $\beta = 0$. Now, assume $\Pr_{\mathcal{M}}\{x^k\} \neq 0$ and set $l = |x|$. Then there exists a path $i_0, i_1, \ldots, i_{kl} \in Q$ such that

$$P_{i_0}(x^k) = \pi_{i_0} M(x_1)_{i_0i_1} M(x_2)_{i_1i_2} \cdot \ldots \cdot M(x_{kl})_{i_{kl-1}i_{kl}} \neq 0,$$

where $x_1x_2\cdots x_{kl} = x^k$. For the pigeonhole principle, in the sequence of states $i_0, i_l, i_{2l}, \ldots, i_{kl}$ there are at least two equal elements. Consider the smallest integers a, b, with $0 \leq a \leq k-1$ and $1 \leq b \leq k$, such that $i_{al} = i_{(a+b)l}$. Then in the

automaton there is a cycle starting from i_{al} and labelled by x^b. As a consequence, for n large enough, we can write x^n as

$$x^n \;=\; x^a \left(x^b\right)^{\lfloor \frac{n-a}{b} \rfloor} x_f,$$

where $x_f = x^{[n-a]_b}$. Then, denoting $\overline{m}_{i_0}(x^a)$ by p_0, $\overline{m}_{i_{al}}(x^b)$ by p and $\overline{m}_{i_{al}}(x_f)$ by p_f, we have

$$P_{i_0}(x^n) \;=\; \pi_{i_0} p_0 p^{\lfloor \frac{n-a}{b} \rfloor} p_f.$$

Now, let $\hat{p_f}$ be the smallest value assumed by p_f for any n, i.e. $\hat{p_f} = \overline{m}_{i_{al}}(x^{b-1})$. Then,

$$\pi_{i_0} p_0 p^{\frac{n-a}{b}} \hat{p_f} \;\leq\; P_{i_0}(x^n) \;\leq\; \pi_{i_0} p_0 p^{\frac{n-a}{b}-1},$$

whence

$$\pi_{i_0} p_0 p^{-\frac{a}{b}} \hat{p_f} \left(p^{\frac{1}{b}}\right)^n \;\leq\; P_{i_0}(x^n) \;\leq\; \pi_{i_0} p_0 p^{-\left(\frac{a}{b}+1\right)} \left(p^{\frac{1}{b}}\right)^n.$$

Therefore, $P_{i_0}(x^n) = \Theta(\alpha^n)$, with $\alpha = p^{\frac{1}{b}}$. Since the initial state i_0 may assume k different values, reasoning as above we can determine k constants $\alpha_1, \ldots, \alpha_k \in [0,1]$ (which are not all null) such that $P_j(x^n) = \Theta(\alpha_j^n)$ for $j = 1, \ldots, k$. Thus, considering $\beta = \max_{1 \leq j \leq k}\{\alpha_j\}$, we have

$$\Pr_{\mathcal{M}}(x^n) \;=\; \Theta(\alpha_1^n) + \Theta(\alpha_2^n) + \cdots + \Theta(\alpha_k^n) \;=\; \Theta(\beta^n).$$

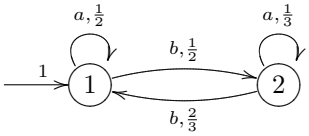
□

The previous lemma plays a role similar to classical pumping lemma in formal languages in the sense that it can be used to show that a given probabilistic model is not Markovian.

Corollary 2.2. *Given a map $\mathcal{P} : A^* \to [0,1]$, assume that for each $n \in \mathbb{N}$ the restriction of $\mathcal{P}$ to the set A^n is a probability function. If there exists a word $x \in A^+$ such that $\mathcal{P}(x^n)$ is not of the order $\Theta(\beta^n)$ for any constant $\beta \geq 0$, then there is no Markovian model $\mathcal{M}$ over A such that $Pr_{\mathcal{M}} = \mathcal{P}$.*

The previous models can generate the traditional Markov sequences of order m over A where the probability of the next symbol occurrence only depends on the previous m symbols. To define these sources in our context we say that a Markovian model $\mathcal{M}$ over A is of *order* m if for every word $w \in A^m$ either there exists j such that $\delta_{\mathcal{M}}(i,w) = j$ for every $i \in Q$ or $\delta_{\mathcal{M}}(i,w) = \perp$ for every $i \in Q$, and m is the smallest integer with such a property.

A relevant case occurs when $m = 1$. In this case, the set of states Q can be reduced to A and $\Pr_{\mathcal{M}}$ is called Markov probability measure in [9]. Also observe that there exist Markovian models that are not of order m for any $m \in N$. For instance, if $\mathcal{M}$ is defined by the following (weighted) finite automaton, then $\delta_{\mathcal{M}}(1, a^n b) \neq \delta_{\mathcal{M}}(2, a^n b)$ for every $n \in \mathbb{N}$.



2.2. Sequential models

A natural extension of the previous model can be obtained by allowing nondeterminism in the corresponding finite state device. In this way the model corresponds to a stochastic sequential machine, as defined in [15], with a unary input alphabet. Moreover, it is characterized by the rational probability measures, i.e. the probability measures on A^* that are rational formal series in $\mathbb{R}_+\langle\langle A\rangle\rangle$ [9].

Formally, we define a *sequential* stochastic model over A as a pair $\mathcal{Q} = (\pi, M)$ where $\pi \in [0,1]^k$ is a stochastic vector and M is a function $M : A \to [0,1]^{k\times k}$ such that $T = \sum_{a\in A} M(a)$ is a stochastic matrix.

As in the previous model, M defines a monoid morphism between A^* and $[0,1]^{k\times k}$. Analogously, the probability of a word $x = x_1x_2\cdots x_n \in A^*$ is

$$\begin{aligned}\Pr_{\mathcal{Q}}(x) &= \pi' M(x_1)M(x_2)\cdots M(x_n)\mathbf{1} = \\ &= \sum_{i_0,i_1,\ldots i_n\in\{1,2,\ldots,k\}} \pi_{i_0} M(x_1)_{i_0i_1} M(x_2)_{i_1i_2}\cdots M(x_n)_{i_{n-1}i_n} .\end{aligned}$$

As in the previous case, (π, T) is the underlying Markov chain and $\Pr_{\mathcal{Q}}$ is a rational formal series taking on values in $[0,1]$. It admits the linear representation $\langle\pi, M, \mathbf{1}\rangle$ and defines a probability function over A^n for every positive integer n. Furthermore, it is easy to see that $\Pr_{\mathcal{Q}}$ is a probability measure on A^* and hence it is a rational probability measure; on the other hand, for every rational probability measure f on A^* there exists a sequential model $\mathcal{Q}$ such that $f = \Pr_{\mathcal{Q}}$ [9].

The pair $\mathcal{Q} = (\pi, M)$ can be interpreted as a finite state automaton equipped with probabilities associated with transitions, the main difference now is that the automaton is nondeterministic. For any $a \in A$, every nonnull entry $M(a)_{ij}$ is the weight of the transition from i to j labeled by a and, for every word x, $\Pr_{\mathcal{Q}}(x)$ is the sum of the weights of all paths labeled by x in the corresponding transition diagram.

Since $\Pr_{\mathcal{Q}}$ is a rational formal series, the series $F_{\mathcal{Q}} = \Phi(\Pr_{\mathcal{Q}})$ is rational in $\mathbb{R}[\![A]\!]$ and it can be given by $F_{\mathcal{Q}} = \pi'(I - \sum_{a\in A} P(a)a)^{-1}\mathbf{1}$. Note that $F_{\mathcal{Q}}$ is the generating function of the probabilities of occurrences of symbols in A. In other words,

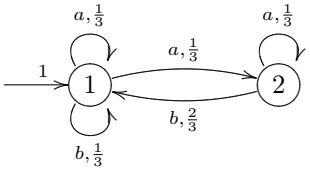
$$F_{\mathcal{Q}} = \sum_{i\in\mathbb{N}^s} F_{\mathcal{Q}}(i)\ \underline{a}^i,$$

where, for all $i \in \mathbb{N}^s$ such that $i_1 + \cdots + i_s = n$, $F_{\mathcal{Q}}(i) = \Pr_{\mathcal{Q}}\{x \in A^n : |x|_{a_1} = i_1, \ldots, |x|_{a_s} = i_s\}$.

Example 2.3. Consider the sequential model $\mathcal{Q} = (\pi, M)$ over the alphabet $A = \{a, b\}$, where $\pi = (1, 0)$,

$$M(a) \;=\; \begin{bmatrix} 1/3 & 1/3 \\ 0 & 1/3 \end{bmatrix} \qquad \text{and} \qquad M(b) \;=\; \begin{bmatrix} 1/3 & 0 \\ 2/3 & 0 \end{bmatrix} ,$$

defining the following weighted automaton:



Here $F_{\mathcal{Q}}(a, b) = 9(a^2 - 6a - ab - 3b + 9)^{-1}$ and, for every integer $n \geq 1$, we have $\Pr_{\mathcal{Q}}(a^n) = (n + 1)3^{-n}$. Hence, by Corollary 2.2, $\Pr_{\mathcal{Q}}$ cannot be the probability function of any Markovian model.

In passing, we observe that sequential models are equivalent to Markov chains with states labeled by symbols of the alphabet. To state this equivalence precisely, let us define a A-colored Markov chain as a triple $\mathrm{CM} = (\xi, S, e)$, where (ξ, S) is a Markov chain over a set of states Q and $e : Q \longrightarrow A$ is a labelling function. This model associates each word $x = x_1 x_2 \cdots x_n$ in A^n with the probability

$$\Pr_{\mathrm{CM}}(x) \;=\; \sum_{\substack{i_0, i_1, \ldots, i_n \in Q \\ e(i_j) = x_j,\ j = 1, \ldots, n}} \pi_{i_0} S_{i_0 i_1} S_{i_1 i_2} \cdots S_{i_{n-1} i_n}.$$

By standard argument, one can prove that, for every sequential model $\mathcal{Q}$ over A, there exists a A-colored Markov chain CM such that, for every word $x \in A^*$, $\Pr_{\mathrm{CM}}(x) = \Pr_{\mathcal{Q}}(x)$; viceversa, for every A-colored Markov chain CM there exists a sequential model $\mathcal{Q}$ over A such that $\Pr_{\mathcal{Q}}(x) = \Pr_{\mathrm{CM}}(x)$, for every $x \in A^*$.

2.3. Rational models

Consider a rational formal series $r \in \mathbb{R}_+\langle\langle A \rangle\rangle$ and, for every positive integer n, assume $r(w) \neq 0$ for some $w \in A^n$. Then r defines a probability function over A^n, given by

$$\Pr_r(x) = \frac{r(x)}{\sum_{w \in A^n} r(w)} \qquad \text{for every } x \in A^n \ . \tag{3}$$

Observe that if r is the characteristic series χ_L of a regular language $L \subseteq A^*$, then $\Pr_r$ represents the uniform probability function over $L \cap A^n$, for each n.

Since r is rational it admits a linear representation (ξ, μ, η) and hence

$$\Pr_r(x) = \frac{\xi' \mu(x) \eta}{\xi' T^n \eta} \qquad \text{for every } x \in A^n, \tag{4}$$

where $T = \sum_{a \in A} \mu(a)$. Also observe that $\Pr_r$ is a sort of Hadamard division of two rational formal series.

It is clear that every sequential model over A is a rational model over the same alphabet. Moreover, also in this case we can define $F_r = \Phi(\mathrm{Pr}_r)$ and we have

$$F_r = \sum_{n=0}^{+\infty} \sum_{i_1+\cdots+i_s=n} F_r(i)\ \underline{a}^i, \qquad \text{where} \qquad F_r(i) = \sum_{|x|_{a_1}=i_1,\ldots,|x|_{a_s}=i_s} \mathrm{Pr}_r(x)\ .$$

Theorem 2.4. *There exists a rational series $r \in \mathbb{R}_+\langle\langle A \rangle\rangle$ such that F_r is not holonomic.*

Proof. Consider the rational series r over the alphabet $\{a, b\}$, defined by the linear representation (ξ, μ, η) such that

$$\xi \ = \ \begin{pmatrix} 1 \\ 0 \end{pmatrix}\ , \quad \mu(a) \ = \ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\ , \quad \mu(b) \ = \ \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}\ , \quad \eta \ = \ \begin{pmatrix} 1 \\ 0 \end{pmatrix}\ .$$

In this case, r is the characteristic series of the language $L = (b + ab)^*$ and we have

$$\mathrm{Pr}_r(x) = \begin{cases} \frac{1}{f_{|x|+1}} & \text{if } x \in L \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Phi(r) = F_r(y, z) = \sum_{k,n\geq 0} \sum_{|x|_a=k,\ |x|_b=n} \mathrm{Pr}_r(x) y^k z^n,$$

where f_n is the nth Fibonacci number. Then, the function $\Phi(r) = F_r(y, z)$ is holonomic only if its section

$$F_r(0, z) \ = \ \sum_{n\geq 0} \mathrm{Pr}_r(b^n) z^n \ = \ \sum_{n\geq 0} \frac{1}{f_{n+1}} z^n$$

is holonomic (see Prop 2.5 in [13]). Now, assume by contradiction that $F_r(0, z)$ is holonomic. Then there exist $k + 1$ polynomials $p_0(n), p_1(n), \ldots, p_k(n)$ satisfying the following linear recurrence equation for every $n \in \mathbb{N}$ large enough:

$$\frac{p_0(n)}{f_n} \ = \ -\sum_{i=1}^{k} \frac{p_i(n)}{f_{n-i}}, \qquad \text{with} \quad p_0(n), p_k(n) \neq 0. \tag{5}$$

We can also assume that $k + 1$ is the smallest number of polynomials satisfying the previous property. The properties of the Fibonacci numbers allow us to write $f_n = f_{n+2} - f_{n+1}$. Thus, each f_{n-i} can be rewritten as $c_i f_n + d_i f_{n-1}$ for suitable $c_i, d_i \in \mathbb{Z}$. Thus, we get

$$\frac{p_0(n)}{f_n} \ = \ \sum_{i=1}^{k} \frac{p_i(n)}{c_i f_n + d_i f_{n-1}}\ .$$

Multiplying both members by f_n, we obtain

$$p_0(n) \;=\; -\sum_{i=1}^{k} \frac{p_i(n)}{c_i + d_i \frac{f_{n-1}}{f_n}}. \tag{6}$$

Since $f_n = \alpha^n + \beta^n$, with $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$, we have

$$\frac{f_{n-1}}{f_n} \;=\; \frac{\alpha^{n-1} + \beta^{n-1}}{\alpha^n + \beta^n} \;=\; \frac{1}{\alpha} + O(\epsilon^n),$$

for some ϵ such that $0 < \epsilon < 1$. Therefore, replacing $c_i + d_i/\alpha$ by e_i, Equation 6 becomes

$$\begin{aligned} p_0(n) \;&=\; -\sum_{i=1}^{k} \frac{p_i(n)}{e_i + d_i O(\epsilon^n)} \;=\; -\sum_{i=1}^{k} \frac{p_i(n)}{e_i \left(1 + \frac{d_i}{e_i} O(\epsilon^n)\right)} \;= \\ &=\; -\sum_{i=1}^{k} \frac{p_i(n)}{e_i} \cdot \left(1 + \frac{d_i}{e_i} O(\epsilon^n) + O(\epsilon^{2n})\right) = \\ &=\; -\sum_{i=1}^{k} \frac{p_i(n)}{e_i} + \left(\sum_{i=1}^{k} \frac{d_i}{e_i} p_i(n)\right) O(\epsilon^n). \end{aligned}$$

Since $p_0(n)$ is a polynomial, this relation is true only if the last sum is identically null. Then, the equation $\sum_{i=1}^{k} \frac{d_i}{e_i} p_i(n) = 0$ must hold, which implies that $p_k(n)$ linearly depends on $p_1(n), \ldots, p_{k-1}(n)$. As a consequence, in Equation 5, $p_k(n)$ can be replaced by a linear combination of $p_1(n), \ldots, p_{k-1}(n)$, obtaining an equation of order $k-1$, which is a contradiction because of the choice of k. □

Proposition 2.5. *The chain of inclusions*

Bernoullian models $\subset$ *Markovian models* $\subset$ *Sequential models* $\subset$ *Rational models*

is strict.

Proof. It is clear that the Markovian models cannot be simulated by Bernoullian models. The second inclusion is strict because of Corollary 2.2 and Example 2.3. Finally, the probability function of sequential models is obviously rational and Theorem 2.4 shows that there exist rational models r such that F_r is not rational and hence they cannot be simulated by sequential models. As a consequence also the last inclusion is strict. □

3. Rational models as conditional sequential models

In this section we show how any rational model can be obtained from an absorbing sequential model by conditioning the process to terminate in a given transient

state. The new model is obtained by adding two states, one for simulating the array η of final weights, the other for equaling the total weight of the transitions from each state. Intuitively, this result associates the rational models with the behaviour of sequential models during the transient phase. We recall that there exist quite natural processes that can be represented by such a behaviour [10].

For our purpose, let us introduce some further notations. Consider a sequential model $\mathcal{Q} = (\pi, M)$ over an alphabet A and let $Q = \{1, 2, \ldots, k\}$ be its set of states. For any word $x \in A^n$ and any $q \in Q$ we denote by $\Pr_{\mathcal{Q}q}(x)$ the probability of generating x after n steps and terminating in q. This value, for $x = x_1 x_2 \cdots x_n$, is given by

$$\Pr_{\mathcal{Q}q}(x) = \pi' \; M(x_1)M(x_2)\cdots M(x_n) \; e_q \;,$$

where e_q is the characteristic vector of q (i.e. the vector having 1 in the entry corresponding to state q and 0 elsewhere).

On the other hand, we denote by $\Pr_{\mathcal{Q}|q}(x)$ the probability of generating x after n steps *conditioned* to the event of terminating the process in q. We can express this measure as

$$\Pr_{\mathcal{Q}|q}(x) = \frac{\Pr_{\mathcal{Q}q}(x)}{\pi' \; T^n \; e_q} = \frac{\pi' \; M(x_1)M(x_2)\cdots M(x_n) \; e_q}{\pi' \; T^n \; e_q},$$

where $T = \sum_{a \in A} M(a)$.

We say that the sequential model $\mathcal{Q}$ is *absorbing* if in the underlying Markov chain there exists just one absorbing state $q \in Q$ and all states $p \neq q$ are transient.

Theorem 3.1. *For every rational series $r \in \mathbb{R}_+\langle\langle A \rangle\rangle$ admitting a linear representation of size k there exists an absorbing sequential model $\mathcal{Q}$ over A of $k+2$ states such that for every $x \in A^+$*

$$Pr_r(x) = Pr_{\mathcal{Q}|q}(x) \;,$$

where q is the absorbing state of $\mathcal{Q}$.

Proof. Let $\langle \xi, \mu, \eta \rangle$ be a linear representation of r of size k and let l be the cardinality of A. We construct an absorbing sequential model $\mathcal{Q} = (\pi, M)$, of dimension $k+2$, where the state $k+2$ is absorbing and the conditional probability $\Pr_{\mathcal{Q}|k+1}$ is equal to $\Pr_r$. The components of the initial vector π are $\pi_i = \xi_i/\overline{\xi}$ if $1 \leq i \leq k$ and zero otherwise, where $\overline{\xi} = \xi' \mathbf{1}$.

In order to define the matrices $M(a)$ for every $a \in A$, let $R(a)$ be the k-vector $\mu(a)(\mathbf{1} + \eta)$ and set

$$u \;=\; \max\{R(a)_i \mid a \in A,\; i = 1, 2, \ldots, k\}\,.$$

For every $i, j \in \{1, \ldots, k\}$, we define

$$M(a)_{ij} = \frac{\mu(a)_{ij}}{lu}, \qquad M(a)_{i\;k+1} = \frac{(\mu(a)\eta)_i}{lu}, \qquad M(a)_{i\;k+2} = \frac{u - R(a)_i}{lu}.$$

The components $M(a)_{k+1\ j}$ and $M(a)_{k+2\ j}$ are equal to l^{-1} if $j = k+2$, and zero otherwise. It is easy to see that the matrix $N = \sum_{a\in A} M(a)$ is stochastic. Indeed, for all $i = 1, \ldots, k$ we have

$$\sum_{j=1}^{k+2} N_{ij} = \sum_{a\in A}\sum_{j=1}^{k+2} M(a)_{ij} = \frac{1}{lu}\sum_{a\in A}\left[(\mu(a)\mathbf{1})_i + (\mu(a)\eta)_i + u - R(a)_i\right] = 1,$$

while $\sum_{j=1}^{k+2} N_{k+1\ j} = \sum_{j=1}^{k+2} N_{k+2\ j} = 1$. Now, by the form of the matrices $M(a)$, the generating process of a word $x = x_1x_2\cdots x_n$ terminating in state $k+1$ never transits through state $k+2$ and has probability

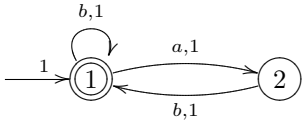
$$\Pr_{\mathcal{Q},k+1}(x) \;=\; \sum_{i=1}^{k} \frac{\xi_i}{\overline{\xi}}\;\frac{\mu(x_1)\cdots\mu(x_{n-1})\;\mu(x_n)\;\eta}{(lu)^n} \;=\; \frac{\xi'\;\mu(x)\;\eta}{\overline{\xi}\;(lu)^n}\;.$$

Moreover, denoting with $T = \sum_{a\in A}\mu(a)$ the stochastic matrix associated to μ, the probability of terminating in state $k+1$ after generating a word of length n is $\pi N^n e_{k+1} = (\overline{\xi}(lu)^n)^{-1}\xi' T^n\eta$, and then

$$\Pr_{\mathcal{Q}|k+1}(x) \;=\; \frac{\Pr_{\mathcal{Q},k+1}(x)}{\pi'\;N^n\;e_{k+1}} \;=\; \frac{\xi'\;\mu(x)\;\eta}{\xi'\;T^n\;\eta} \;=\; \Pr_r(x).$$

□

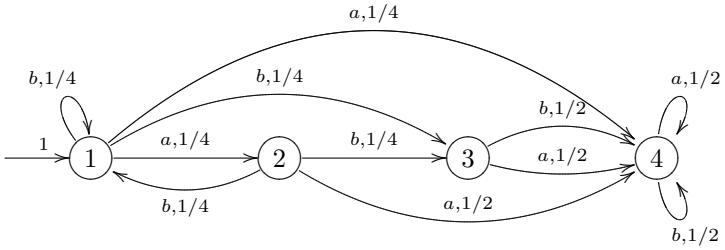
Example 3.2. Consider the linear representation (ξ, μ, η) and the series r defined in the proof of Theorem 2.4. We recall that r is the characteristic series of $L = (b + ab)^*$, while (ξ, μ, η) corresponds to the following automaton:



Applying Theorem 3.1 we get the sequential model $\mathcal{Q} = (\pi, M)$ such that

$$\pi = \begin{pmatrix} 1\\0\\0\\0 \end{pmatrix} \quad M(a) = \begin{pmatrix} 0 & 1/4 & 0 & 1/4\\ 0 & 0 & 0 & 1/2\\ 0 & 0 & 0 & 1/2\\ 0 & 0 & 0 & 1/2 \end{pmatrix} \quad M(b) = \begin{pmatrix} 1/4 & 0 & 1/4 & 0\\ 1/4 & 0 & 1/4 & 0\\ 0 & 0 & 0 & 1/2\\ 0 & 0 & 0 & 1/2 \end{pmatrix}.$$

The corresponding weighted automaton is described by the following picture:



As a matter of fact, it is not difficult to see that $\Pr_{\mathcal{Q}|3}(x) = \Pr_r(x)$, for every $x \in A^*$.

4. Primitive models

As in the theory of Markov chains, the asymptotic behaviour of our models depends on the properties of the primitive cases. In this section we consider sequential and rational models under a primitivity hypothesis and study some standard parameters defined with respect to a symbol $a \in A$, as for instance, the time of first occurrence of a and its number of occurrences in a word generated by the process. We give some properties easily deducible from the definitions and extend certain results appeared in the literature.

4.1. Limit probability of symbol occurrence

A sequential model $\mathcal{Q} = (\pi, M)$ over A is primitive if the matrix $T = \sum_{a\in A} M(a)$ is primitive. As a consequence, by equation (2), $T^n = \mathbf{1}u' + \mathrm{O}(\varepsilon^n)$ where u is the left eigenvector of T corresponding to the eigenvalue 1 normed so that $u'\mathbf{1} = 1$ and $0 < \varepsilon < 1$. Let $x_1x_2\cdots x_n \in A^n$ be a word generated by $\mathcal{Q}$. Then, for every $a \in A$, we have

$$\Pr(x_n = a) = \pi' T^{n-1} M(a)\mathbf{1} = u'M(a)\mathbf{1} + \mathrm{O}(\varepsilon^n) \ .$$

Thus, the set of values $\{u'M(a)\mathbf{1} \mid a \in A\}$ defines the limit distribution of occurrence of symbols in A as the number of steps increases. Note that it does not depend on the initial probabilities π_i's and, intuitively, it plays the analogous of the stationary distribution in ergodic Markov chains.

The previous analysis can be extended to the rational models. Consider a rational formal series $r \in \mathbb{R}_+\langle\langle A\rangle\rangle$ and let (ξ, μ, η) be a linear representation of r. Also define $T = \sum_{x\in A} \mu(x)$. For every symbol $a \in A$ and every word $x = x_1x_2\cdots x_n \in A^n$ generated at random in this model we have

$$\Pr_r(x_n = a) = \frac{\xi' T^{n-1}\mu(a)\eta}{\xi' T^n \eta} \ , \tag{7}$$

$$\Pr_r(x_m = a) = \frac{\xi' T^{m-1}\mu(a)T^{n-m}\eta}{\xi' T^n \eta} \ , \qquad \text{for any} \quad m < n \ . \tag{8}$$

We say that (ξ, μ, η) is primitive if T is primitive. In this case let λ be the Perron-Frobenius eigenvalue of T and let u, v be its left and right eigenvectors normed so that $u'v = 1$. We know that $T^n = \lambda^n(vu' + \mathrm{O}(\varepsilon^n))$ for some $0 < \varepsilon < 1$, and hence from the previous equations we get

$$\lim_{n\to+\infty} \mathrm{Pr}_r(x_n = a) = \frac{u'\mu(a)\eta}{\lambda u'\eta} \ ,$$
$$\lim_{m\to+\infty} \left(\lim_{n\to+\infty} \mathrm{Pr}_r(x_m = a) \right) = \frac{u'\mu(a)v}{\lambda} \ .$$

Now, let us define two probability functions over A, i.e.

$$\alpha : A \longrightarrow [0,1], \quad \alpha(a) = \frac{u'\mu(a)\eta}{\lambda u'\eta} \qquad \forall a \in A \ ,$$
$$\beta : A \longrightarrow [0,1], \quad \beta(a) = \frac{u'\mu(a)v}{\lambda} \qquad \forall a \in A \ .$$

Thus, α represents the limit distribution of symbol occurrence at the end of a word. Note that it does not depend on the initial vector ξ. Analogously, β represents the limit distribution of symbol occurrence in the middle of a long word. It is independent of both the initial and the final vector.

4.2. Fundamental matrix in primitive rational models

It is well-known that in every ergodic Markov chain the number of passages through a given state in the first n steps has a Gaussian limit distribution [10, Sect. 4.3.5]. The same property can be proved for the number of occurrences of a given symbol in words generated by primitive rational models and, more generally, by rational models having a unique dominant component, that is a primitive component with a maximum eigenvalue. In [4,8] more general conditions are given that guarantee a Gaussian limit distribution for the same quantity in rational models. Here, we want to give a further analogy between Markov chains and rational models concerning the fundamental matrix Z given in Section 1.

Consider again a primitive rational model (ξ, μ, η) and let T, λ, u, v be defined as in the previous subsection. Then we know that, for every $n \in \mathbb{N}$, there exists a real matrix $C(n)$ such that $T^n = \lambda^n(vu' + C(n))$ and $C(n) = \mathrm{O}(\varepsilon^n)$, for some $0 < \varepsilon < 1$. Therefore the matrix $C = \sum_{n\geq 0} C(n)$ is well defined and one can prove the following property

Proposition 4.1. *The matrix C defined above satisfies the relation*

$$C = \left(I - \left(\frac{T}{\lambda} - vu' \right) \right)^{-1} \ .$$

Proof. First recall that, for every square matrix X, if $X^n \to 0$ then $(I-X)^{-1}$ exists and $(I-X)^{-1} = \sum_n X^n$. In our case, for every positive integer n, we have

$$\begin{aligned}\left(\frac{T}{\lambda} - vu'\right)^n &= \frac{T^n}{\lambda^n} + \sum_{k=0}^{n-1} \binom{n}{k} \frac{T^k}{\lambda^k} (-1)^{n-k} vu' = \\ &= \frac{T^n}{\lambda^n} + \sum_{k=0}^{n-1} \binom{n}{k} (-1)^{n-k} vu' = \frac{T^n}{\lambda^n} - vu' = C(n) \ .\end{aligned}$$

This implies $(T/\lambda - vu')^n \to 0$ and hence by the property above the matrix $(I-(T/\lambda - vu'))^{-1} = \sum_{n\geq 0}(T/\lambda - vu'))^n$ exists, which proves the result. □

The matrix C appears in the asymptotic expression of the variance of the random variable $y_n(a)$ that represents the number of occurrences of a in a word of length n generated by rational model defined by (ξ, μ, η) [3]. Indeed one can prove that the mean value of $y_n(a)$ is given by $E(y_n(a)) = \beta(a)n + \mathrm{O}(1)$, while its variance is

$$\mathrm{Var}(y_n(a)) = \gamma(a)n + \mathrm{O}(1) \ , \quad \text{where} \quad \gamma(a) = \left(\beta(a) - \beta(a)^2 + 2\frac{u'\mu(a)C\mu(a)v}{\lambda^2}\right).$$

One can also prove that $\gamma(a) \neq 0$ if and only if $T \neq \mu(a) \neq 0$ and, in this case, the random variable $(y_n(a) - \beta(a)n)/\sqrt{\gamma(a)n}$ has a Gaussian limit distribution of mean value 0 and variance 1.

4.3. Time of first symbol occurrence

This quantity can be easily evaluated in case of primitive sequential models. Let $\mathcal{Q} = (\pi, M)$ be defined as in Section 4.1 and, for any $a \in A$, let τ_a denote the position of the first occurrence of a. Defining B as the matrix $B = T - M(a)$, for every integer $i > 0$ we have $\Pr(\tau_a = i) = \pi' B^{i-1} M(a)\mathbf{1}$.

Proposition 4.2. *Let $\mathcal{Q} = (\pi, M)$ be a primitive sequential model over A and set $a \in A$. If $M(a) \neq 0$ then*

$$\begin{aligned} E(\tau_a) &= \pi'(I-B)^{-1}\mathbf{1} \ , \\ Var(\tau_a) &= \pi'(I-B)^{-1}(I+B-\mathbf{1}\pi')(I-B)^{-1}\mathbf{1} \ . \end{aligned}$$

Proof. Since $M(a) \neq 0$, B is equivalent to the restriction of a stochastic matrix to a transient class. Then, by equation (1), $(I-B)^{-1} = \sum_{i\geq 0} B^i$ is well-defined. Analogously, $(I-B)^{-2} = \sum_{i\geq 0} iB^{i-1}$ and thus we can write

$$E(\tau_a) = \sum_{i\geq 0} i\pi' B^{i-1} M(a)\mathbf{1} = \pi'(I-B)^{-2} M(a)\mathbf{1} \ .$$

This proves the first result, since $(M(a)+B)\mathbf{1}=\mathbf{1}$ and hence $M(a)\mathbf{1}=(I-B)\mathbf{1}$. Moreover, the variance $Var(\tau_a)=E((\tau_a^2)-(E(\tau_a))^2)$ satisfies the equality

$$\begin{aligned} Var(\tau_a) &= \sum_{i\geq 0} i^2\pi' B^{i-1}M(a)\mathbf{1} \;-\; \left(\pi'(I-B)^{-1}\mathbf{1}\right)^2 \;= \\ &= \pi'(I-B)^{-1}(B+I)(I-B)^{-1}\mathbf{1} \;-\; \pi'(I-B)^{-1}\mathbf{1}\pi'(I-B)^{-1}\mathbf{1}\;, \end{aligned}$$

which implies the second equation. □

By a similar computation one can prove that the moment generating function of τ_a, defined by $E(e^{t\tau_a})=\sum_{i\geq 0} e^{it}\Pr(\tau_a=i)$, satisfies the identity

$$E(e^{t\tau_a})=\pi' e^t(I-e^tB)^{-1}M(a)\mathbf{1}\;.$$

5. Models for pattern statistics

The major problem in pattern statistics is to estimate the frequency of pattern occurrences in a random text. A formal model for such a statistics is given by a language of patterns $L\subseteq A^*$ and a function $\mathcal{P}: A^*\longrightarrow[0,1]$ defining a probability function on each subset A^n, for $n>0$. The associated statistics $O_n(L,\mathcal{P})$ is defined as the number of occurrences of strings of L in a text x generated at random in A^n with probability $\mathcal{P}(x)$. Here an occurrence is a position where a pattern terminates in the text x. Hence $O_n(L,\mathcal{P})$ is a random variable taking on values in $\{0,1,\ldots,n\}$.

In [3] it is proved that for every regular set $L\subseteq A^*$ and every Markovian model $\mathcal{M}$ of order 1 there exists a rational formal series $s\in\mathbb{R}_+\langle\langle\{a,b\}\rangle\rangle$ such that $O_n(L,\Pr_{\mathcal{M}})$ and $O_n(a,\Pr_s)$ have the same distribution for every integer $n>0$. That is a sort of reduction property from Markovian models of order 1 to rational models: the frequency of regular patterns in the former model can be reduced to the frequency of a single symbol in the latter. Here we extends this result by showing two analogous reductions concerning the rational and the sequencial models, respectively. This also implies that the frequency of regular patterns in any Markovian model is reducible to the frequency of a single symbol in a sequential model, so stressing the role of nondeterminism in such a relationship.

Theorem 5.1. *For every regular language $L\subseteq A^*$ and every rational formal series $r\in\mathbb{R}_+\langle\langle A\rangle\rangle$ there exists a rational formal series $s\in\mathbb{R}_+\langle\langle\{a,b\}\rangle\rangle$ such that, for every integer $n>0$ and every $k=0,1,\ldots,n$*

$$Prob\{O_n(L,Pr_r)=k\}=Prob\{O_n(a,Pr_s)=k\}\;.$$

Proof. Let $\langle Q,A,p,\delta,F\rangle$ be the deterministic automaton recognizing the language A^*L and let $\langle\xi,\mu,\eta\rangle$ be a linear representation of r, with $\xi,\eta\in\mathbb{R}_+^m$, $\mu: A\longrightarrow\mathbb{R}_+^{m\times m}$. Moreover, let l be the cardinality of A. We define the linear representation $\langle\rho,\nu,\tau\rangle$ as follows:

$$\rho,\tau\in\mathbb{R}_+^{Q'},\;\nu:\{a,b\}\longrightarrow\mathbb{R}_+^{Q'\times Q'},$$

where $Q' = Q \times \{1, \ldots, m\} \times A$ and

$$\begin{aligned}
\rho_{(q,j,\sigma)} &= \begin{cases} \xi_j/l & \text{if } q = p \\ 0 & \text{otherwise} \end{cases}, \\
\tau_{(q,j,\sigma)} &= \eta_j \ , \\
\nu(a)_{(q,j,\sigma)(q',j',\sigma')} &= \begin{cases} \mu(\sigma')_{j,j'} & \text{if } \delta(q,\sigma') = q' \quad \text{and} \quad q' \in F \\ 0 & \text{otherwise} \end{cases}, \\
\nu(b)_{(q,j,\sigma)(q',j',\sigma')} &= \begin{cases} \mu(\sigma')_{j,j'} & \text{if } \delta(q,\sigma') = q' \quad \text{and} \quad q' \notin F \\ 0 & \text{otherwise} \end{cases}.
\end{aligned}$$

We show that $\langle \rho, \nu, \tau \rangle$ is the linear representation we are looking for. Let $f : A^+ \longrightarrow \{a,b\}^+$ be the function such that, for every word $\sigma_1\sigma_2\cdots\sigma_n \in A^+$, $f(\sigma_1\sigma_2\cdots\sigma_n) = x_1x_2\cdots x_n$, where $x_i = a$ if $\delta(p, \sigma_1\sigma_2\cdots\sigma_i) \in F$, and $x_i = b$ otherwise. We now prove that, for every $x \in \{a,b\}^+$, $\rho' \ \nu(x)\tau = \sum_{\omega \in f^{-1}(x)} \xi' \mu(\omega) \eta$. Indeed,

$$\rho' \ \nu(x)\tau \ = \sum_{\substack{q_0,\ldots,q_n \in Q \\ 1 \le i_0,\ldots,i_n \le m \\ \sigma_0,\ldots,\sigma_n \in A}} \rho_{(q_0,i_0,\sigma_0)} \left(\prod_{j=1}^{n} \nu(x_j)_{(q_{j-1},i_{j-1},\sigma_{j-1})(q_j,i_j,\sigma_j)} \right) \tau_{(q_n,i_n,\sigma_n)}.$$

By construction, the last sum is equal to

$$\sum_{\substack{q_0 = p \\ 1 \le i_0,\ldots,i_n \le m \\ \sigma_0,\ldots,\sigma_n \in A}} \rho_{(q_0,i_0,\sigma_0)} \left(\prod_{\delta(q_{j-1},\sigma_j)=q_j} \nu(x_j)_{(q_{j-1},i_{j-1},\sigma_{j-1})(q_j,i_j,\sigma_j)} \right) \tau_{(q_n,i_n,\sigma_n)},$$

where, once $\sigma_1\sigma_2\cdots\sigma_n \in A^n$ is fixed, the sequence $q_0, q_1, \ldots, q_n$ is defined by $q_0 = p$ and $q_j = \delta(q_{j-1}, \sigma_j)$, for $j = 1, \ldots, n$. The terms of this sum are different from zero only if $\sigma_1\sigma_2\cdots\sigma_n \in f^{-1}(x)$. Thus, since $\rho_{(p,i_0,\sigma_0)} = \xi_{i_0}/l$, we have

$$\begin{aligned}
\rho' \ \nu(x)\tau \ &= \sum_{\substack{1 \le i_0,\ldots,i_n \le m \\ \sigma_1\cdots\sigma_n \in f^{-1}(x)}} \xi_{i_0} \left(\prod_{\delta(q_{j-1},\sigma_j)=q_j} \mu(\sigma_j)_{i_{j-1},i_j} \right) \eta_{i_n} \ = \\
&= \sum_{\sigma_1\cdots\sigma_n \in f^{-1}(x)} \xi' \ \mu(\sigma_1\sigma_2\cdots\sigma_n) \ \eta.
\end{aligned}$$

Therefore,

$$\sum_{x \in \{a,b\}^n} \rho' \ \nu(x)\tau \ = \ \sum_{\omega \in A^n} \xi' \ \mu(\omega)\eta$$

and hence, for every n and k, we have

$$\frac{\sum_{\omega\in A^n,\ |\omega|_L=k} \xi'\ \mu(\omega)\eta}{\sum_{\omega\in A^n} \xi'\ \mu(\omega)\eta} = \frac{\sum_{x\in\{a,b\}^n,\ |x|_a=k} \rho'\ \nu(x)\tau}{\sum_{x\in\{a,b\}^n} \rho'\ \nu(x)\tau} ,$$

which proves the result. □

By adapting the previous proof to sequential models, we obtain the following statement.

Theorem 5.2. *For every regular language $L \subseteq A^*$ and every sequential model $\mathcal{Q}$ over A there exists a sequential model $\mathcal{N}$ over $\{a, b\}$ such that, for every integer $n > 0$ and every $m = 0, 1, \ldots, n$,*

$$Prob\{O_n(L, Pr_{\mathcal{Q}}) = m\} = Prob\{O_n(a, Pr_{\mathcal{N}}) = m\} .$$

We conclude observing that the construction given in the last proof yields a nondeterministic automaton and hence it cannot be used to show the same reduction between Markovian models. However, it shows that the frequency of a regular set of patterns in a word generated by a Markovian model is always equivalent to the frequency of a single symbol in a binary text generated in a sequential model.

Acknowledgments.
We warmly thank Jean Mairesse for interesting and precious discussions about the subject of this work.

References

[1] I. N. Bernstein, Modules over a ring of differential operators, study of the fundamental solutions of equations with constant coefficients, *Functional Anal. Appl.* vol. 5 (1971), pages 1-16 (Russian); pages 89-101 (English).

[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. *Theoret. Comput. Sci.*, 302(1-3):431–456, 2003.

[4] A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati. Local limit properties for pattern statistics and rational models. *Theory of Computing Systems*, 39 (1): 209–235, 2006.

[5] J. Bourdon and B. Vallée. Generalized pattern matching statistics. *Mathematics and computer science II: algorithms, trees, combinatorics and probabilities*. Proc. of Versailles Colloquium, Birkhauser, 249–265, 2002.

[6] A. Denise. Génération aléatoire et uniforme de mots de langages rationnels. *Theoret. Comput. Sci.*, 159(1):43–63, 1996.

[7] P. Flajolet, P. Zimmerman, and B. Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theoret. Comput. Sci.*, 132(1-2):1–35, 1994.

[8] M. Goldwurm and V. Lonati. Pattern occurrences in multicomponent models. Proc. 22nd STACS, LNCS n. 3404, 680–692, 2005.

[9] G. Hansel and D. Perrin. Rational probability measures. *Theoret. Comput. Sci.*, 65 : 171–188, 1989 (french version in *Mots*, M. Lothaire ed., Hermes, 1990, pp. 335–357).
[10] M. Iosifescu. *Finite Markov Processes and Their Applications*, J. Wiley and Sons, 1980.
[11] P. Jacket and W. Szpankowski. Analytic approach to pattern matching, Ch.7 in M. Lothaire, *Applied Combinatorics on Words*, Cambridge University Press, 2005.
[12] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*, Van Nostrand, 1960.
[13] L. Lipshitz. *D*-finite power series, *Journal of Algebra*, 122 : 353–373, 1989.
[14] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoret. Comput. Sci.*, 287(2):593–617, 2002.
[15] A. Paz. *Introduction to Probabilistic Automata*, Academic Press, 1971.
[16] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22 (4):621–649, 1998.
[17] A. Salomaa and M. Soittola. *Automata-Theoretic Aspects of Formal Power Series*, Springer-Verlag, 1978.
[18] E. Seneta. *Non-negative Matrices and Markov Chains*, Springer–Verlag, 1981.
[19] R. P. Stanley. Differentiably finite power series, *European Journal of Combinatorics*, 1 : 175–188, 1980.

Communicated by (The editor will be set by the publisher).