Analysis of patterns and minimal embeddings of non-Markovian sequences

Manuel.Lladser@Colorado.EDU

Department of Applied Mathematics
University of Colorado
Boulder

AofA - April 13 2008

NOTATION & TERMINOLOGY.

$\mathcal{A}$ is a finite **alphabet**

$\mathcal{A}^*$ is the set of all words of finite length

A **language** is a set $\mathcal{L} \subset \mathcal{A}^*$

$X = (X_n)_{n\geq 1}$ is a sequence of $\mathcal{A}$-valued random variables

X may be **non-Markovian**

$X_1 \cdots X_l$ models a **random word of length** l

PARADIGM.

For various probabilistic models for X and languages $\mathcal{L}$ the **frequency statistics of** $\mathcal{L}$ *are asymptotically normal.*

$$S_n^{\mathcal{L}} := \begin{pmatrix} \textit{number of prefixes in } X_1 \cdots X_n \\ \textit{that belong to the language } \mathcal{L} \end{pmatrix}$$

The paradigm applies for:

- generalized patterns $\oplus$ i.i.d. models [BenKoch93]
- simple patterns $\oplus$ stationary Markovian models [RegSzp98]
- primitive patterns $\oplus$ k-order Markovian models [NicSalFla02, Nic03]
- primitive patterns $\oplus$ nice dynamical sources [BouVal02, BouVal06]
- hidden patterns $\oplus$ i.i.d. models [FlaSpaVal06]

THE MARKOV CHAIN EMBEDDING TECHNIQUE.

IF X is a homogeneous Markov chain

IF $\mathcal{L}$ is a regular language

IF $G = (V, \mathcal{A}, f, q, T)$ is a DFA that recognizes $\mathcal{L}$

IF the *embedding of* X *into* G i.e. the stochastic process $X_n^G := f(q, X_1 \cdots X_n)$ is a first-order homogenous Markov chain

THEN

$$S_n^{\mathcal{L}} = \left(\begin{array}{c} \textit{number of visits the embedded process} \\ X^G \ \textit{makes to}\ T\ \textit{in the first}\ n\textit{-steps} \end{array} \right)$$

EXAMPLE.

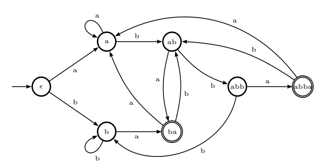
Consider a 1-st order Markov chain X such that

$$P[X_1 = a] = \mu; \qquad P[X_1 = b] = (1-\mu);$$

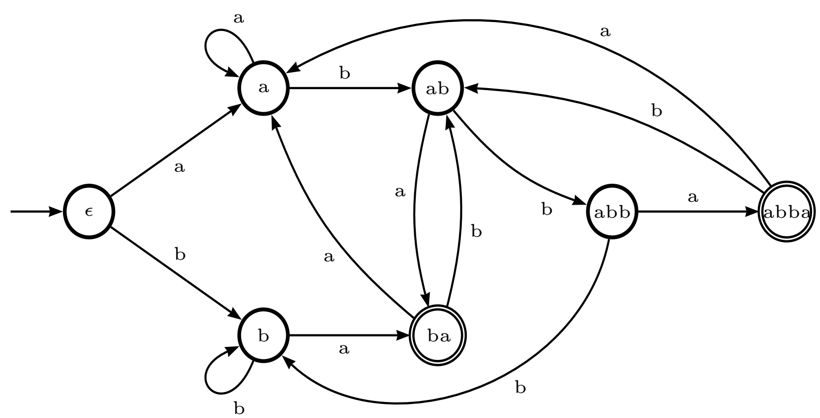
$$P[X_{n+1} = a \mid X_n = a] = p; \qquad P[X_{n+1} = b \mid X_n = a] = (1-p);$$

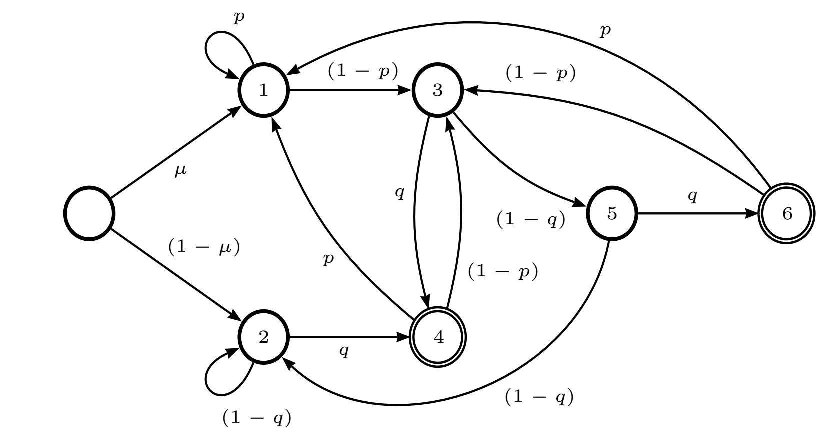
$$P[X_{n+1} = a \mid X_n = b] = q; \qquad P[X_{n+1} = b \mid X_n = b] = (1-q).$$

Then the embedding of X into the Aho-Corasick automaton



that recognizes matches with the regular expression $\{a,b\}^*\{ba, abba\}$ i.e. all words of the form $x = ...ba$ or $x = ...abba$ is a 1-st order Markov chain.


ϵ
a
b
ab
ba
abb
abba


1
2
3
4
5
6
μ
(1 − μ)
p
(1 − p)
q
(1 − q)

What about a completely general sequence X?

EXAMPLE. A seemingly unbiassed coin.

Let $0 < p < 1/2$

Consider the random binary sequence $X = (X_n)_{n \geq 1}$ such that

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

Question. *Is there a Markovian structure where X can be embedded into for analyzing the asymptotic distribution of the frequency statistics of a given language?*

GENERAL SETTING.

Given

- a possibly **non-Markovian** sequence X
- a possibly **non-regular** language $\mathcal{L}$
- a **transformation** $R : \mathcal{A}^* \to \mathcal{S}$

define X^R to be the stochastic process

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

Question 1. *What conditions are necessary and sufficient in order for X^R to be Markovian?*

Question 2. *Given a pattern $\mathcal{L}$, is there a transformation R such that X^R is Markovian but also informative of the distribution of the frequency statistics of $\mathcal{L}$?*

REMARK.

The Markovianity or non-Markovianity of

$$X_n^R := R(X_1 \cdots X_n), \quad n \geq 1$$

does not really depend on the range of R

The above motivates to think of $R : \mathcal{A}^* \to \mathcal{S}$ as an **equivalence relation** over $\mathcal{A}^*$:

$$u \, R \, v \Longleftrightarrow R(u) = R(v)$$

- $R(u)$ is the unique equivalence class of R that contains u
- $c \in R$ means that c is an equivalence class of R

DEFINITION. X *is embedable w.r.t.* R provided that for all $u, v \in \mathcal{A}^*$ and $c \in R$, if $u\,R\,v$ then

$$\sum_{\alpha \in \mathcal{A}:R(u\alpha)=c} P[X = u\alpha... \mid X = u...] = \sum_{\alpha \in \mathcal{A}:R(v\alpha)=c} P[X = v\alpha... \mid X = v...]$$

DEFINITION. X *is embedable w.r.t.* R provided that for all $u, v \in \mathcal{A}^*$ and $c \in R$, if $u \, R \, v$ then

$$\sum_{\alpha \in \mathcal{A} : R(u\alpha) = c} P[X = u\alpha ... \mid X = u...] = \sum_{\alpha \in \mathcal{A} : R(v\alpha) = c} P[X = v\alpha ... \mid X = v...]$$

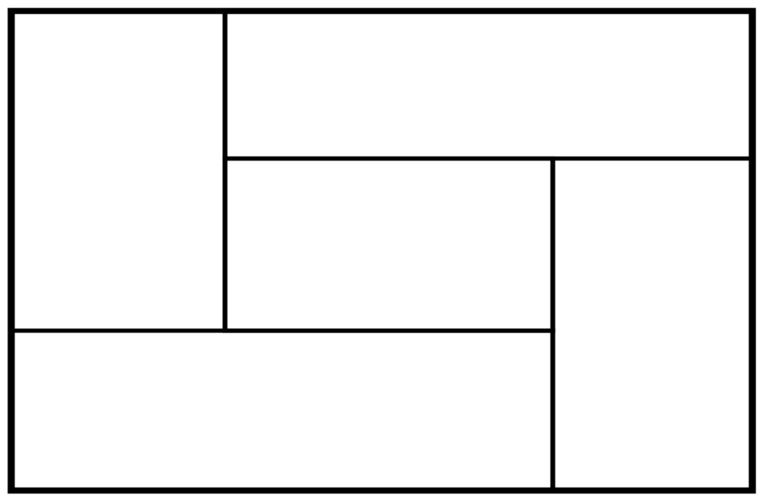


Figure. *Schematic partition of* $\{0, 1, 2\}^*$ *into equivalence classes*

DEFINITION. X *is embedable w.r.t.* R provided that for all $u, v \in \mathcal{A}^*$ and $c \in R$, if $u\,R\,v$ then

$$\sum_{\alpha \in \mathcal{A} : R(u\alpha) = c} P[X = u\alpha... \mid X = u...] = \sum_{\alpha \in \mathcal{A} : R(v\alpha) = c} P[X = v\alpha... \mid X = v...]$$

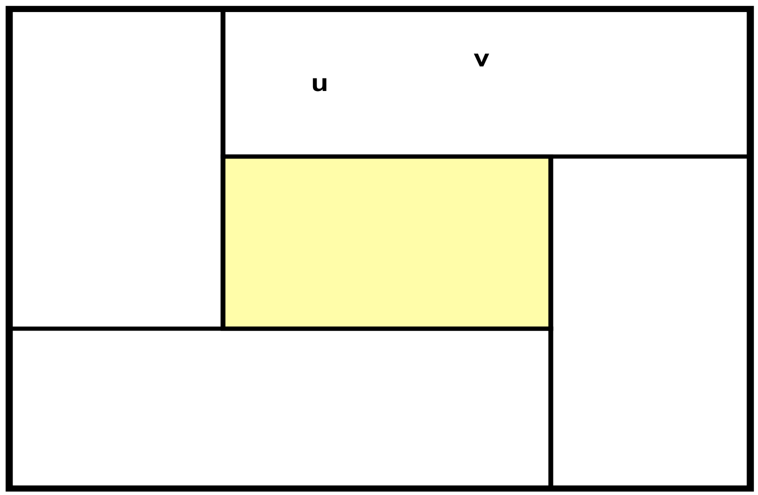


Figure. *Schematic partition of* $\{0, 1, 2\}^*$ *into equivalence classes*

DEFINITION. X *is embedable w.r.t.* R provided that for all $u, v \in \mathcal{A}^*$ and $c \in R$, if $u\,R\,v$ then

$$\sum_{\alpha\in\mathcal{A}:R(u\alpha)=c} P[X = u\alpha\ldots \mid X = u\ldots] = \sum_{\alpha\in\mathcal{A}:R(v\alpha)=c} P[X = v\alpha\ldots \mid X = v\ldots]$$

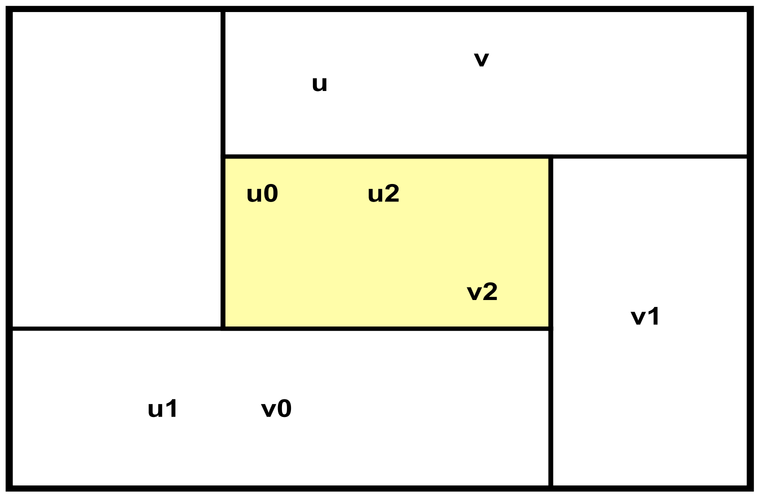


Figure. *Schematic partition of* $\{0, 1, 2\}^*$ *into equivalence classes*

DEFINITION. *X is embedable w.r.t. R* provided that for all $u, v \in \mathcal{A}^*$ and $c \in R$, if $u\,R\,v$ then

$$\sum_{\alpha\in\mathcal{A}:R(u\alpha)=c} P[X = u\alpha ... \mid X = u...] = \sum_{\alpha\in\mathcal{A}:R(v\alpha)=c} P[X = v\alpha ... \mid X = v...]$$

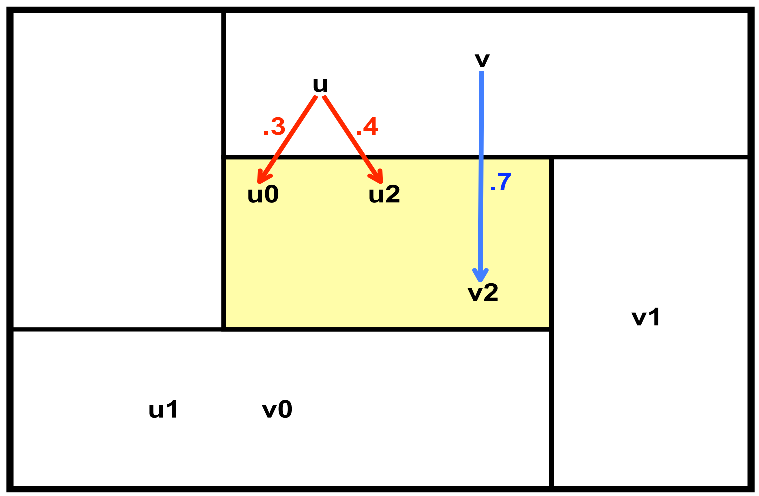
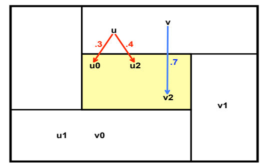


Figure. *Schematic partition of* $\{0, 1, 2\}^*$ *into equivalence classes*



THEOREM A. *X is embedable w.r.t. R if and only if, for $x \in \mathcal{A}^*$, if we condition on having $X = x$... then the stochastic process*

$$X_n^R := R(X_1 \cdots X_n), \quad n \geq |x|,$$

is a first-order homogeneous Markov chain with transition probabilities that do not depend on x

THEOREM B. *For each equivalence relation R in $\mathcal{A}^*$, there exists a unique coarsest refinement R' of R w.r.t. which X is embedable*

APPLICATION/QUESTION. *What is the smallest state-space for studying the frequency statistics of a language $\mathcal{L}$ in X?*

$$\begin{array}{rcccccccl}
\longrightarrow\ X & = & a & b & b & a & b & \ldots & \text{(original sequence)} \\
\longrightarrow\ X^R & = & 1 & 0 & 0 & 1 & 0 & \ldots & \text{(non-Markovian encoding)} \\
X^{R'} & = & 0 & 4 & 6 & 3 & 4 & \ldots & \text{(optimal Markovian encoding)} \\
X^Q & = & 6 & 3 & 18 & 15 & 10 & \ldots & \text{(any other Markovian encoding)}
\end{array}$$

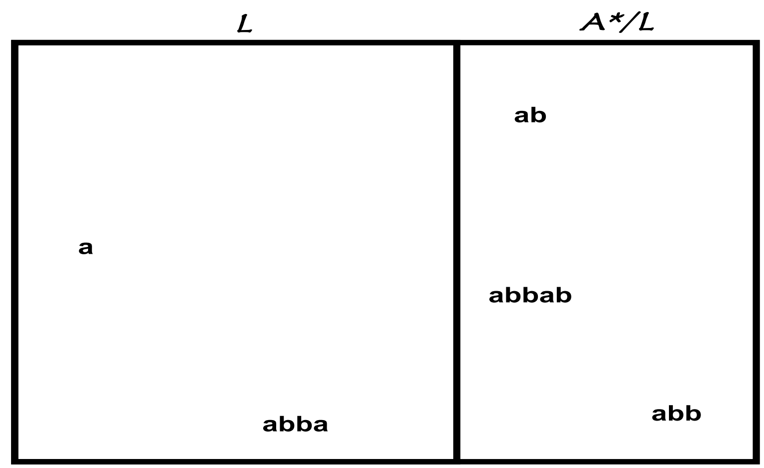


Figure. *Partition $R = \{\mathcal{L}, \mathcal{A}^* \setminus \mathcal{L}\}$ s.t. X^R is non-Markovian*

APPLICATION/QUESTION. *What is the smallest state-space for studying the frequency statistics of a language $\mathcal{L}$ in X?*

$$
\begin{array}{rllllllll}
\longrightarrow & X & = & a & b & b & a & b & \ldots & \text{(original sequence)} \\
 & X^R & = & 1 & 0 & 0 & 1 & 0 & \ldots & \text{(non-Markovian encoding)} \\
\longrightarrow & X^{R'} & = & 0 & 4 & 6 & 3 & 4 & \ldots & \text{(optimal Markovian encoding)} \\
 & X^Q & = & 6 & 3 & 18 & 15 & 10 & \ldots & \text{(any other Markovian encoding)}
\end{array}
$$

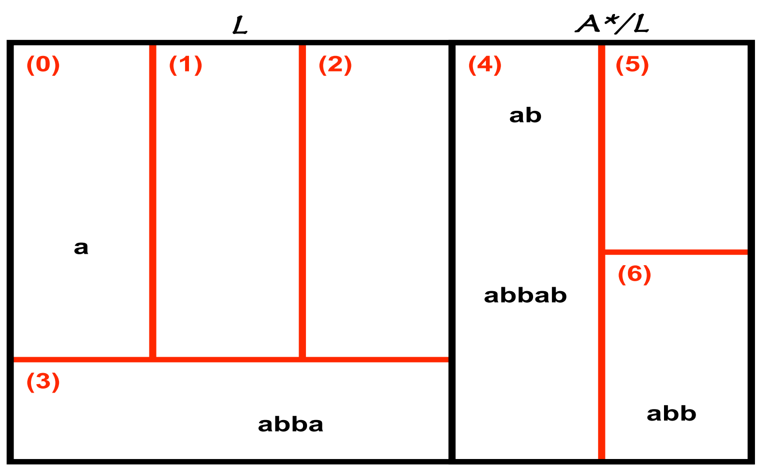


Figure. *Coarsest refinement R' of R w.r.t. which X is embedable*

APPLICATION/QUESTION. *What is the smallest state-space for studying the frequency statistics of a language $\mathcal{L}$ in X?*

$$
\begin{array}{lllllllll}
\longrightarrow & X & = & a & b & b & a & b & \dots & \text{(original sequence)} \\
 & X^R & = & 1 & 0 & 0 & 1 & 0 & \dots & \text{(non-Markovian encoding)} \\
 & X^{R'} & = & 0 & 4 & 6 & 3 & 4 & \dots & \text{(optimal Markovian encoding)} \\
\longrightarrow & X^Q & = & 6 & 3 & 18 & 15 & 10 & \dots & \text{(any other Markovian encoding)}
\end{array}
$$

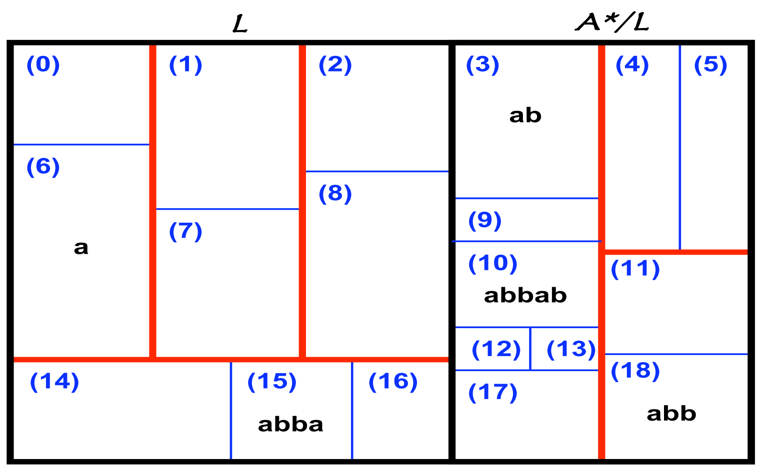


Figure. *Arbitrary refinement Q of R w.r.t. which X is embedable*

REMARK. The optimal refinement R' of R such that $X^{R'}$ is embedable is obtained through a limiting process: this makes it almost impossible to characterize de equivalence classes of R'

Motivated by this we will introduce an embedding which—while not as optimal—it is analytically tractable (!)

DEFINITION. The **Markov relation** induced by X into $\mathcal{A}^*$ is the equivalence relation defined as

$$uR^X v \Leftrightarrow (\forall w \in \mathcal{A}^*) : P[X = uw...\,|\,X = u...] = P[X = vw...\,|\,X = v...]$$

DEFINITION. The **Markov relation** induced by X into $\mathcal{A}^*$ is the equivalence relation defined as

$$uR^X v \Leftrightarrow (\forall w \in \mathcal{A}^*) : P[X = uw... \,|\, X = u...] = P[X = vw... \,|\, X = v...]$$

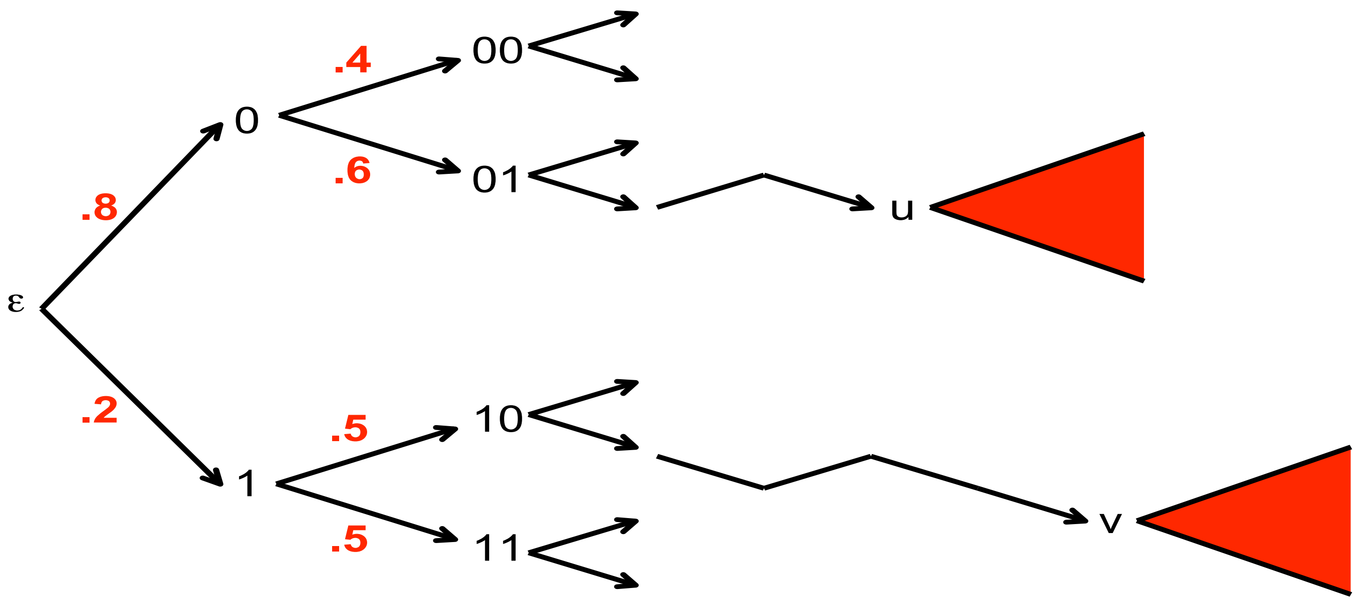
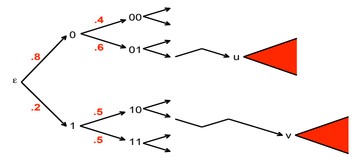


Figure. *Weighted tree visualization of definition with* $\mathcal{A} = \{0, 1\}$



An equivalence relation R is said to be **right-invariant** if for all $u, v \in \mathcal{A}^*$ and $\alpha \in \mathcal{A}$:

$$R(u) = R(v) \Longrightarrow R(u\alpha) = R(v\alpha)$$

THEOREM C. *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*

EXAMPLE. Back to the seemingly unbiassed coin.

For $0 < p < 1/2$, define

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

We aim to understand the frequency statistics of

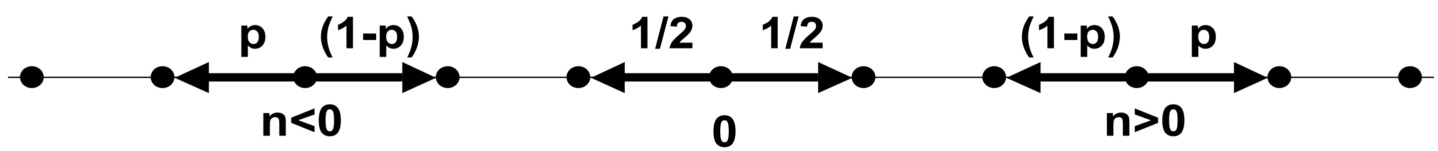
$$\begin{aligned} \mathcal{L}_1 &= \{0,1\}^*\{1\}, \\ \mathcal{L}_2 &= \{0\}^*\{1\}\{0\}^*(\{1\}\{0\}^*\{1\}\{0\}^*)^* \end{aligned}$$

within X

PROPOSITION. *$R : \{0,1\}^* \to \mathbb{Z}$ defined as*

$$R(x) \;=\; 2\left\{\sum_{i=1}^{|x|} x_i - \frac{|x|}{2}\right\} \;=\; \sum_{i=1}^{|x|} x_i - \sum_{i=1}^{|x|}(1 - x_i)$$

is a right-invariant refinement of R^X. In particular, $X_n^R := R(X_1 \cdots X_n)$ is a first-order homogeneous Markov chain



X^R is **recurrent**, with **period** 2. Because $0 < p < 1/2$, X^R is **positive recurrent**; in particular, there exists a stationary distribution π.
Observe that

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

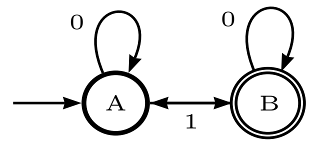
$$S_n^{\mathcal{L}_1} = \sum_{i=1}^{n} X_i$$

COROLLARY A. *If U and V are $\mathbb{Z}$-valued random variables such that*

$$\begin{aligned} P[U = n] &= 2 \cdot \pi(n), \quad n = 0(\bmod 2); \\ P[V = n] &= 2 \cdot \pi(n), \quad n = 1(\bmod 2); \end{aligned}$$

then for $\mathcal{L}_1 := \{0,1\}^*\{1\}$ it applies that

$$\begin{aligned} \lim_{\substack{n\to\infty \\ n=0(\bmod 2)}} 2n \cdot \left\{ \frac{S_n^{\mathcal{L}_1}}{n} - \frac{1}{2} \right\} &\stackrel{d}{=} U; \\ \lim_{\substack{n\to\infty \\ n=1(\bmod 2)}} 2n \cdot \left\{ \frac{S_n^{\mathcal{L}_1}}{n} - \frac{1}{2} \right\} &\stackrel{d}{=} V. \end{aligned}$$

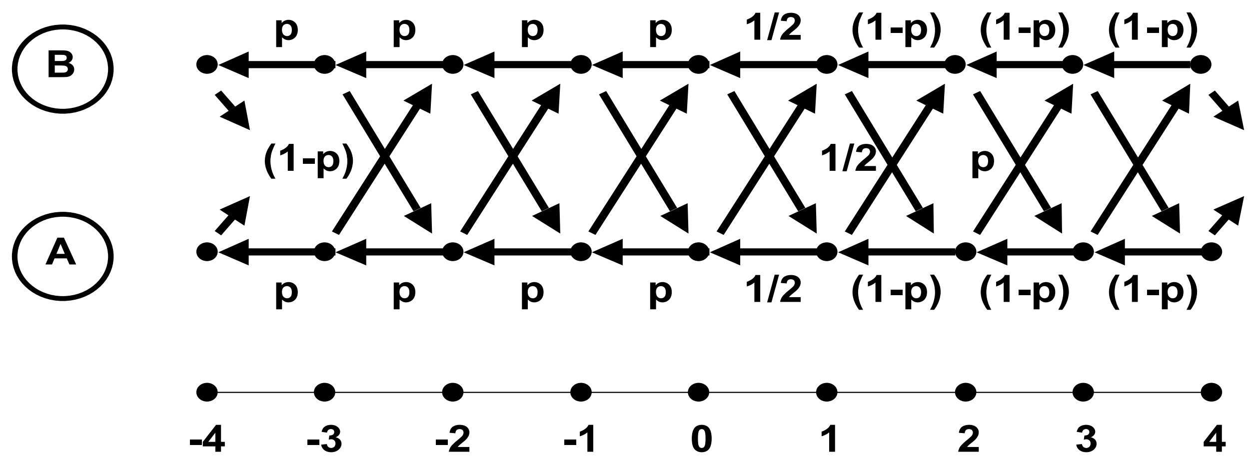
$\mathcal{L}_2$ is recognized by the automaton: 

According to the Mihill-Nerode theorem, $Q : \{0,1\}^* \to \{A, B\}$ defined as

$$Q(x) := \left(\begin{array}{l} \text{state in the automaton where the path} \\ \text{associated with } x \text{ ends when starting at } A \end{array} \right)$$

is right-invariant

Hence $R \times Q$ is also right-invariant and a refinement of R^X. In particular, $X_n^{R \times Q} := (X_n^R, X_n^Q)$ is a first-order homogeneous Markov chain



$X^{R\times Q}$ is **positive recurrent**, with **period** 4. Returning times to a state have finite second moment. This allows to use the central limit theorem for additive functionals of Markov chains to obtain the following result.

COROLLARY B. *There exists $\sigma > 0$ such that*

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

where W is a standard Normal random variable

CONCLUSION. For the same non-Markovian sequence X, non-Gaussian (discrete w/phases) and Gaussian limits are obtained for the frequency statistics of different regular languages

(More details in the 2008 ANALCO proceedings.)

... Thank you (!)