Pattern Matching on Correlated Sources

Jérémie Bourdon and Brigitte Vallée

LINA, Nantes and GREYC/CNRS, Caen

march 23, 2006

Random Sources (of texts)

A random source is a process that construct random words. Each words w is produced with probability p_w such that $\sum_{|w|=k} p_w = 1$

- Uniform sources: each symbol is produced independently with a uniformly probability.
- Memoryless sources: each symbol is produced independently with a fixed probability.
- (Hidden) Markov chains: each symbol is produced with a bounded memory on the past.
- Dynamical Sources, mixing sources,...: the memory is not bounded.

Random Sources (of texts)

A random source is a process that construct random words. Each words w is produced with probability p_w such that $\sum_{|w|=k} p_w = 1$

- Uniform sources: each symbol is produced independently with a uniformly probability.
- Memoryless sources: each symbol is produced independently with a fixed probability.
- (Hidden) Markov chains: each symbol is produced with a bounded memory on the past.
- Dynamical Sources, mixing sources,...: the memory is not bounded.

Random Sources (of texts)

A random source is a process that construct random words. Each words w is produced with probability p_w such that $\sum_{|w|=k} p_w = 1$

- Uniform sources: each symbol is produced independently with a uniformly probability.
- Memoryless sources: each symbol is produced independently with a fixed probability.
- (Hidden) Markov chains: each symbol is produced with a bounded memory on the past.
- Dynamical Sources, mixing sources,...: the memory is not bounded.

Random Sources (of texts)

A random source is a process that construct random words. Each words w is produced with probability p_w such that $\sum_{|w|=k} p_w = 1$

- Uniform sources: each symbol is produced independently with a uniformly probability.
- Memoryless sources: each symbol is produced independently with a fixed probability.
- (Hidden) Markov chains: each symbol is produced with a bounded memory on the past.
- Dynamical Sources, mixing sources,...: the memory is not bounded.

Random Sources (of texts)

A random source is a process that construct random words. Each words w is produced with probability p_w such that $\sum_{|w|=k} p_w = 1$

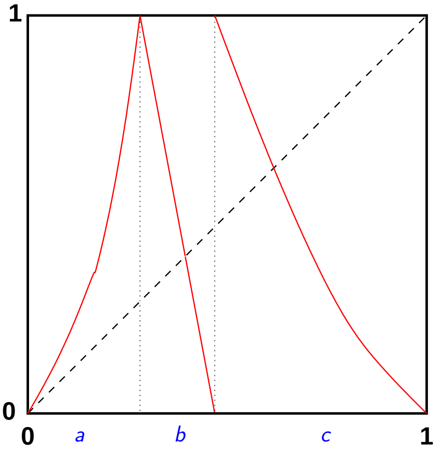
- Uniform sources: each symbol is produced independently with a uniformly probability.
- Memoryless sources: each symbol is produced independently with a fixed probability.
- (Hidden) Markov chains: each symbol is produced with a bounded memory on the past.
- Dynamical Sources, mixing sources,...: the memory is not bounded.

Random Sources (of texts)

A random source is a process that construct random words. Each words w is produced with probability p_w such that $\sum_{|w|=k} p_w = 1$

- Uniform sources: each symbol is produced independently with a uniformly probability.
- Memoryless sources: each symbol is produced independently with a fixed probability.
- (Hidden) Markov chains: each symbol is produced with a bounded memory on the past.
- Dynamical Sources, mixing sources, . . . : the memory is not bounded.

Dynamical sources



$M(x) = caccab\ldots$

Deterministic mechanism:

1) an alphabet Σ

2) an encoding function σ

3) A shift function T

Random choice:

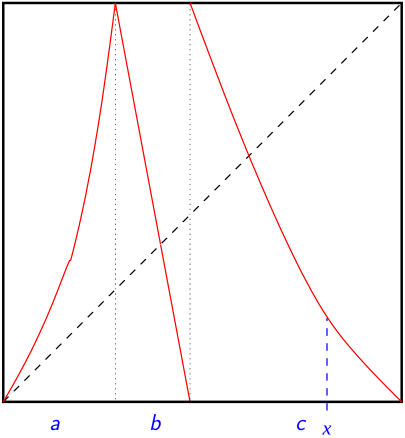
4) An initial density f

Word produced:

$$M(x) := (\sigma(x), \sigma(Tx), \sigma(T^2x), \ldots)$$

Fundamental intervals:

$$I_w = \{x | M(x) \text{ begins with } w\}.$$



$M(x) = caccab\ldots$

Deterministic mechanism:

1) an alphabet Σ

2) an encoding function σ

3) A shift function T

Random choice:

4) An initial density f

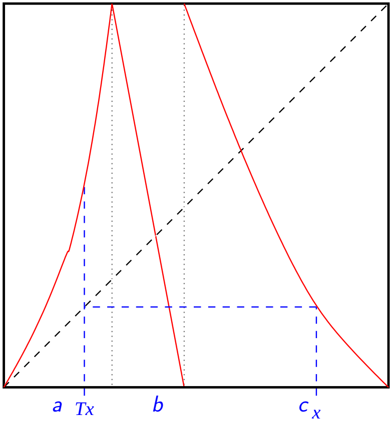
Word produced:

$$M(x) := (\sigma(x), \sigma(Tx), \sigma(T^2x), \ldots)$$

Fundamental intervals:

$$I_w = \{x | M(x) \text{ begins with } w\}.$$

Dynamical sources



$M(x) = caccab\ldots$

Deterministic mechanism:

1) an alphabet Σ

2) an encoding function σ

3) A shift function T

Random choice:

4) An initial density f

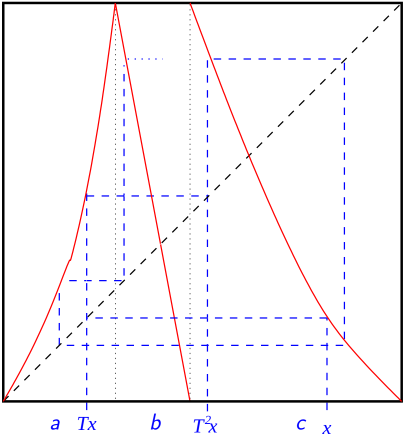
Word produced:

$$M(x) := (\sigma(x), \sigma(Tx), \sigma(T^2x), \ldots)$$

Fundamental intervals:

$$I_w = \{x | M(x) \text{ begins with } w\}.$$

Dynamical sources



$M(x) = caccab\ldots$

Deterministic mechanism:

1) an alphabet Σ

2) an encoding function σ

3) A shift function T

Random choice:

4) An initial density f

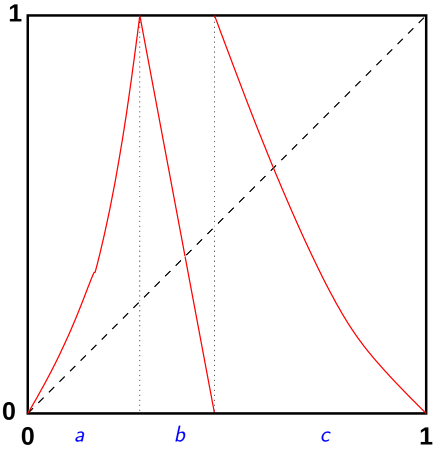
Word produced:

$$M(x) := (\sigma(x), \sigma(Tx), \sigma(T^2x), \ldots)$$

Fundamental intervals:

$$I_w = \{x | M(x) \text{ begins with } w\}.$$

Dynamical sources



$M(x) = caccab\ldots$

Deterministic mechanism:

1) an alphabet Σ

2) an encoding function σ

3) A shift function T

Random choice:

4) An initial density f

Word produced:

$$M(x) := (\sigma(x), \sigma(Tx), \sigma(T^2x), \ldots)$$

Fundamental intervals:

$$I_w = \{x | M(x) \text{ begins with } w\}.$$

Density transformer operator

f_0 is the initial density on $[0, 1]$

X R.V. of density f_0

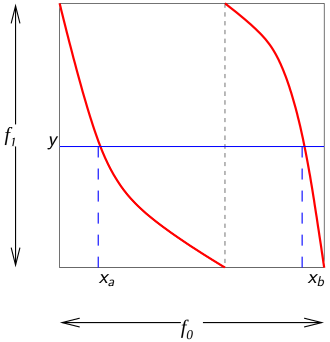
$\downarrow \quad T$

$T\ X$ R.V. of density f_1 (??)

$\downarrow \quad T$

$T^2\ X$ R.V. of density f_2 (??)

$\downarrow \quad T \quad \cdots$



$x_a := h_a(y)$

$x_b := h_b(y)$

Density transformer operator

f_0 is the initial density on $[0,1]$

X R.V. of density f_0

$\downarrow \quad T$

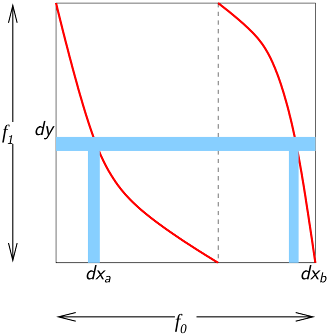
$T\ X$ R.V. of density f_1 (??)

$\downarrow \quad T$

$T^2\ X$ R.V. of density f_2 (??)

$\downarrow \quad T \quad \cdots$

$$f_1(y) = \sum_{m \in \Sigma} |h'_m(y)| f_0 \circ h_m(y) =: \mathbf{G}[f_0](y)$$



$x_a := h_a(y)$

$x_b := h_b(y)$

$\frac{dx_a}{dy} = h'_a(y)$

Dynamical sources and Information Theory

$$p_w := \int_{\mathcal{I}_w} f(t)dt =$$

$$M(z,u) = \sum p_w u^{C(w)} z^{|w|} \quad \longleftrightarrow \quad \mathbf{M}(z,u) = \sum \mathbf{G}_w u^{C(w)} z^{|w|}$$

$$p_{w \cdot w'} = p_w p_{w'} \quad \longleftrightarrow \quad \mathbf{G}_{w \cdot w'} = \mathbf{G}_{w'} \circ \mathbf{G}_w$$

$$p_w := \int_{\mathcal{I}_w} f(t)dt = \int_0^1 |h'_w| f \circ h_w(t) dt, \quad h_w = (T^{|w|})\|_{\mathcal{I}_w}^{-1}$$

$$M(z,u) = \sum p_w u^{C(w)} z^{|w|} \quad \leftrightarrow \quad \mathbf{M}(z,u) = \sum \mathbf{G}_w u^{C(w)} z^{|w|}$$

$$p_{w \cdot w'} = p_w p_{w'} \quad \leftrightarrow \quad \mathbf{G}_{w \cdot w'} = \mathbf{G}_{w'} \circ \mathbf{G}_w$$

Dynamical sources and Information Theory

$$p_w := \int_{\mathcal{I}_w} f(t)dt = \int_0^1 \mathbf{G}_w[f](t)dt$$

$$M(z,u) = \sum p_w u^{C(w)} z^{|w|} \qquad \leftrightarrow \qquad \mathbf{M}(z,u) = \sum \mathbf{G}_w u^{C(w)} z^{|w|}$$

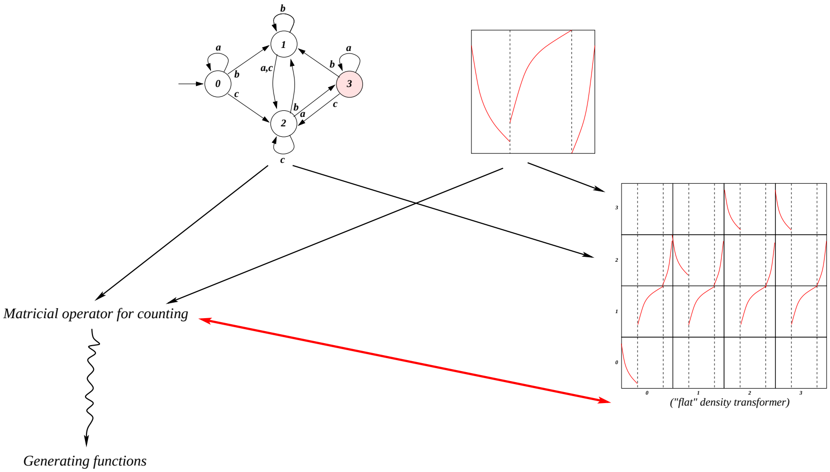
$$p_{w\cdot w'} = p_w p_{w'} \qquad \leftrightarrow \qquad \mathbf{G}_{w\cdot w'} = \mathbf{G}_{w'} \circ \mathbf{G}_w$$

Nice “decomposable” Sources

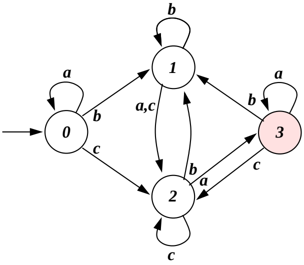
A Dynamical Source is called "decomposable" if its density transformer, when acting on an adapted Banach space, posses a unique dominant eigenvalue separated from the remainder of the spectrum by a “spectral gap”.
When the alphabet is finite, this property is satisfied when branches are expansives and when the source is topologically mixing.

$$\mathbf{G}^n = \lambda^n \mathbf{P} + \mathbf{N}^n, \qquad (\lambda = 1).$$

(on the functionnal space $BV(\mathcal{I})$ endowed with the norm $||f|| = \sup |f| + V(f)$).

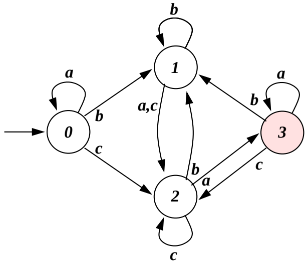

Matricial operator for counting
Generating functions
("flat" density transformer)

We first construct the DFA that recognize $\Sigma^\star\mathcal{R}$



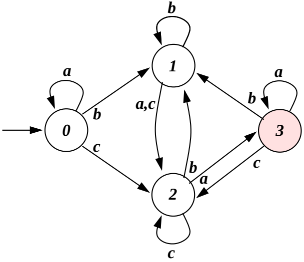
1 - A matricial operator for counting

We first construct the DFA that recognize $\color{red}{\Sigma^\star}\mathcal{R}$



$$\mathcal{T} = \begin{pmatrix} \{a\} & \emptyset & \emptyset & \emptyset \\ \{b\} & \{b\} & \{b\} & \{b\} \\ \{c\} & \{a,c\} & \{c\} & \{c\} \\ \emptyset & \emptyset & \{a\} & \{a\} \end{pmatrix}$$

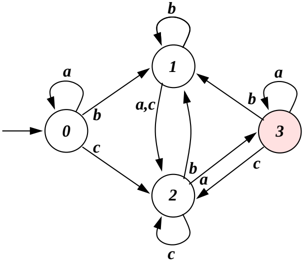
We first construct the DFA that recognize $\Sigma^\star\mathcal{R}$



$$T = \begin{pmatrix} p_a & 0 & 0 & 0 \\ p_b & p_b & p_b & p_b \\ p_c & p_a + p_c & p_c & p_c \\ 0 & 0 & u p_a & u p_a \end{pmatrix}$$

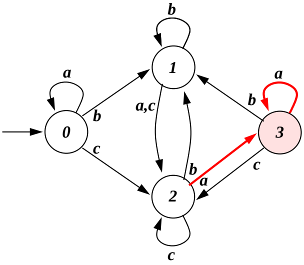
1 - A matricial operator for counting

We first construct the DFA that recognize $\Sigma^\star\mathcal{R}$



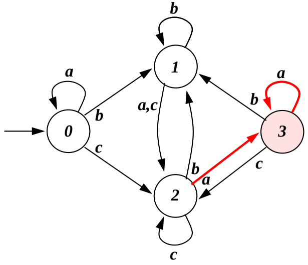
$$\mathbb{T} = \begin{pmatrix} \mathbf{G}_a & 0 & 0 & 0 \\ \mathbf{G}_b & \mathbf{G}_b & \mathbf{G}_b & \mathbf{G}_b \\ \mathbf{G}_c & \mathbf{G}_a + \mathbf{G}_c & \mathbf{G}_c & \mathbf{G}_c \\ 0 & 0 & u\mathbf{G}_a & u\mathbf{G}_a \end{pmatrix}$$

We first construct the DFA that recognize $\Sigma^\star\mathcal{R}$



$$\mathbb{T}(u) = \begin{pmatrix} \mathbf{G}_a & 0 & 0 & 0 \\ \mathbf{G}_b & \mathbf{G}_b & \mathbf{G}_b & \mathbf{G}_b \\ \mathbf{G}_c & \mathbf{G}_a + \mathbf{G}_c & \mathbf{G}_c & \mathbf{G}_c \\ 0 & 0 & u\mathbf{G}_a & u\mathbf{G}_a \end{pmatrix}$$

We first construct the DFA that recognize $\Sigma^\star\mathcal{R}$

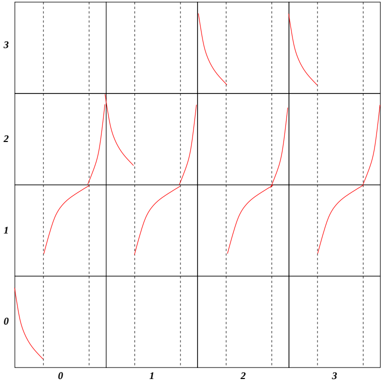


$$\mathbf{M}(z,u) = \sum_n (1,\cdots,1) \begin{pmatrix} \mathbf{G}_a & 0 & 0 & 0 \\ \mathbf{G}_b & \mathbf{G}_b & \mathbf{G}_b & \mathbf{G}_b \\ \mathbf{G}_c & \mathbf{G}_a + \mathbf{G}_c & \mathbf{G}_c & \mathbf{G}_c \\ 0 & 0 & u\mathbf{G}_a & u\mathbf{G}_a \end{pmatrix}^n \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} z^n$$

2 - A mixed source

$\mathcal{A}$ automaton with r states, $\delta(m,i)$ its transition function.
The mixed source is defined on interval $]0,r[$, for alphabet $\Sigma \times \{1,\ldots,r\}$, by

$$\mathcal{I}_{m,i} = \mathcal{I}_m + i, \qquad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(m,i), \qquad h_{m,i}(t) = h_m(t - \delta(m,i)) + i$$



2 - A mixed source

$\mathcal{A}$ automaton with r states, $\delta(m, i)$ its transition function.
The mixed source is defined on interval $]0, r[$, for alphabet $\Sigma \times \{1, \ldots, r\}$, by

$$\mathcal{I}_{m,i} = \mathcal{I}_m + i, \qquad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(m, i), \qquad h_{m,i}(t) = h_m(t - \delta(m, i)) + i$$

- The mixed alphabet is finite (of size $|\Sigma| \times r$)
- The mixed branches are expansives
- The mixed source is topologically mixing (if any state of the automaton can be reached for any other state)
- $\mathfrak{G}$ acts on $BV(]0, r[)$ and decomposes as

$$\mathfrak{G} = \lambda \mathfrak{P} + \mathfrak{N}$$

2 - A mixed source

$\mathcal{A}$ automaton with r states, $\delta(m,i)$ its transition function. The mixed source is defined on interval $]0,r[$, for alphabet $\Sigma \times \{1,\ldots,r\}$, by

$$\mathcal{I}_{m,i} = \mathcal{I}_m + i, \qquad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(m,i), \qquad h_{m,i}(t) = h_m(t - \delta(m,i)) + i$$

- The mixed alphabet is finite (of size $|\Sigma| \times r$)
- The mixed branches are expansives
- The mixed source is topologically mixing (if any state of the automaton can be reached for any other state)
- $\mathfrak{G}$ acts on $BV(]0,r[)$ and decomposes as

$$\mathfrak{G} = \lambda\mathfrak{P} + \mathfrak{N}$$

$\mathcal{A}$ automaton with r states, $\delta(m, i)$ its transition function. The mixed source is defined on interval $]0, r[$, for alphabet $\Sigma \times \{1, \ldots, r\}$, by

$$\mathcal{I}_{m,i} = \mathcal{I}_m + i, \qquad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(m, i), \qquad h_{m,i}(t) = h_m(t - \delta(m, i)) + i$$

- The mixed alphabet is finite (of size $|\Sigma| \times r$)
- The mixed branches are expansives
- The mixed source is topologically mixing (if any state of the automaton can be reached for any other state)
- $\mathfrak{G}$ acts on $BV(]0, r[)$ and decomposes as

$$\mathfrak{G} = \lambda \mathfrak{P} + \mathfrak{N}$$

2 - A mixed source

$\mathcal{A}$ automaton with r states, $\delta(m, i)$ its transition function. The mixed source is defined on interval $]0, r[$, for alphabet $\Sigma \times \{1, \dots, r\}$, by

$$\mathcal{I}_{m,i} = \mathcal{I}_m + i, \qquad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(m, i), \qquad h_{m,i}(t) = h_m(t - \delta(m, i)) + i$$

- The mixed alphabet is finite (of size $|\Sigma| \times r$)
- The mixed branches are expansives
- The mixed source is topologically mixing (if any state of the automaton can be reached for any other state)
- $\mathfrak{G}$ acts on $BV(]0, r[)$ and decomposes as

$$\mathfrak{G} = \lambda \mathfrak{P} + \mathfrak{N}$$

2 - A mixed source

$\mathcal{A}$ automaton with r states, $\delta(m, i)$ its transition function.
The mixed source is defined on interval $]0, r[$, for alphabet $\Sigma \times \{1, \ldots, r\}$, by

$$\mathcal{I}_{m,i} = \mathcal{I}_m + i, \qquad \mathcal{J}_{m,i} = \mathcal{J}_m + \delta(m, i), \qquad h_{m,i}(t) = h_m(t - \delta(m, i)) + i$$

- The mixed alphabet is finite (of size $|\Sigma| \times r$)
- The mixed branches are expansives
- The mixed source is topologically mixing (if any state of the automaton can be reached for any other state)
- $\mathfrak{G}$ acts on $BV(]0, r[)$ and decomposes as

$$\mathfrak{G} = \lambda\mathfrak{P} + \mathfrak{N}$$

3 - Matricial operator vs "flat" operator

$\mathfrak{G}$ and $\mathbb{T}$ are conjugated by Ψ,

$$\mathfrak{G} = \Psi \circ \mathbb{T} \circ \Psi^{-1},$$

où $\Psi : (BV(\mathcal{I}))^r \to BV(]0, r[)$,

$$\Psi({}^t(g_1, \ldots, g_r))(x) = \sum_{i=1}^{r} \mathbb{1}_{[i-1,i]}(x) \cdot g_i(x - i + 1)$$

$$\mathbb{T}(u) = \lambda(u)\mathbb{P}(u) + \mathbb{N}(u)$$

Analytical perturbation

3 - Matricial operator vs “flat” operator

$\mathfrak{G}$ and $\mathbb{T}$ are conjugated by Ψ,

$$\mathfrak{G} = \Psi \circ \mathbb{T} \circ \Psi^{-1},$$

où $\Psi : (BV(\mathcal{I}))^r \to BV(]0, r[)$,

$$\Psi({}^t(g_1, \ldots, g_r))(x) = \sum_{i=1}^{r} 1\!\!1_{[i-1,i]}(x) \cdot g_i(x - i + 1)$$

$$\mathbb{T}(u) = \lambda(u)\mathbb{P}(u) + \mathbb{N}(u)$$

Analytical perturbation

Results

$$\mathbb{E}\left[C_n\right] = \lambda'(1)c_1 n + o(n),$$
$$\mathbb{V}\left[C_n\right] = (\lambda''(1) + \lambda'(1) - (\lambda'(1))^2)c_2 n + o(n)$$

C_n follow asymptotically a gaussian law.