PATTERN MATCHING STATISTICS

FOR DYNAMICAL SOURCES

Brigitte Vallée, GREYC (CNRS and Université de Caen, France)

Joint work with Jérémie Bourdon, LINA (CNRS and Université de Nantes)

Statistics for pattern matching problems.

How frequent is a fixed pattern in a "typical" text ?

In order to discern....

meaningful informations from statistically expected phenomena

Which pattern?

A string, a set of strings, a sequence of patterns, a regular expression

Which text?

A random text emitted by a fixed source, more or less correlated:
a memoryless source, a Markov chain or a dynamical source.

Which measure?

– Number of occurrences Ω, number of occurrence positions C

– Mean value or distribution of these parameters.

Here, a distributional study of two main parameters

$$Y := C(\mathcal{E}) \text{ or } Y = \Omega(\mathcal{W})$$

For a text w, $Y(w)$ is the number of

– occurrence positions of a regular expression $\mathcal{E}$ in the text w

$$[\text{case } Y = C(\mathcal{E})]$$

– occurrences of a finite set of words $\mathcal{W}$ in the text w

$$[\text{case } Y = \Omega(\mathcal{W})]$$

. . . when the text w is emitted by a "dynamical" source.

Main Result. *$Y_n := Y|_{\Sigma^n}$ follows an asymptotic gaussian law. Moreover, there are precise estimates for the mean value and the variance of Y_n:* $\mathbb{E}[Y_n] \sim a_Y \cdot n, \qquad \mathbb{V}[Y_n] \sim b_Y \cdot n$ *where a_Y and b_Y are mathematically well-defined constants, (easily) computable for a (good) dynamical system.*

Generalization of well-known results already obtained when the source has a bounded memory (memoryless source, Markov chain).

Tool 1. Probabilistic tools, generating functions.

A source creates the text by emitting symbols from alphabet Σ.
For $w \in \Sigma^\star$, p_w is the probability that the source emits a prefix w.

Parameter Y. $Y(w)$ is the "cost" of the text w wrt a given pattern.
$Y : \Sigma^\star \to \mathbb{N}, \quad Y_n := Y|_{\Sigma^n}.$ Here $Y := C(\mathcal{E})$ or $Y := \Omega(\mathcal{W})$.

Bivariate generating function of parameter Y

$$F_Y(z,u) = \sum_{w\in\Sigma^\star} p_w \cdot u^{Y(w)} \cdot z^{|w|} = \sum_{n\geq 0} z^n \cdot \left(\sum_{w\in\Sigma^n} p_w \cdot u^{Y_n(w)} \right)$$

Fact. *If the moment generating function of* Y_n

$$\mathbb{E}[\exp(tY_n)] := \sum_{w\in\Sigma^n} p_w \cdot \exp[tY_n(w)] = [z^n]F_Y(z,e^t)$$

behaves as a n*-th "uniform quasi-power",*

this entails an asymptotic gaussian law for Y_n.

Tool 2. Automata. (for generating the pattern)

An automaton is defined by $(\Sigma, \mathcal{Q}, \mathcal{F}, s, \delta)$:

– the alphabet Σ,

– the set of states $\mathcal{Q} := \{0, 1, 2, \ldots, r-1\}$,

with the final states $\mathcal{F} \subset \mathcal{Q}$ and the initial state $s = 0$

– the transition function $\delta : \mathcal{Q} \times \Sigma \to \mathcal{Q}$, represented by the transition matrix $\mathcal{T}$ defined by

$$\mathcal{T}_{i,j} := \{m \in \Sigma; \quad \delta(i,m) = j\}, \quad \text{for } 0 \leq i, j \leq r-1.$$

The language $\mathcal{L}$, the language $\mathcal{L}_n := \mathcal{L} \cap \Sigma^n$, recognized by the automaton, are defined as

$$\mathcal{L}_n = S \cdot \mathcal{T}^n \cdot F, \quad \mathcal{L} = S \cdot \mathcal{T}^\star \cdot F,$$

where F is the final (column) vector, and s the initial (row) vector

$$F :=^t (f_1, \ldots, f_r), \quad \text{with } f_i = 1 \text{ iff } i \in \mathcal{F}, \qquad S = (1, 0, \ldots, 0).$$

Study of $C(\mathcal{E})$. Minimal automaton $\mathcal{A}$ which recognizes $\mathcal{L} = \Sigma^\star \cdot \mathcal{E}$.

Decomposition of $\mathcal{A}$ into the acyclic graph of its strongly connected components (SCC).

$\mathcal{E}$ is said to be simple if $\mathcal{A}$ has a unique SCC which contains $\mathcal{F}$.

In this case, $\mathcal{Q} = \mathcal{X} + \mathcal{Y}$, with $\mathcal{F} \subset \mathcal{Y}$.
If $\mathcal{X} \neq \oslash$, then $s \in \mathcal{X}$ and the transition matrix $\mathcal{T}$ decomposes as,

$$\mathcal{T} = \begin{pmatrix} \mathcal{M} & \mathcal{U} \\ 0 & \mathcal{R} \end{pmatrix} \quad \text{with} \quad \mathcal{M} := \mathcal{T}|_{\mathcal{X}\times\mathcal{Y}},\ \mathcal{R} := \mathcal{T}|_{\mathcal{Y}\times\mathcal{Y}},\ \mathcal{U} := \mathcal{T}|_{\mathcal{X}\times\mathcal{Y}},$$

$$\text{and} \quad \Sigma^\star = S_{\mathcal{X}} \cdot \mathcal{M}^\star \cdot [\mathbf{1}_{\mathcal{X}} + \mathcal{U} \cdot \mathcal{R}^\star \cdot \mathbf{1}_{\mathcal{Y}}].$$

If $\mathcal{X} = \oslash$, then $\mathcal{T} := \mathcal{R}$ and $\Sigma^\star = S \cdot \mathcal{R}^\star \cdot \mathbf{1}_{\mathcal{Y}}$.

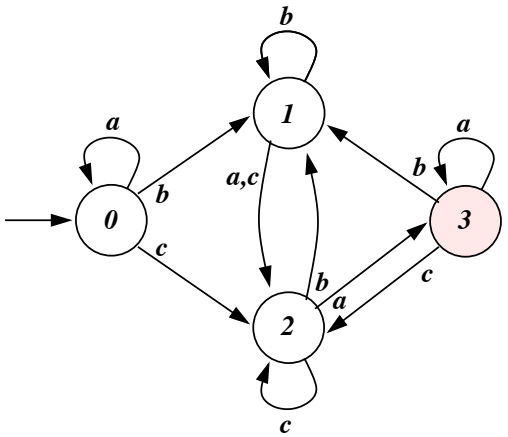
In both cases, for studying $C(\mathcal{E})$

the graph $(\mathcal{Y}, \mathcal{R})$ is called the useful part of the automaton.

Marked automaton for $C(\mathcal{E})$. We now "mark" the transitions of $\mathcal{T}$ which arrive at the final states and define two "marked" matrices $\mathcal{R}(u), \mathcal{U}(u)$ by the relations ($[\![\cdot]\!]$is Iverson's bracket)

$$\mathcal{R}(u)_{i,j} = u^{[\![j\in\mathcal{F}]\!]} \cdot \mathcal{R}_{i,j}, \qquad \mathcal{U}(u)_{i,j} = u^{[\![j\in\mathcal{F}]\!]} \cdot \mathcal{U}_{i,j},$$

Example : $\mathcal{E} = (ba|c)^+a^+$**.**



$${}^t\mathcal{R} := \begin{pmatrix} \{b\} & \{b\} & \{b\} \\ \{a,c\} & \{c\} & \{c\} \\ \oslash & \{a\} & \{a\} \end{pmatrix}, \qquad {}^t\mathcal{M} := \begin{pmatrix} \{a\} \end{pmatrix}, \qquad {}^t\mathcal{U} := \begin{pmatrix} \{b\} \\ \{c\} \\ \oslash \end{pmatrix}.$$

$${}^t\mathcal{R}(u) := \begin{pmatrix} \{b\} & \{b\} & \{b\} \\ \{a,c\} & \{c\} & \{c\} \\ \oslash & u\{a\} & u\{a\} \end{pmatrix}, \qquad {}^t\mathcal{U}(u) := {}^t\mathcal{U}$$

Case $\Omega(\mathcal{W})$ with $\ell+1 :=$ the maximum length of $w \in \mathcal{W}$.

The de Bruijn automaton $\mathcal{B}$ relative to alphabet Σ and length ℓ describes a "sliding window" of length ℓ. It reads a symbol at each stage, and keeps in memory the last ℓ letters read.

Set of States $\mathcal{Q} = \Sigma^\ell$ = Initial states = Final States.
Transition function $\delta(b, m) = c$: when the symbol m is read, in a state $b \in \Sigma^\ell$, one erases the left symbol of b, which provides a word denoted by $\tau(b)$, and $c := \tau(b) \cdot m$.

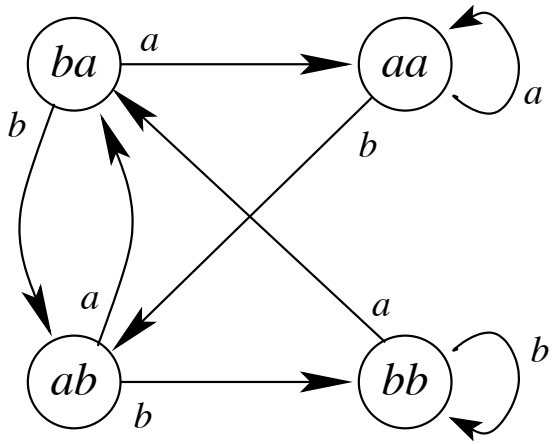
Matrix $\mathcal{B}$ defined by $\mathcal{B}_{b,c} = \{m, bm \in \Sigma c\}$

The number of new occurrences of $\mathcal{W}$ which arise when reading the last letter m in state b is $\phi(b, m) = \Omega(b \cdot m) - \Omega(b)$
We define the "marked" matrices $\mathcal{B}(u), \mathcal{S}(u)$ by

$$\mathcal{B}(u)_{b,c} := u^{\phi(b,m)} \cdot \mathcal{B}_{b,c} = u^{\Omega(bm)-\Omega(b)} \cdot \{m, bm \in \Sigma c\},$$

$$\mathcal{S}(u)_b := u^{\Omega(b)} \cdot \{b\}.$$

The De Bruijn automaton with $\ell = 2$ and $\Sigma = \{a, b\}$, its transition matrix, its use for $\mathcal{W} = \{ab, aab, aba\}$ with the marked matrix.



$$
{}^t\mathcal{B} = \begin{array}{c} \\ aa \\ ab \\ ba \\ bb \end{array}
\begin{array}{c}
\begin{array}{cccc} aa & ab & ba & bb \end{array} \\
\begin{pmatrix}
\{a\} & \oslash & \{a\} & \oslash \\
\{b\} & \oslash & \{b\} & \oslash \\
\oslash & \{a\} & \oslash & \{a\} \\
\oslash & \{b\} & \oslash & \{b\}
\end{pmatrix}
\end{array}.
$$

$$
{}^t\mathcal{B}(u) = \begin{pmatrix}
\{a\} & \oslash & \{a\} & \oslash \\
u^2 \cdot \{b\} & \oslash & u \cdot \{b\} & \oslash \\
\oslash & u \cdot \{a\} & \oslash & \{a\} \\
\oslash & \{b\} & \oslash & \{b\}
\end{pmatrix},
\qquad
{}^t\mathcal{S}(u) = \begin{pmatrix}
\{aa\} \\ u\{ab\} \\ \{ba\} \\ \{bb\}
\end{pmatrix}
$$

In both cases [$Y = C(\mathcal{E})$ or $Y = \Omega(\mathcal{W})$, the weighted language

$$\mathcal{L}_Y := \sum_{w \in \Sigma^\star} \{w\} \cdot u^{Y(w)}$$

admits an alternative expression which involves marked matrices, via their quasi-inverses denoted with a star: $\mathcal{A}^\star := \sum_{n \geq 0} \mathcal{A}^n$.

In case of $Y = \Omega(\mathcal{W})$, $\quad \mathcal{L}_Y := {}^t\mathcal{S}(u) \cdot \mathcal{B}(u)^\star \cdot \mathbf{1}_{\Sigma^\ell}$
involves the quasi-inverse of the marked matrix $\mathcal{B}(u)$ related to the de Bruijn automaton.

In case of $Y = C(\mathcal{E})$, $\quad \mathcal{L}_Y = (1, 0, \ldots 0) \cdot \mathcal{M}^\star \cdot [\mathbf{1}_{\mathcal{X}} + \mathcal{U}(u) \cdot \mathcal{R}(u)^\star \cdot \mathbf{1}_{\mathcal{Y}}]$
or $\quad \mathcal{L}_Y = (1, 0, \ldots 0) \cdot \mathcal{R}(u)^\star \cdot \mathbf{1}_{\mathcal{Y}}$
involves the quasi-inverse of the marked matrix $\mathcal{R}(u)$ related to the useful part of the automaton of $\Sigma^\star \mathcal{E}$

Tool 3. Dynamical sources (for generating the text)

A dynamical system $(\mathcal{I}, \mathcal{S})$ is defined by four elements:

(a) a finite alphabet Σ,

(b) a topological partition of $\mathcal{I} :=]0,1[$ with open intervals $\mathcal{I}_{m, m\in\Sigma}$,

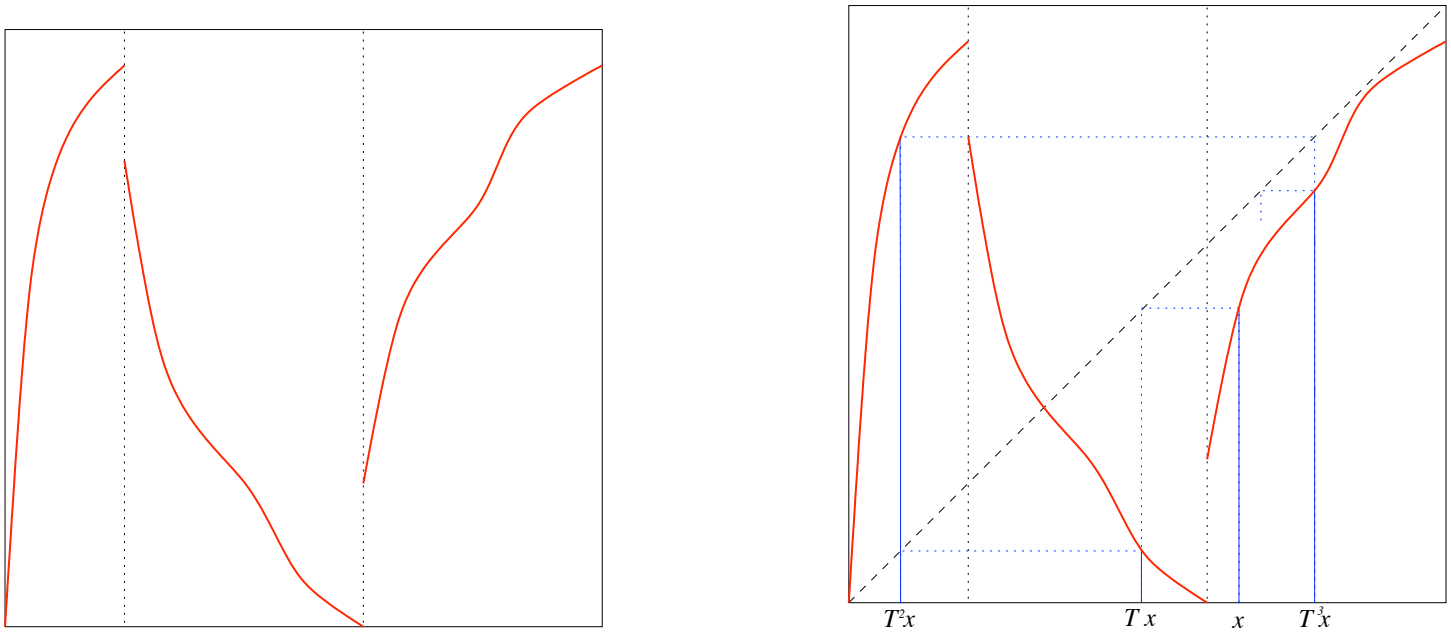
(c) an encoding mapping σ equal to m on each $\mathcal{I}_m$,

(d) a shift mapping S s.t. $S|_{\mathcal{I}_m}$ is a bijection of class $\mathcal{C}^2$ from $\mathcal{I}_m$ to $\mathcal{J}_m := S(\mathcal{I}_m)$.

This gives rise to a source: on an input x of $\mathcal{I}$, it outputs the word

$$M(x) := (\sigma x, \sigma Sx, \sigma S^2 x, \ldots).$$

When an initial density f is chosen on $\mathcal{I}$,
this induces (via the mapping M) a probabilistic model on Σ^∞.



A dynamical system, with $\Sigma = \{a, b, c\}$ and a word $M(x) = (c, b, a, c \ldots)$.

Correlations between symbols due to

– the geometry of the branches [position of $S(\mathcal{I}_m)$ wrt $\mathcal{I}_\ell$] describes the set $s(m)$ of possible successors of the symbol m.

Particular cases: – Complete systems $S(\mathcal{I}_m) = \mathcal{I}$

– Markovian systems $S(\mathcal{I}_m) =$ union of some $\mathcal{I}_\ell$

give rise to a finite characterization of $s(m)$.

Topological mixing.

$$\forall (b,e) \in \Sigma^2,\ \exists n_0 \geq 1 \ st\ \forall n \geq n_0,\ one\ has:\ \mathcal{I}_b \cap S^{-n}(\mathcal{I}_e) \neq \emptyset$$

"There is a word of length n which begins with b and ends with e".

– the shape of the branches [derivatives of the branches]: description of the evolution of the distribution.

Less correlated systems correspond to systems with affine branches.

Expansiveness.

$$\exists \delta < 1 \ st\ ,\ \forall m \in \Sigma,\ \forall x \in \mathcal{I},\ one\ has:\ |h'_m(x)| \leq \delta < 1.$$

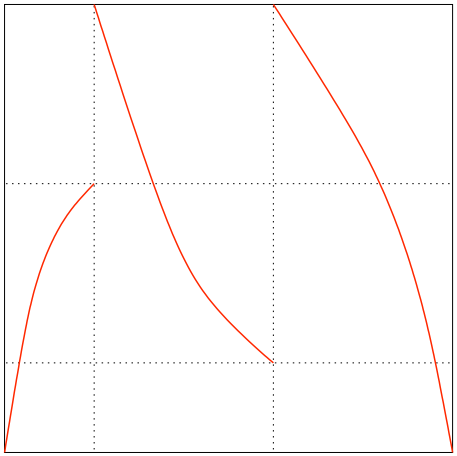
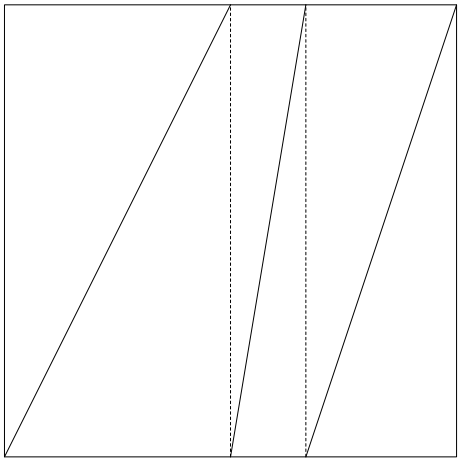
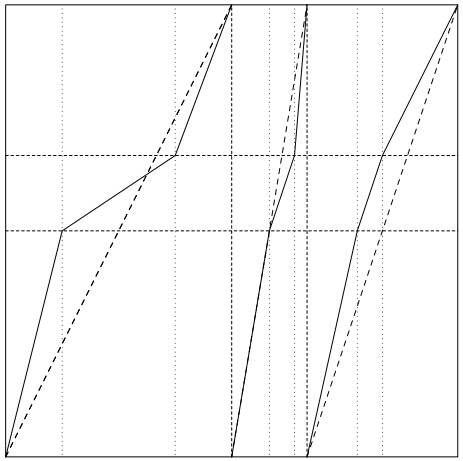
Particular cases...

A memoryless source:= a complete system with affine branches and uniform initial density

A Markov chain:= a Markovian system with affine branches, with an initial density which is constant on each $\mathcal{I}_m$.

Examples.

A Markovian system, a memoryless source, a Markov chain.

Generating operators for probabilities (I) Density transformer

Given an initial density $f = f_0$ on $\mathcal{I}$, what is the density f_1 on $\mathcal{I}$ after one iteration? the density f_n after n iterations?

$$f_1(x) = \sum_{m \in \Sigma} |h'_m(x)| \cdot f_0 \circ h_m(x) \cdot \mathbb{1}_{\mathcal{J}_m}(x) = \mathbf{G}[f_0](x),$$

where h_m is the inverse branch of $S|_{\mathcal{I}_m} : \mathcal{I}_m \to \mathcal{J}_m$.

$\mathbf{G}$ is called the density transformer.

More generally, $f_n(x) = \mathbf{G}^n[f_0](x)$ with

$$\mathbf{G}^n[f](x) = \sum_{w \in \Sigma^n} |h'_w(x)| \cdot f \circ h_w(x) \cdot \mathbb{1}_{\mathcal{J}_w}(x)$$

Here, the inverse branches of S^n are indexed by Σ^n, and for any $w := m_1 m_2 \dots m_n$, the mapping $h_w := h_{m_1} \circ h_{m_2} \circ \dots \circ h_{m_n}$ is a $\mathcal{C}^2$ bijection from $\mathcal{J}_w$ to $\mathcal{I}_w$.

If w is not produced by the source, then $\mathcal{J}_w = \oslash$.

For $w \in \Sigma^\star$, the component operator $\mathbf{G}_{[w]}$ defined as

$$\mathbf{G}_{[w]}[f](x) := |h'_w(x)| \cdot f \circ h_w(x) \cdot \mathbb{1}_{\mathcal{J}_w}(x)$$

has two main properties:

– *Generation.* $\mathbf{G}_{[w]}$ generates the probability p_w

$$p_w = \int_{\mathcal{I}_w} f(t)dt = \int_{\mathcal{I}} |h'_w(x)| \cdot f \circ h_w(x) \cdot \mathbb{1}_{\mathcal{J}_w}(x)dx = \int_{\mathcal{I}} \mathbf{G}_{[w]}[f](x)dx.$$

– *Multiplicativity.* The relation $\mathbf{G}_{[w \cdot w']} = \mathbf{G}_{[w']} \circ \mathbf{G}_{[w']}$ generalizes the equality $p_{w \cdot w'} = p_w \cdot p_w$, no longer true when the source has memory.

The generating operator of a language $\mathcal{L}$ is just

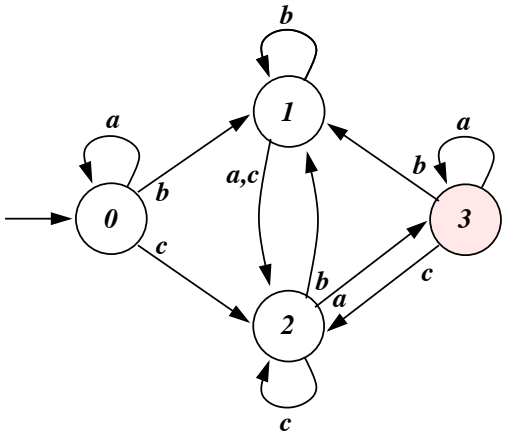
$$\mathbf{L}(z) := \sum_{w \in \mathcal{L}} z^{|w|} \cdot \mathbf{G}_{[w]},$$

and the generating operator of $\mathcal{L} \times \mathcal{M}$ is $\mathbf{M}(z) \circ \mathbf{L}(z)$.

For generating both the source and the pattern, we use the marked matrix operator $\mathbb{A}(u)$, obtained from the marked automaton $\mathcal{A}(u)$ by replacing languages by the generating operators of languages.

First example:

The marked matrix operators for $C(\mathcal{E})$ with $\mathcal{E} = (ba|c)^+a^+$.



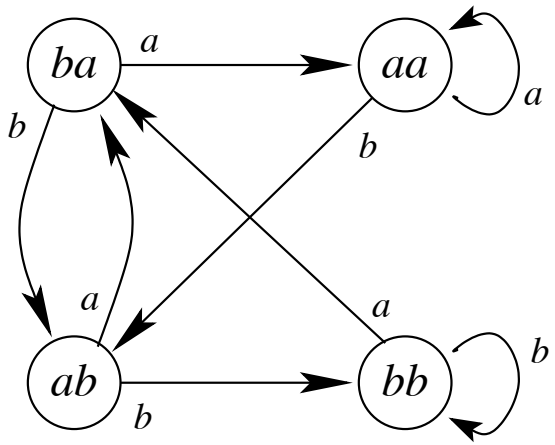
$$ {}^t\mathcal{R}(u) := \begin{pmatrix} \{b\} & \{b\} & \{b\} \\ \{a,c\} & \{c\} & \{c\} \\ \oslash & u\{a\} & u\{a\} \end{pmatrix}, $$

$$ {}^t\mathcal{M} := (\,\{a\}\,), \qquad {}^t\mathcal{U}(u) := \begin{pmatrix} \{b\} \\ \{c\} \\ \oslash \end{pmatrix}. $$

$$ \mathbb{R}(u) := \begin{pmatrix} \mathbf{G}_{[b]} & \mathbf{G}_{[b]} & \mathbf{G}_{[b]} \\ \mathbf{G}_{[a]} + \mathbf{G}_{[c]} & \mathbf{G}_{[c]} & \mathbf{G}_{[c]} \\ 0 & u \cdot \mathbf{G}_{[a]} & u \cdot \mathbf{G}_{[a]} \end{pmatrix}, \quad \mathbb{U}(u) := \begin{pmatrix} \mathbf{G}_{[b]} \\ \mathbf{G}_{[c]} \\ 0 \end{pmatrix} $$

Second example:

The marked matrix operators for $C(\mathcal{W})$ with $\mathcal{W} = \{ab, aab, aba\}$.



$$^t\mathcal{B}(u) = \begin{pmatrix} \{a\} & \oslash & \{a\} & \oslash \\ u^2 \cdot \{b\} & \oslash & u \cdot \{b\} & \oslash \\ \oslash & u \cdot \{a\} & \oslash & \{a\} \\ \oslash & \{b\} & \oslash & \{b\} \end{pmatrix}$$

$$\mathbb{B}(u) = \begin{pmatrix} \mathbf{G}_{[a]} & 0 & \mathbf{G}_{[a]} & 0 \\ u^2 \cdot \mathbf{G}_{[b]} & 0 & u \cdot \mathbf{G}_{[b]} & 0 \\ 0 & u \cdot \mathbf{G}_{[a]} & 0 & \mathbf{G}_{[a]} \\ 0 & \mathbf{G}_{[b]} & 0 & \mathbf{G}_{[b]} \end{pmatrix}, \qquad \mathbb{S}(u) = \begin{pmatrix} \mathbf{G}_{[aa]} \\ u\mathbf{G}_{[ab]} \\ \mathbf{G}_{[ba]} \\ \mathbf{G}_{[bb]} \end{pmatrix}$$

With multiplicative properties of the marked matrix operator, there is a translation of equalities between languages into equalities between generating operators.

Previously, with marked automata $\mathcal{A}(u)$

$$\mathcal{L}_Y := \sum_{w \in \Sigma^\star} \{w\} \cdot u^{Y(w)} = \mathcal{S}(u) \cdot \mathcal{A}(u)^\star \cdot \mathbf{1}$$

Now, with marked matrix operators $\mathbb{A}(u)$

$$\sum_{w \in \Sigma^\star} \mathbf{G}_{[w]} \cdot u^{Y(w)} \cdot z^{|w|} = {}^t\mathbf{1} \cdot (I - z\mathbb{A}(u))^{-1} \circ \mathbb{S}(z, u)$$

And, finally, with generation properties of probabilities,

$$F_Y(z, u) = \sum_{w \in \Sigma^\star} p_w \cdot u^{Y(w)} \cdot z^{|w|} = \int_{\mathcal{I}} {}^t\mathbf{1} \cdot (I - z\mathbb{A}(u))^{-1} \circ \mathbb{S}(z, u)[g](t)dt$$

The **asymptotic behaviour** of $\mathbb{E}[\exp(tY_n)] = [z^n]F_Y(z, e^t)$ is related to **singularities** of $z \mapsto (I - z\mathbb{A}(e^t))^{-1}$.

The density transformer $\mathbf{G}$ encapsulates various properties of the dynamical system. For most of DS, $\mathbf{G}$ has an eigenvalue equal to 1, and $z = 1$ is a singularity of $(I - z\mathbf{G})^{-1}$.

Fact. *If the system is topologically mixing and expansive,*

Topological mixing. *[geometry of the branches]*

$$\forall (b, e) \in \Sigma^2, \exists n_0 \geq 1 \text{ st } \forall n \geq n_0, \text{ one has: } \mathcal{I}_b \cap S^{-n}(\mathcal{I}_e) \neq \emptyset$$

"There is a word of length n which begins with b and ends with e".

Expansiveness. *[shape of the branches]*

$$\exists \delta < 1 \text{ st } , \forall m \in \Sigma, \forall x \in \mathcal{I}, \text{ one has: } |h'_m(x)| \leq \delta < 1.$$

Then $\mathbf{G}$ has good properties: when acting on a convenient functional space, it has a unique dominant eigenvalue equal to 1 and the quasi-inverse $(I - z\mathbf{G})^{-1}$ has a pôle at $z = 1$

Remark. For t near 0, $(I - z\mathbb{A}(e^t))^{-1}$ is a perturbation of $(I - z\mathbb{A}(1))^{-1}$, whose behaviour is quite similar to $(I - z\mathbf{G})^{-1}$.

Then, for t near 0 and z near 1, $z \mapsto (I - z\mathbb{A}(e^t))^{-1}$ has also a pôle at $z = 1/\lambda(e^t)$, where $\lambda(u)$ is the dominant eigenvalue of $\mathbb{A}(u)$.

Finally, $$[z^n](I - z\mathbb{A}(e^t))^{-1} = \Theta(\lambda^n(e^t)),$$
and the moment generating function $\mathbb{E}[\exp[tY_n]$ behaves as a n-th "quasi-power"...

We have exhibited the asymptotic gaussian law for Y_n.