Pattern Matching on Correlated Sources

Jérémie Bourdon and Brigitte Vallée

LINA, Nantes and GREYC/CNRS, Caen

march 23, 2006
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.
A random source is a process that construct random words. Each word $w$ 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.
A random source is a process that constructs random words. Each word \( w \) 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.
A random source is a process that construct random words. Each word $w$ 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.
A random source is a process that constructs random words. Each word $w$ 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.
A random source is a process that constructs random words. Each word \( w \) 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

Deterministic mechanism:
1) an alphabet $\Sigma$
2) an encoding function $\sigma$
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

Deterministic mechanism:
1) an alphabet $\Sigma$
2) an encoding function $\sigma$
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 $\Sigma$
2) an encoding function $\sigma$
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

Deterministic mechanism:
1) an alphabet $\Sigma$
2) an encoding function $\sigma$
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 $\Sigma$
2) an encoding function $\sigma$
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:**
$$I_w = \{x | M(x) \text{ begins with } w\}.$$
$f_0$ is the initial density on $[0, 1]$

$X$ R.V. of density $f_0$

$\downarrow T$

$T \times X$ R.V. of density $f_1$

$\downarrow T$

$T^2 \times X$ R.V. of density $f_2$

$\downarrow T \cdots$

$x_a := h_a(y)$

$x_b := h_b(y)$

$f_0$
Density transformer operator

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

$X$ R.V. of density $f_0$

\[ \downarrow T \]

$TX$ R.V. of density $f_1$ (??)

\[ \downarrow T \]

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

\[ \downarrow T \ldots \]

\[ f_1(y) = \sum_{m \in \Sigma} |h'_m(y)| f_0 \circ h_m(y) =: G[f_0](y) \]
Dynamical sources and Information Theory

\[ p_w := \int_{I_w} f(t) dt = \]

\[ M(z, u) = \sum p_w u^{C(w)} z^{|w|} \quad \leftrightarrow \quad M(z, u) = \sum G_w u^{C(w)} z^{|w|} \]

\[ p_{w \cdot w'} = p_w p_{w'} \quad \leftrightarrow \quad G_{w \cdot w'} = G_{w'} \circ G_w \]
Dynamical sources and Information Theory

\[ p_w := \int_{I_w} f(t) dt = \int_0^1 |h'_w| f \circ h_w(t) dt, \quad h_w = (T^{|w|})_{II w}^{-1} \]

\[ M(z, u) = \sum p_w u^c(w) z^{|w|} \quad \leftrightarrow \quad M(z, u) = \sum G_w u^c(w) z^{|w|} \]

\[ p_w \cdot p'_w = p_w p'_w \quad \leftrightarrow \quad G_w \cdot G'_w = G'_w \circ G_w \]
\[ p_w := \int_{\mathcal{I}_w} f(t) \, dt = \int_0^1 G_w[f](t) \, dt \]

\[ M(z, u) = \sum p_w u^{C(w)} z^{|w|} \quad \leftrightarrow \quad M(z, u) = \sum G_w u^{C(w)} z^{|w|} \]

\[ p_{w \cdot w'} = p_w p_{w'} \quad \leftrightarrow \quad G_{w \cdot w'} = G_{w'} \circ G_w \]
A Dynamical Source is called "decomposable" if its density transformer, when acting on an adapted Banach space, possesses 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.

\[ G^n = \lambda^n P + N^n, \quad (\lambda = 1). \]

(on the functionnal space \( BV(I) \) endowed with the norm \( \| f \| = \sup |f| + V(f) \)).
Core of the method

Matricial operator for counting

Generating functions

Jérémie Bourdon and Brigitte Vallée

Pattern Matching on Correlated Sources
We first construct the DFA that recognize $\Sigma^* \mathcal{R}$.
We first construct the DFA that recognize $\Sigma^* \mathcal{R}$.
We first construct the DFA that recognize $\Sigma^* \mathcal{R}$.
We first construct the DFA that recognize $\Sigma^*\mathcal{R}$.
We first construct the DFA that recognize $\Sigma^* R$

\[
\mathbb{P}(u) = \begin{pmatrix}
G_a & 0 & 0 & 0 \\
G_b & G_b & G_b & G_b \\
G_c & G_a + G_c & G_c & G_c \\
0 & 0 & uG_a & uG_a
\end{pmatrix}
\]
We first construct the DFA that recognize $\Sigma^*\mathcal{R}$

\[ M(z, u) = \sum_{n} (1, \cdots, 1) \left( \begin{array}{cccc} G_a & 0 & 0 & 0 \\ G_b & G_b & G_b & G_b \\ G_c & G_a + G_c & G_c & G_c \\ 0 & 0 & uG_a & uG_a \end{array} \right)^n \left( \begin{array}{c} 1 \\ 0 \\ \vdots \\ 0 \end{array} \right) z^n \]
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

$$
I_{m,i} = I_m + i, \quad J_{m,i} = J_m + \delta(m, i), \quad h_{m,i}(t) = h_m(t - \delta(m, i)) + i
$$

![Graph showing the mixed source with $r$ states and transition function $\delta(m, i)$]

Jérémie Bourdon and Brigitte Vallée

Pattern Matching on Correlated Sources
A mixed source

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

$$I_{m,i} = I_{m} + i, \quad J_{m,i} = J_{m} + \delta(m, i), \quad h_{m,i}(t) = h_{m}(t - \delta(m, i)) + i$$

- 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)
- $G$ acts on $BV([0, r[)$ and decomposes as

$$G = \lambda P + N$$
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

\[
I_{m,i} = I_m + i, \quad J_{m,i} = J_m + \delta(m, i), \quad h_{m,i}(t) = h_m(t - \delta(m, i)) + i
\]

- 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)
- \( \mathcal{G} \) acts on \( BV([0, r[) \) and decomposes as

\[
\mathcal{G} = \lambda \mathcal{B} + \mathcal{N}
\]
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

\[
I_{m,i} = I_{m} + i, \quad J_{m,i} = J_{m} + \delta(m, i), \quad h_{m,i}(t) = h_{m}(t - \delta(m, i)) + i
\]

- 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)
- \( \mathcal{G} \) acts on \( BV([0, r[) \) and decomposes as

\[
\mathcal{G} = \lambda \mathcal{P} + \mathcal{N}
\]
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

\[
I_{m,i} = I_m + i, \quad J_{m,i} = J_m + \delta(m, i), \quad h_{m,i}(t) = h_m(t - \delta(m, i)) + i
\]

- 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)

\( G \) acts on \( BV([0, r[) \) and decomposes as

\[
G = \lambda P + N
\]
2 - A mixed source

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

\[
I_{m,i} = I_{m+i}, \quad J_{m,i} = J_{m+i} + \delta(m, i), \quad h_{m,i}(t) = h_m(t-\delta(m, i)) + i
\]

- 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)
- \( \mathcal{G} \) acts on \( BV(]0, r[) \) and decomposes as

\[
\mathcal{G} = \lambda \mathcal{P} + \mathcal{N}
\]
\( \mathcal{G} \) and \( \mathcal{T} \) are conjugated by \( \Psi \),

\[ \mathcal{G} = \Psi \circ \mathcal{T} \circ \Psi^{-1}, \]

\( \Psi : (BV(I))^r \to BV([0, r]) \),

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

\[ \mathcal{T}(u) = \lambda(u) \mathcal{P}(u) + \mathcal{N}(u) \]

Analytical perturbation
\( \mathcal{G} \) and \( T \) are conjugated by \( \Psi \),

\[
\mathcal{G} = \Psi \circ T \circ \Psi^{-1},
\]
où \( \Psi : (BV(\mathcal{I}))^r \rightarrow BV([0, r[), \)

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

\[
T(u) = \lambda(u)P(u) + N(u)
\]

Analytical perturbation
\[ \mathbb{E} [C_n] = \lambda' (1)c_1 n + o(n), \]

\[ \mathbb{V} [C_n] = (\lambda''(1) + \lambda'(1) - (\lambda'(1))^2)c_2 n + o(n) \]

\[ C_n \text{ follow asymptotically a gaussian law.} \]