

# Markovian embeddings of general random strings

Manuel E. Lladser<sup>\*†</sup>

## Abstract

Let  $\mathcal{A}$  be a finite set and  $X$  a sequence of  $\mathcal{A}$ -valued random variables. We do not assume any particular correlation structure between these random variables; in particular,  $X$  may be a non-Markovian sequence. An adapted embedding of  $X$  is a sequence of the form  $R(X_1)$ ,  $R(X_1, X_2)$ ,  $R(X_1, X_2, X_3)$ , etc where  $R$  is a transformation defined over finite length sequences. In this extended abstract we characterize a wide class of adapted embeddings of  $X$  that result in a first-order homogeneous Markov chain. We show that any transformation  $R$  has a unique coarsest refinement  $R'$  in this class such that  $R'(X_1)$ ,  $R'(X_1, X_2)$ ,  $R'(X_1, X_2, X_3)$ , etc is Markovian. (By refinement we mean that  $R'(u) = R'(v)$  implies  $R(u) = R(v)$ , and by coarsest refinement we mean that  $R'$  is a deterministic function of any other refinement of  $R$  in our class of transformations.) We propose a specific embedding that we denote as  $R^X$  which is particularly amenable for analyzing the occurrence of patterns described by regular expressions in  $X$ . A toy example of a non-Markovian sequence of 0's and 1's is analyzed thoroughly: discrete asymptotic distributions are established for the number of occurrences of a certain regular pattern in  $X_1, \dots, X_n$  as  $n \rightarrow \infty$  whereas a Gaussian asymptotic distribution is shown to apply for another regular pattern.

## 1 Introduction

Imagine a gambling scenario where two players name different patterns of heads and tails and flip a coin until one of the patterns is observed. The difficulty in determining which pattern will be the first to be observed or how often will each pattern be observed when the coin is tossed infinitely often depends greatly on the complexity of the patterns involved as well as the probabilistic model of coin-tossing.

In a more general context we may consider a finite non-empty set  $\mathcal{A}$  and an infinite sequence of  $\mathcal{A}$ -valued random variables  $X = (X_n)_{n \geq 1}$ .

We call  $\mathcal{A}$  the *alphabet* and refer to its elements as *characters*. We use the script  $\mathcal{A}^*$  to denote the set of all finite length sequences of elements in  $\mathcal{A}$ . We refer to

elements in  $\mathcal{A}^*$  as *words* or sometimes *strings*. In our context a *pattern* is any non-empty set  $\mathcal{L} \subset \mathcal{A}^*$ .

The *frequency statistics* associated with a pattern  $\mathcal{L}$  are the random variables

$$S_n^{\mathcal{L}} := \sum_{i=1}^n \mathbb{I}[X_1 \cdots X_i \in \mathcal{L}],$$

with  $n \geq 1$ . Above  $X_1 \cdots X_i$  is a shortcut notation for the random word  $(X_1, \dots, X_i)$ . Furthermore,  $\mathbb{I}[\cdot]$  denotes the Iverson's bracket i.e. a quantity defined to be 1 whenever the statement within the brackets is true, and 0 otherwise. Equivalently,  $\mathbb{I}[X_1 \cdots X_i \in \mathcal{L}]$  is the indicator function of the event  $[X_1 \cdots X_i \in \mathcal{L}]$ .

To fix some ideas about the above definition consider a nonempty set  $\mathcal{W} \subset \mathcal{A}^*$ . If  $\mathcal{W}$  is *suffix reduced* i.e. no word in  $\mathcal{W}$  is a proper suffix of another word in  $\mathcal{W}$  and  $\mathcal{L} = \mathcal{A}^* \mathcal{W}$  i.e.  $x \in \mathcal{L}$  if and only if  $x$  has a suffix in  $\mathcal{W}$  then  $S_n^{\mathcal{L}}$  corresponds to the number of substrings of  $X_1 \cdots X_n$  that belong to  $\mathcal{W}$ .

Much of the efforts undertaken in the literature for studying patterns in random strings have been about determining the exact and/or asymptotic distribution of the frequency statistics of one or more patterns. Various techniques have been utilized for this effect for different types of patterns as well as probabilistic models for  $X$ . These have included *martingale methods* [15, 21], *combinatorial methods* [12, 2], *renewal arguments* and *formal language recursions* [23, 10], *large deviations* [22], and *symbolic dynamics* [5, 6], among many others.

Perhaps the most commonly used technique for analyzing patterns in random strings is the (*finite*) *Markov chain embedding technique* (MCET). It originated in [11, 3] as a technique for analyzing patterns described by a finite set of words in the context of memoryless sequences. The case of first-order homogeneous Markov sequences was provided in [4]. An important extension of the technique to consider general *Markovian models* i.e. when  $X$  is a  $k$ -th order homogeneous Markov chain and *regular patterns* i.e. sets of words described by regular expressions is due to [19] and the follow up work in [18]. Minimality considerations about the automata needed for analyzing the frequency statistics of regular patterns under Markovian models were addressed in [16].

Broadly speaking, the Markov chain embedding

<sup>\*</sup>The University of Colorado, Department of Applied Mathematics, PO Box 526 UCB, Boulder, CO 80309-0526, The United States

<sup>†</sup>e-mail: manuel.lladser@colorado.edu

technique consists in transforming  $X$  into a first-order homogeneous Markov chain that is informative of a pattern of interest. In [17] the embedding of  $X$  into a *deterministic finite automaton* (DFA)  $G = (V, \mathcal{A}, f, q, T)$  that recognizes a regular pattern  $\mathcal{L}$  is defined as the infinite sequence of  $V$ -valued random variables <sup>1</sup>

$$(1.1) \quad X_n^G := f(q, X_1 \cdots X_n).$$

We refer to  $X^G = (X_n^G)_{n \geq 1}$  as the *embedding of  $X$  into  $G$* . If  $X^G$  is a first-order homogeneous Markov chain then  $S_n^{\mathcal{L}}$  has the same distribution as the number of visits that  $X^G$  makes to  $T$  in the first  $n$  steps. In this case, *transfer matrix methods* of the type implemented in [19, 18] can be used to determine the generating function associated with the frequency statistics of the regular pattern  $\mathcal{L}$ .

In order for the MCET to work the embedding  $X^G$  must be Markovian. This requires a level of compatibility between the distribution of  $X$  and the transition function of  $G$ . To fix ideas consider a first-order homogeneous Markov chain  $X$  with state space  $\{a, b\}$  and the regular languages  $\mathcal{L}_1 = \{a, b\}^* \{ba\}$  and  $\mathcal{L}_2 = \{a, b\}^* \{abba\}$ . If  $AC$  is the Aho-Corasick automaton [1] associated with the keywords ‘ $ba$ ’ and ‘ $abba$ ’ (see Figure 1) then  $X^{AC}$  is known to be a first-order homogeneous Markov chain because  $AC$  conveys a memory length of order one [16] i.e. each state accessible from the initial state is informative of the last character processed by the automaton. In particular, if

$$\begin{aligned} P[X_1 = a] &= \mu; \\ P[X_1 = b] &= (1 - \mu); \\ P[X_{n+1} = a \mid X_n = a] &= p; \\ P[X_{n+1} = b \mid X_n = a] &= (1 - p); \\ P[X_{n+1} = a \mid X_n = b] &= q; \\ P[X_{n+1} = b \mid X_n = b] &= (1 - q); \end{aligned}$$

then

$$(1.2) \quad P[S_n^{\mathcal{L}_1} = k_1, S_n^{\mathcal{L}_2} = k_2] = [x^{k_1} y^{k_2}] \mu^T \cdot P_{x,y}^{n-1} \cdot \mathbf{1}$$

where

$$\mu^T := [ \mu \quad (1 - \mu) \quad 0 \quad 0 \quad 0 \quad 0 ];$$

$$P_{x,y} := \begin{bmatrix} p & 0 & (1-p) & 0 & 0 & 0 \\ 0 & (1-q) & 0 & x \cdot q & 0 & 0 \\ 0 & 0 & 0 & x \cdot q & (1-q) & 0 \\ p & 0 & (1-p) & 0 & 0 & 0 \\ 0 & (1-q) & 0 & 0 & 0 & xy \cdot q \\ p & 0 & (1-p) & 0 & 0 & 0 \end{bmatrix};$$

$$\mathbf{1}^T := [ 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 ].$$

<sup>1</sup> $V$  denotes the (finite) state space of  $G$ ,  $f : V \times \mathcal{A}^* \rightarrow V$  its transition function,  $q \in V$  its initial state, and  $T \subset V$  its set of terminal states.

The row-vector  $\mu^T$  corresponds to the initial distribution of  $X^{AC}$  according to the labels assigned in Figure 1.  $P_{x,y}$  is the probability transition matrix of  $X^{AC}$  but after multiplying by the (marker) variable  $x$  each column associated with a state in  $AC$  that contributes to an occurrence of the keyword ‘ $ba$ ’. Similarly the variable  $y$  marks matches with the keyword ‘ $abba$ ’. In particular, using (1.2) we obtain that the moment generating function associated with the random vector  $(S_n^{\mathcal{L}_1}, S_n^{\mathcal{L}_2})$  is given by the formula

$$\begin{aligned} \sum_{k_1, k_2 \geq 0, n \geq 1} P[S_n^{\mathcal{L}_1} = k_1, S_n^{\mathcal{L}_2} = k_2] x^{k_1} y^{k_2} z^n \\ = z \cdot \mu^T \cdot (\mathbb{I} - z \cdot P_{x,y})^{-1} \cdot \mathbf{1}, \end{aligned}$$

where  $\mathbb{I}$  is the  $6 \times 6$  identity matrix.

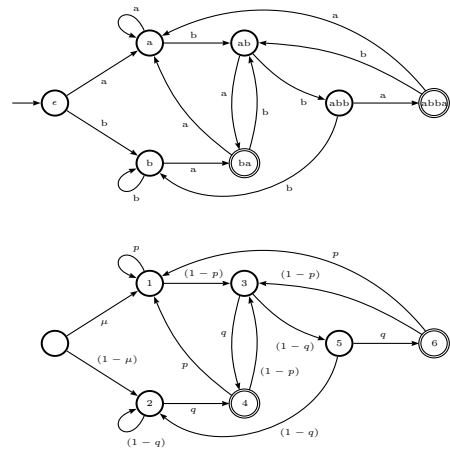


Figure 1: Top, Aho-Corasick automaton  $AC$  that recognizes the regular language  $\{a, b\}^* \{ba, abba\}$ .  $\epsilon$  is the initial state.  $ba$  and  $abba$  are terminal states. Transitions into state  $ba$  correspond to matches with keyword ‘ $ba$ ’ that do not contribute to a match with the keyword ‘ $abba$ ’. Transitions into state  $abba$  correspond to matches with ‘ $ba$ ’ and ‘ $abba$ ’. Bottom, embedding of a general first-order homogeneous Markov chain  $X$  with state space  $\{a, b\}$  into  $AC$ .

The automaton  $AC$  also conveys a memory length of order 2; in particular, the embedding of any second-order homogeneous Markov chain  $X$  into  $G$  is also Markovian [16]. On the contrary, for a general third-order homogeneous Markov chain  $X$  over  $\{a, b\}$ , the Aho-Corasick automaton  $AC$  in Figure 1 may not be suitable for a Markovian embedding: starting from the initial state it is possible to reach state  $a$  using the text ‘ $aaaaa$ ’ but also ‘ $babaa$ ’, in particular, state  $a$  is not informative of the last three characters fed into the automaton. This nuisance can be resolved by

duplicating states until each state becomes informative of the last three characters fed into the automaton [19, 18], or by considering the product between  $AC$  and a de Bruijn like graph [16].

In this extended abstract we develop some theoretical bases for a general MCET to consider arbitrary probabilistic models and patterns. By arbitrary probabilistic models we mean possibly non-Markovian models. By arbitrary patterns we mean possibly non-regular ones. The starting point of our approach is the observation that the embedding in (1.1) may be extrapolated as a sequence of the form  $R(X_1)$ ,  $R(X_1X_2)$ ,  $R(X_1X_2X_3)$ , etc where  $R$  is a transformation defined over  $\mathcal{A}^*$  and taking values in a certain space  $\mathcal{C}$ . In compact form we write  $X^R$  to refer to the transformed sequence i.e.

$$X_n^R := R(X_1 \cdots X_n).$$

Two natural questions in this context are:

- (a) *what conditions are necessary and sufficient in order for  $X^R$  to be a first-order homogeneous Markov chain for a possibly non-Markovian sequence  $X$ ?, and*
- (b) *given a possibly non-regular pattern  $\mathcal{L}$ , is there an  $R$  such that  $X^R$  is a Markov chain informative of the frequency statistics of  $\mathcal{L}$  but also with the ‘least’ number of states?*

Observe that the actual range of  $R$  is not really relevant for answering the above questions: any set in one-to-one correspondence with it may be regarded as the range without affecting the Markovianity of  $X^R$  nor the number of states that this chain could potentially visit. In particular, the level curves of  $R$  are the most natural choice for the range of the transformation. Equivalently, since the level curves of  $R$  form a partition of  $\mathcal{A}^*$ , we may think of  $R$  as an equivalence relation defined over  $\mathcal{A}^*$ . Henceforth  $c \in R$  will mean that  $c$  is an equivalence class of  $R$  and, for  $x \in \mathcal{A}^*$ ,  $R(x)$  will denote the unique equivalence class that contains  $x$ . (This notation allows to think of  $R$  simultaneously as an equivalence relation and as a transformation.) In particular,  $R(x) = R(y)$  is equivalent to having  $xRy$  i.e. that  $x$  and  $y$  are in the same equivalence class of  $R$ . Furthermore,  $x \in c$  is equivalent to having  $R(x) = c$ .

To fix some ideas we consider some examples. If  $R_1$  is the relation defined as  $xR_1y$  if and only if  $x = y$  then  $R_1(x) = \{x\}$ . We refer to this relation as the *identity relation*. Observe that the equivalence classes of  $R_1$  correspond to the level curves of any one-to-one transformation defined over  $\mathcal{A}^*$ . On the other hand, if  $R_2$  is such that  $xR_2y$  for all  $x, y \in \mathcal{A}^*$  then  $R_2(x) = \mathcal{A}^*$ . We call  $R_2$  the *coarsest relation*. The equivalence classes

of  $R_2$  correspond to the level curves of any constant transformation defined over  $\mathcal{A}^*$ .

**1.1 Outline of the paper.** In §2 we introduce a class of equivalence relations which are guarantee to produce Markovian embeddings. A characterization of the equivalence relations in this class is provided in Theorem 2.1. Our result resembles the characterization of strong lumpability (also called strong state space aggregation) for Markov chains [14], but in the more general context of possibly non-Markovian sequences. Theorem 2.2 asserts that it is always possible to refine the equivalence classes of a given relation so as to obtain an equivalence relation in the class of transformations characterized by Theorem 2.1, but with the least possible number of equivalence classes. Our result touches bases with the algorithm proposed in [20] and implemented in [8] for finding the optimal strongly lumpable refinement of a partition of the state space of a Markov chain. At the end of §2 we introduce an equivalence relation whose equivalence classes are defined in terms of the distribution of  $X$ . Some key properties of this relation are presented in Theorem 2.3. (The proofs of our results in §2 are omitted from this extended abstract and will be provided in the final version of the paper.) In §3 we review briefly some basic limit theorems for Markov chains and see how the results of §2 fit for analyzing the frequency statistics of general patterns in general random strings. The end of §3 is devoted to a case study of a non-Markovian sequence and the frequency statistics of two regular patterns. The study reveals new phenomena which has not been previously observed, even in the context of probabilistic dynamical models [6].

## 2 Main definitions and results

For an integer  $n > 0$ ,  $\mathcal{A}^n$  denotes the set of all sequences of  $n$  elements in  $\mathcal{A}$ . We define  $\mathcal{A}^0 = \{\epsilon\}$ , where  $\epsilon \notin \mathcal{A}$  is by definition the *empty word*. We also define  $|x| = n$  whenever  $x \in \mathcal{A}^n$ . We call  $|x|$  the *length of  $x$* . Observe that  $x = \epsilon$  if and only if  $|x| = 0$ . Define  $\mathcal{A}^* := \cup_{n \geq 0} \mathcal{A}^n$ .

In what follows,  $X = (X_n)_{n \geq 1}$  is a given infinite sequence of  $\mathcal{A}$ -valued random variables defined on a common probability space  $(\Omega, \mathcal{F}, P)$ . For  $x \in \mathcal{A}^*$ ,  $|x| > 0$ , we use the notation  $[X = x...]$  as a shortcut of the event  $[X_1 \cdots X_{|x|} = x]$ . We define  $[X = \epsilon...] = \Omega$ . The *support of  $X$*  is the set

$$\text{supp}(X) := \{x \in \mathcal{A}^* : P[X = x...] > 0\}.$$

Observe that the above set is nonempty because  $\epsilon \in \text{supp}(X)$ . Furthermore,  $X_1 \cdots X_n \in \text{supp}(X)$  almost surely for all  $n \geq 1$ .

Henceforth  $R$  will denote an equivalence relation

defined over  $\text{supp}(X)$ . For technical purposes we extend any such relation to an equivalence relation over  $\mathcal{A}^*$  by defining

$$R(x) := \mathcal{A}^* \setminus \text{supp}(X), \quad x \notin \text{supp}(X).$$

**DEFINITION 2.1.** We will say that  $X$  is embedable w.r.t.  $R$  provided that for all  $x, y \in \text{supp}(X)$  and  $c \in R$ , if  $R(x) = R(y)$  then

$$\begin{aligned} & \sum_{\alpha \in \mathcal{A}: R(x\alpha) = c} P[X = x\alpha \dots \mid X = x \dots] \\ &= \sum_{\alpha \in \mathcal{A}: R(y\alpha) = c} P[X = y\alpha \dots \mid X = y \dots]. \end{aligned}$$

Observe that the left- and right-hand side above are identically zero when  $c \cap \text{supp}(X) = \emptyset$ . As a result, to check embedability we may always assume that  $c \subset \text{supp}(X)$ .

It is direct to check that  $X$  is embedable w.r.t. the restriction to  $\text{supp}(X)$  of the identity relation and the coarsest relation. In particular there exist equivalence relations w.r.t. which  $X$  is embedable. Observe however that none of these equivalence relations takes into account the actual distribution of  $X$  in their definition. An important instance of such a relation is provided in section §2.3.

**2.1 Characterization of embedability.** Before stating our first result we need introduce some notation. In what follows, the script  $\mu$  is reserved to denote a probability measure defined over  $\text{supp}(X)$ . Since this last set is countable, we will write  $\mu(x)$  instead of  $\mu(\{x\})$ . Furthermore, we use the notation  $X^\mu = (X_n^\mu)_{n \geq 0}$  to refer to the unique stochastic process in  $\text{supp}(X)$  such that

$$(2.3) \quad X_0^\mu \stackrel{d}{=} \mu,$$

and for all  $n \geq 1$  and  $x \in \text{supp}(X)$  such that  $\mu(x) > 0$

$$(2.4) \quad \begin{aligned} & X_{|x|+1} \cdots X_{|x|+n} \mid [X_1 \cdots X_{|x|} = x] \\ & \stackrel{d}{=} X_1^\mu \cdots X_n^\mu \mid [X_0^\mu = x]. \end{aligned}$$

Due to condition (2.3),  $X_0^\mu$  is a  $\text{supp}(X)$ -valued random variable. Condition (2.4) is equivalent to requesting that  $\mu$ -almost surely for all  $x$ , the conditional distribution of the stochastic process  $(X_n^\mu)_{n \geq 1}$  given that  $X_0^\mu = x$  is the same as the conditional distribution of  $(X_{n+|x|})_{n \geq 1}$  given the event  $[X = x \dots]$ . In particular, for all  $n \geq 1$ ,  $X_n^\mu$  is an  $\mathcal{A}$ -valued random variable.

**THEOREM 2.1.**  $X$  is embedable w.r.t.  $R$  if and only if, for all  $\mu$ , the stochastic process  $Y = (Y_n)_{n \geq 0}$ , with  $Y_n := R(X_0^\mu \cdots X_n^\mu)$ , is a first-order homogeneous Markov chain with probability transitions that do not depend upon  $\mu$ .

Observe that if  $\mu = \delta_\epsilon$  (the probability mass function at  $\epsilon$ ) then  $(X_n^\mu)_{n \geq 1}$  has the same distribution as  $X$ . In particular, the following result is an immediate consequence of the above theorem.

**COROLLARY 2.1.** If  $X$  is embedable w.r.t.  $R$  then the stochastic process  $X^R$  is a first-order homogeneous Markov chain.

**2.2 Coarsest embedable refinement.** For a given equivalence relation  $R$  defined over  $\text{supp}(X)$  the process  $X$  may or may not be embedable w.r.t.  $R$ . However the following result asserts that it is always possible to refine the equivalence classes of  $R$  so as to obtain an equivalence relation with the ‘least’ number of equivalence classes w.r.t. which  $X$  is embedable.

**THEOREM 2.2.** For each equivalence relation  $R$  defined over  $\text{supp}(X)$ , there exists a unique coarsest refinement  $R'$  of  $R$  w.r.t. which  $X$  is embedable.

**2.3 Markov relation induced by  $X$ .** An equivalence relation  $R$  is said to be *right-invariant* if for all  $x, y \in \mathcal{A}^*$  the condition  $R(x) = R(y)$  implies that  $R(x\alpha) = R(y\alpha)$ , for all  $\alpha \in \mathcal{A}$ . In particular, for a right-invariant relation  $R$  it applies that

$$R(x) = R(y) \iff (\forall z \in \mathcal{A}^*) : R(xz) = R(yz).$$

The identity relation as well as the coarsest-relation are trivial examples of right-invariant relations w.r.t. which  $X$  is embedable. Another example of such a relation but that takes into account the actual distribution of  $X$  is provided by the following definition.

**DEFINITION 2.2.** The *Markov relation induced by  $X$*  is the equivalence relation  $R^X$  over  $\text{supp}(X)$  defined as follows:  $xR^X y$  if and only if

$$\begin{aligned} & (\forall z \in \mathcal{A}^*) : P[X = xz \dots \mid X = x \dots] \\ &= P[X = yz \dots \mid X = y \dots]. \end{aligned}$$

We extend  $R^X$  to an equivalence relation to the whole of  $\mathcal{A}^*$  by setting

$$R^X(x) := \mathcal{A}^* \setminus \text{supp}(X), \quad x \notin \text{supp}(X).$$

**THEOREM 2.3.**  $R^X$  is a right-invariant relation. Furthermore,  $X$  is embedable w.r.t. any right-invariant equivalence relation that is a refinement of  $R^X$ ; in particular,  $X$  is embedable w.r.t.  $R^X$ .

### 3 Frequency statistics of patterns

There are two ways in which the results of the previous section fit nicely for analyzing the frequency statistics of a pattern  $\mathcal{L}$  in the infinite sequence  $X$ . We may first induce in  $\text{supp}(X)$  the equivalence relation

$$R := \{\text{supp}(X) \cap \mathcal{L}, \text{supp}(X) \setminus \mathcal{L}\}.$$

The equivalence classes of the above relation correspond to the indicator function of  $\mathcal{L}$  when restricted to the support of  $X$ . In particular, there is no warranty that the embedding  $X^R$  is Markovian at all. However, according to Theorem 2.2 and Corollary 2.1,  $Y = (Y_n)_{n \geq 1}$  with  $Y_n := R'(X_1 \cdots X_n)$  is a first-order homogeneous Markov chain with the least number of equivalence classes among all equivalence relations w.r.t. which  $X$  is embedable (see Figure 2). In particular, if we define

$$T := \{c \in R' : c \subset \text{supp}(X) \cap \mathcal{L}\}$$

then

$$(3.5) \quad S_n^{\mathcal{L}} \stackrel{d}{=} \sum_{i=1}^n \mathbb{I}[Y_i \in T],$$

i.e. the distribution of  $S_n^{\mathcal{L}}$  corresponds to the number of visits that  $Y$  makes to the states in  $T$  in the first  $n$  steps. If  $Y$  is irreducible and positive recurrent and  $\pi$  denotes the unique stationary distribution of this chain then the strong law for additive functionals of Markov chains [9] implies that

$$(3.6) \quad \lim_{n \rightarrow \infty} \frac{S_n^{\mathcal{L}}}{n} = \sum_{c \in T} \pi(c),$$

almost surely, regardless of the initial distribution of  $Y$ . If in addition to the above conditions there exists  $c_0$  such that

$$\sigma_0^2 := E \left( \left\{ \sum_{i=0}^{\tau_0-1} \mathbb{I}[Y_i \in T] - \tau_0 \cdot \sum_{c \in T} \pi(c) \right\}^2 \middle| Y_0 = c_0 \right)$$

is strictly positive and finite, where

$$\tau_0 := \min\{n \geq 1 : Y_n = c_0\},$$

then the central limit theorem for additive functionals of Markov chains [9] implies that there exists  $\sigma > 0$  such that

$$(3.7) \quad \lim_{n \rightarrow \infty} \frac{S_n^{\mathcal{L}} - n \cdot \sum_{c \in T} \pi(c)}{\sqrt{n}} \stackrel{d}{=} \sigma \cdot W,$$

where  $W$  is a standard Normal random variable<sup>2</sup>. We remark that the condition  $E(\tau_0^2 | Y_0 = c_0) < +\infty$  is sufficient to have  $\sigma_0^2 < +\infty$ .

<sup>2</sup> $\tau_0$  is usually called the *first-return time* of  $Y$  to  $c_0$ .

In most situations of interest  $R'$  will consist of a countable number of equivalence classes and the embedding  $X^{R'}$  will not be analytically tractable. A slight improvement in this direction may be attained by letting the alphabet play a more direct role in the embedding. For this we induce in  $\mathcal{A}^*$  the *Mihill-Nerode equivalence relation* defined as

$$x R_{\mathcal{L}} y \iff (\forall z \in \mathcal{L}) : [xz \in \mathcal{L} \iff yz \in \mathcal{L}].$$

The relation  $R_{\mathcal{L}}$  is clearly right-invariant. (If  $\mathcal{L}$  is a regular pattern then  $R_{\mathcal{L}}$  has a finite number of equivalence classes due to the Mihill-Nerode Theorem [13].) In particular, if  $R$  is any right-invariant refinement of  $R^X$  then so is the product between  $R$  and  $R_{\mathcal{L}}$  i.e. the equivalence relation  $R_{\mathcal{L}}^X$  defined over  $\mathcal{A}^*$  as follows

$$x R_{\mathcal{L}}^X y \iff [x R y \text{ and } x R_{\mathcal{L}} y].$$

Observe that if  $c_1 \in R$  and  $c_2 \in R_{\mathcal{L}}$  then  $(c_1 \cap c_2)$  is an equivalence class of  $R_{\mathcal{L}}^X$  provided that the intersection is not empty. Conversely, any equivalence class of  $R_{\mathcal{L}}^X$  is of the form  $(c_1 \cap c_2)$  for a unique  $c_1 \in R$  and  $c_2 \in R_{\mathcal{L}}$ . Due to this the equivalence classes of  $R_{\mathcal{L}}^X$  are in one-to-one correspondence with ordered-pairs of the form  $(c_1, c_2) \in R \times R_{\mathcal{L}}$  such that  $(c_1 \cap c_2) \neq \emptyset$ . This property motivates the terminology of product relation.

Since  $R_{\mathcal{L}}^X$  is a right-invariant refinement of  $R^X$ , it follows from Theorem 2.3 and Corollary 2.1 that the embedding  $Y_n = R_{\mathcal{L}}^X(X_1 \cdots X_n)$  can be regarded as a first-order homogeneous Markov chain with state space  $R \times R_{\mathcal{L}}$  (see Figure 3). Identity (3.5) will also apply for this embedding but with

$$T := \{(c_1, c_2) \in R \times R_{\mathcal{L}} : c_2 \subset \text{supp}(X) \cap \mathcal{L}\},$$

and so will (3.6) and (3.7) provided that the necessary technical conditions are verified. To make explicit the probability transitions of this chain consider the transition function  $f : R \times R_{\mathcal{L}} \times \mathcal{A} \rightarrow R \times R_{\mathcal{L}}$  defined as

$$f(c_1, c_2, \alpha) := (R(x\alpha), R_{\mathcal{L}}(y\alpha)),$$

provided that  $x \in c_1$  and  $y \in c_2$ . (The above definition does not depend upon the selection of  $x$  or  $y$  because  $R$  and  $R_{\mathcal{L}}$  are right-invariant.) The probability transitions of  $Y$  are then easily found to be

$$\begin{aligned} P[Y_{n+1} = (c'_1, c'_2) | Y_n = (c_1, c_2)] \\ = \sum_{\alpha \in \mathcal{A} : f(c_1, c_2, \alpha) = (c'_1, c'_2)} P[X = x\alpha \dots | X = x \dots], \end{aligned}$$

provided that  $x \in c_1 \cap \text{supp}(X)$ . (The probabilities in the summation above do not depend on the selection of  $x$  because  $R$  is a refinement of  $R^X$ .)

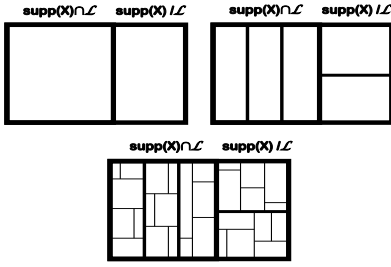


Figure 2: Top-left, schematic representation of partition  $R$  induced in  $\text{supp}(X)$  by a possibly non-regular pattern  $\mathcal{L}$ . Top-right, schematic representation of what could be the coarsest refinement  $R'$  of  $R$  w.r.t. which  $X$  is embeddable. ( $R'$  may have a countable number of equivalence classes.) Bottom, schematic representation of a general refinement of  $R$  w.r.t. which  $X$  is embeddable.

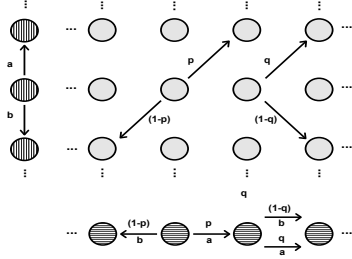


Figure 3: Most-left vertical axis, possible labeled transitions associated with one of the equivalence classes of  $R_{\mathcal{L}}$ , with  $\mathcal{A} = \{a, b\}$ . Each circle represents an equivalence class of the right-invariant relation  $R_{\mathcal{L}}$ . Bottom axis, possible probability and labeled transitions associated with two different states of  $X^R$  for a right-invariant refinement  $R$  of  $R^X$ . Here  $0 \leq p, q \leq 1$  and each circle represents an equivalence class of  $R$ . Middle-grid, state space of the embedding of  $X$  through the product relation  $R_{\mathcal{L}}^X$  between  $R$  and  $R_{\mathcal{L}}$ . The probability transitions associated with two of the states is displayed in accordance with the information available for  $R$  and  $R_{\mathcal{L}}$ .

**3.1 A toy non-Markovian model.** Let  $0 < p < 1/2$  be a given parameter. Consider a sequence  $X = (X_n)_{n \geq 1}$  of  $\{0, 1\}$ -valued random variables such that

$$(3.8) \quad X_{n+1} \stackrel{d}{=} \begin{cases} \text{Bernoulli}(p) & , \frac{1}{n} \sum_{i=1}^n X_i > \frac{1}{2}; \\ \text{Bernoulli}(1/2) & , \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{2}; \\ \text{Bernoulli}(1-p) & , \frac{1}{n} \sum_{i=1}^n X_i < \frac{1}{2}. \end{cases}$$

The binary process  $X$  self-regulates the cumulative averages  $\sum_{i=1}^n X_i/n$  so as to keep them tied to a 50% value. In particular,  $X$  evolves in a non-Markovian way. Consider the regular patterns

$$\begin{aligned} \mathcal{L}_1 &= \{0, 1\}^* \{1\}, \\ \mathcal{L}_2 &= \{0\}^* \{1\} \{0\}^* (\{1\} \{0\}^* \{1\} \{0\}^*)^*. \end{aligned}$$

Both of these patterns are *primitive* i.e. each is recognized by a DFA whose associated digraph is irreducible and aperiodic. In particular, for binary Markovian sequences [19, 18] and more generally for sequences produced by nice probabilistic dynamical sources [6] the frequency statistics of  $\mathcal{L}_1$  and  $\mathcal{L}_2$  ought to be asymptotically Normal. However, as the following result shows, non-Gaussian asymptotic distributions may be also encountered for primitive patterns under more general non-Markovian models.

**PROPOSITION 3.1.** *If  $0 < p < 1/2$  then*

$$(3.9) \quad \pi(n) := \begin{cases} \frac{1-2p}{2(1-p)} & , n = 0; \\ \frac{1-2p}{4p(1-p)} \left(\frac{p}{1-p}\right)^{|n|} & , n \neq 0; \end{cases}$$

satisfies that

$$\sum_{n=0(\bmod 2)} \pi(n) = \sum_{n=1(\bmod 2)} \pi(n) = \frac{1}{2}.$$

Furthermore, the following applies.

(a) *If  $U$  and  $V$  are  $\mathbb{Z}$ -valued random variables such that  $P[U = n] = 2\pi(n)$ , for  $n = 0(\bmod 2)$ , and  $P[V = n] = 2\pi(n)$ , for  $n = 1(\bmod 2)$ , then*

$$\begin{aligned} \lim_{n \rightarrow \infty} 2n \cdot \left\{ \frac{S_n^{\mathcal{L}_1}}{n} - \frac{1}{2} \right\} &\stackrel{d}{=} U; \\ \lim_{n \rightarrow \infty} 2n \cdot \left\{ \frac{S_n^{\mathcal{L}_2}}{n} - \frac{1}{2} \right\} &\stackrel{d}{=} V. \end{aligned}$$

(b) *If  $W$  is a standard Normal random variable then there exists  $\sigma > 0$  such that*

$$\lim_{n \rightarrow \infty} \sqrt{n} \cdot \left\{ \frac{S_n^{\mathcal{L}_2}}{n} - \frac{1}{2} \right\} \stackrel{d}{=} \sigma \cdot W.$$

Observe that  $\text{supp}(X) = \{0, 1\}^*$ . To show the proposition consider the transformation  $R : \{0, 1\}^* \rightarrow \mathbb{Z}$  defined as

$$(3.10) \quad \begin{aligned} R(x) &:= 2 \left\{ \sum_{i=1}^{|x|} x_i - \frac{|x|}{2} \right\}; \\ &= \sum_{i=1}^{|x|} x_i - \sum_{i=1}^{|x|} (1 - x_i). \end{aligned}$$

Above it is understood that  $R(\epsilon) = 0$ . In particular,  $R(xy) = R(x) + R(y)$ , for all  $x, y \in \{0, 1\}^*$ , and hence the level curves of  $R$  define a right-invariant equivalence relation on  $\{0, 1\}^*$ . On the other hand, for  $x, z \in \{0, 1\}^*$  with  $|z| > 0$  it applies from (3.8) that

$$P[X = xz\dots | X = x\dots] = p^{M(x,z)} \cdot (1-p)^{N(x,z)} \cdot \left(\frac{1}{2}\right)^{|z|-M(x,z)-N(x,z)},$$

where

$$\begin{aligned} M(x, z) &:= \sum_{i=1}^{|z|} \mathbb{I}[z_i = 0] \cdot \mathbb{I}[R(z_1 \dots z_{i-1}) < -R(x)] \\ &\quad + \sum_{i=1}^{|z|} \mathbb{I}[z_i = 1] \cdot \mathbb{I}[R(z_1 \dots z_{i-1}) > -R(x)]; \\ N(x, z) &:= \sum_{i=1}^{|z|} \mathbb{I}[z_i = 0] \cdot \mathbb{I}[R(z_1 \dots z_{i-1}) > -R(x)] \\ &\quad + \sum_{i=1}^{|z|} \mathbb{I}[z_i = 1] \cdot \mathbb{I}[R(z_1 \dots z_{i-1}) < -R(x)]; \end{aligned}$$

where it is understood that  $z_1 \dots z_0 = \epsilon$ . From the above identity it is immediate that if  $R(x) = R(y)$  then  $R^X(x) = R^X(y)$ . In particular, since  $R$  is right-invariant, Theorem 2.3 and Corollary 2.1 imply that  $X^R$  is a first-order homogeneous Markov chain with state space  $\mathbb{Z}$ . To obtain the probability transitions of  $X^R$  notice that

$$X_{n+1}^R = X_n^R + R(X_{n+1}).$$

Since  $X_n^R$  and  $(\sum_{i=1}^n X_i - n/2)$  have the same sign, (3.8) implies that

$$X_{n+1}^R - X_n^R = \begin{cases} (+1) & \text{w.p. } p \text{ if } X_n^R > 0; \\ (-1) & \text{w.p. } (1-p) \text{ if } X_n^R > 0; \\ (+1) & \text{w.p. } 1/2 \text{ if } X_n^R = 0; \\ (-1) & \text{w.p. } 1/2 \text{ if } X_n^R = 0; \\ (+1) & \text{w.p. } (1-p) \text{ if } X_n^R < 0; \\ (-1) & \text{w.p. } p \text{ if } X_n^R < 0. \end{cases}$$

The process  $X^R$  is recurrent and has period 2. It is positive recurrent because  $p \cdot 1 + (1-p) \cdot (-1) < 0$  and  $(1-p) \cdot 1 + p \cdot (-1) > 0$ . In particular,  $X^R$  has a unique stationary distribution  $\pi$  supported over  $\mathbb{Z}$ . Indeed, since  $X^R$  is a birth-death chain we may use the detailed balance condition [9] to obtain the explicit formula in (3.9). Since

$$S_n^{\mathcal{L}_1} = \sum_{i=1}^n X_i,$$

part (a) in the proposition is direct from (3.9) and (3.10), and the well-known result on convergence in distribution of a Markov chain to its stationary distribution for the periodic case [9].

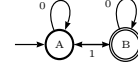


Figure 4: Deterministic finite automaton with initial state  $A$  and terminal state  $B$  that recognizes the binary regular language  $\{0\}^*\{1\}\{0\}^*(\{1\}\{0\}^*\{1\}\{0\}^*)^*$ .

To show part (b) consider the DFA displayed in Figure 4. This automaton is the minimal automaton that recognizes  $\mathcal{L}_2$ . In particular, the equivalence classes of  $R_{\mathcal{L}_2}$  are in one-to-one correspondence with the states of this automaton [13]. Furthermore, since  $R_{\mathcal{L}_2}^X$  is a right-invariant refinement of  $R^X$ ,  $Y_n = R_{\mathcal{L}_2}^X(X_1 \dots X_n)$  is a first-order homogeneous Markov chain. Without loss of generality we may assume that  $Y = (Y_n)_{n \geq 1}$  has  $\mathbb{Z} \times \{A, B\}$  as its state space.  $Y$  is irreducible and has period 4. It is also positive recurrent because if we define

$$(3.11) \quad \pi(n, A) := \pi(n, B) := \frac{\pi(n)}{2}$$

then

$$\begin{aligned} &\sum_{(m,i) \in \mathbb{Z} \times \{A,B\}} \pi(m, i) \cdot P[Y_2 = (n, A) | Y_1 = (m, i)] \\ &= \sum_{(m,i) = (n-1, B), (n+1, A)} \pi(m, i) \cdot P[Y_2 = (n, A) | Y_1 = (m, i)], \\ &= \frac{1}{2} \sum_{m=n-1, n+1} \pi(m) \cdot P[X_2^R = n | X_1^R = m], \\ &= \frac{1}{2} \sum_{m \in \mathbb{Z}} \pi(m) \cdot P[X_2^R = n | X_1^R = m], \\ &= \frac{\pi(n)}{2}, \\ &= \pi(n, A). \end{aligned}$$

Similarly, we obtain that

$$\begin{aligned} &\sum_{(m,i) \in \mathbb{Z} \times \{A,B\}} \pi(m, i) \cdot P[Y_2 = (n, B) | Y_1 = (m, i)] \\ &= \pi(n, B), \end{aligned}$$

which shows that (3.11) is the stationary distribution of  $Y$ . In particular,

$$(3.12) \quad \lim_{n \rightarrow \infty} \frac{S_n^{\mathcal{L}_2}}{n} = \sum_{n \in \mathbb{Z}} \pi(n, B) = \frac{1}{2},$$

almost surely. On the other hand, observe that if  $Y_0 = (0, A)$  then  $Y_n = (0, A)$  if and only if  $X_n^R = 0$  and  $n$  is divisible by four. Let  $\tau$  denote the first-return time of  $Y$  to  $(0, A)$  conditioned on having  $Y_0 = (0, A)$ . Using the Chomsky and Schützenberger method [7], the moment generating function of  $\tau$  may be determined explicitly to find that its radius of convergence is given by  $1/\sqrt{4p(1-p)}$ . Since this last quantity is strictly greater than one for  $0 < p < 1/2$ , the tail distribution of  $\tau$  decays exponentially fast and therefore  $\tau$  must have a finite second-moment. Part (b) in the proposition is now a direct consequence of (3.7) and (3.12).

## References

- [1] A. V. Aho and M. J. Corasick. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June 1975.
- [2] F. Bassino, J. Clément, J. Fayolle, and P. Nicodème. Counting occurrences for a finite set of words: an inclusion-exclusion approach. 2007. Proceedings of the 2007 Conference on Analysis of Algorithms.
- [3] E. A. Bender and F. Kochman. The distribution of subword counts is usually normal. *Eur. J. Comb.*, 14(4):265–275, 1993.
- [4] J. D. Biggins and C. Cannings. Markov renewal processes, counters and repeated sequences in Markov chains. *Adv. Appl. Prob.*, 19:521–545, 1987.
- [5] J. Bourdon and B. Vallée. Generalized pattern matching statistics. In *Colloquium on Mathematics and Computer Science : Algorithms and Trees*, Trends in Mathematics, pages 249–265. Birkhauser, 2002.
- [6] J. Bourdon and B. Vallée. Pattern matching statistics on correlated sources. In *Proc. of the 7th Latin American Symposium on Theoretical Informatics (LATIN’06)*, pages 224–237, Valdivia, Chile, 2006.
- [7] N. Chomsky and M. P. Schützenberger. The algebraic theory of context-free languages. *Computer Programming and Formal Languages*, pages 118–161, 1963.
- [8] S. Derisavi, H. Hermanns, and W. H. Sanders. Optimal state-space lumping in markov chains. *Inf. Process. Lett.*, 87(6):309–315, 2003.
- [9] R. Durrett. *Probability: theory and examples*. Duxbury Press, third edition, 2004.
- [10] P. Flajolet, W. Szpankowski, and B. Vallée. Hidden word statistics. *J. ACM*, 53(1):147–183, 2006.
- [11] H. U. Gerber and S.-Y. R. Li. The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stochastic Processes and their Applications*, 11(1):101–108, 1981.
- [12] L. J. Guibas and A. M. Odlyzko. Periods in strings. *J. Comb. Theory, Ser. A*, 30(1):19–42, 1981.
- [13] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 1979.
- [14] J. Kemeny and J. L. Snell. *Finite Markov Chains*. Van Nostrand Reinhold LTD., 1960.
- [15] S.-Y. R. Li. A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *The Annals of Probability*, 8(6):1171–1176, 1980.
- [16] M. Lladser. Minimal Markov chain embeddings of pattern problems. 2006. Proceedings of the 2007 Information Theory and Applications Workshop, University of California, San Diego.
- [17] M. Lladser, M. D. Betterton, and R. Knight. Multiple pattern matching: A Markov chain approach. *Journal of Mathematical Biology*, 56:51–92, 2008.
- [18] P. Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. *Fundamenta Informaticae*, 56(1-2):71–88, 2003.
- [19] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–617, 2002.
- [20] R. Paige and R. E. Tarjan. Three partition refinement algorithms. *SIAM J. Comput.*, 1987.
- [21] V.I. Pozdnyakov and M. Kulldorff. Waiting times for patterns and a method of gambling teams. *American Mathematical Monthly*, 113(2):134–143, 2006.
- [22] M. Régnier and A. Denise. Rare events and conditional events on random strings. *DMTCS*, 6(2):191–214, 2004.
- [23] M. Régnier and W. Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algoritmica*, 22(4):631–649, 1998.