Abres des contextes et sources dynamiques

P. Cénac, B. Chauvin, F. Paccaut*, N. Pouyanne

*LAMFA, Amiens

24 mars 2010

Plan de l'exposé

Processus à mémoire infinie

L'arbre des contextes a été introduit par Rissanen en 1983 ("A universal data compression system") pour modéliser une source générale. L'algorithme CONTEXT permet à partir de la source, d'estimer les contextes et les probabilités de transition. Une fois l'arbre des contextes reconstruit, que peut-on en dire ? Existence d'une mesure stationnaire, propriétés de cette mesure (mélange, récurrence)

Sources dynamiques

- Source (au sens de la théorie de l'information) : mécanisme qui produit des mots infinis sur un alphabet.
- Source dynamique : le mécanisme repose sur un système dynamique.
- Source à contextes : le mécanisme repose sur un arbre des contextes probabilisé.

Sources dynamiques

- Source (au sens de la théorie de l'information) : mécanisme qui produit des mots infinis sur un alphabet.
- Source dynamique : le mécanisme repose sur un système dynamique.
- Source à contextes : le mécanisme repose sur un arbre des contextes probabilisé.

Sources dynamiques

- Source (au sens de la théorie de l'information) : mécanisme qui produit des mots infinis sur un alphabet.
- Source dynamique : le mécanisme repose sur un système dynamique.
- Source à contextes : le mécanisme repose sur un arbre des contextes probabilisé.

Tries

- Tries généraux : hauteur, nombre de noeuds internes, longueur de cheminement. Méthodes de combinatoire analytique (Flajolet, Vallée, Fayolle) pour les sources dynamiques, méthodes probabilistes basées sur les propriétés de mélange de la source (Jacquet, Szpankowski).
- Trie des suffixes : niveau de saturation.

But de l'exposé

- Relier sources dynamiques et sources à contextes.
- Existence et unicité de mesures invariantes pour les VOMC. Propriétés de mélange.
- Fonctions génératrices de la r-ième apparition d'un motif pour une source à contexte.

Definition (arbre des contextes probabilisé)

Un sous arbre de l'arbre complet est dit saturé si tous les noeuds internes ont 2 descendants. Un arbre des contextes est un sous arbre saturé de l'arbre complet dont la frontière est finie ou dénombrable. Un arbre des contextes probabilisé est une paire

$$\left(\mathcal{T}, (p_u)_{u\in\partial\overline{\mathcal{T}}}\right)$$

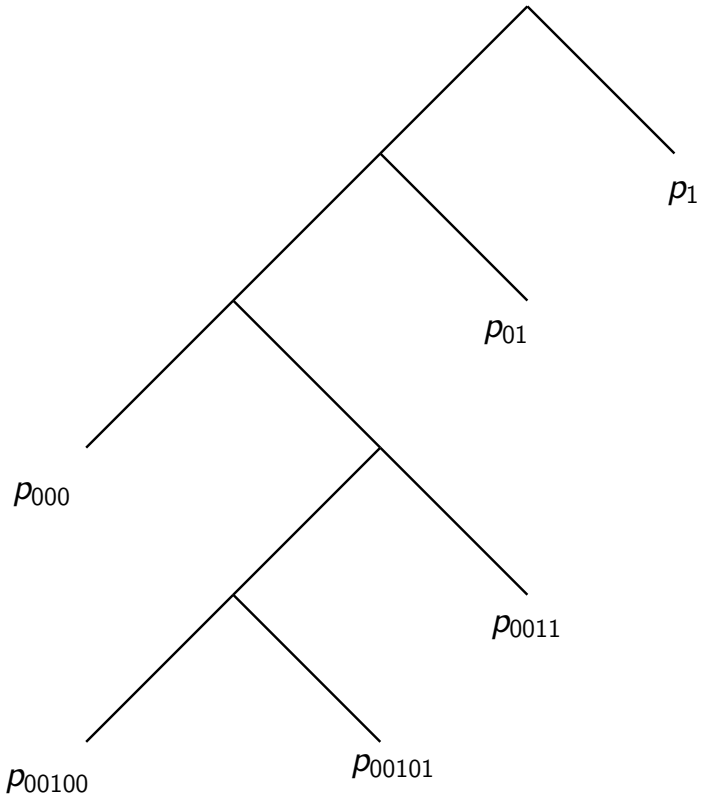
où $\mathcal{T}$ est un arbre des contextes et $(p_u)_{u\in\partial\overline{\mathcal{T}}}$ est une famille de mesures de probabilité sur $\{0,1\}$, indexée par l'ensemble $\partial\overline{\mathcal{T}}$ de toutes les feuilles de $\mathcal{T}$.

Definition (VOMC associée)

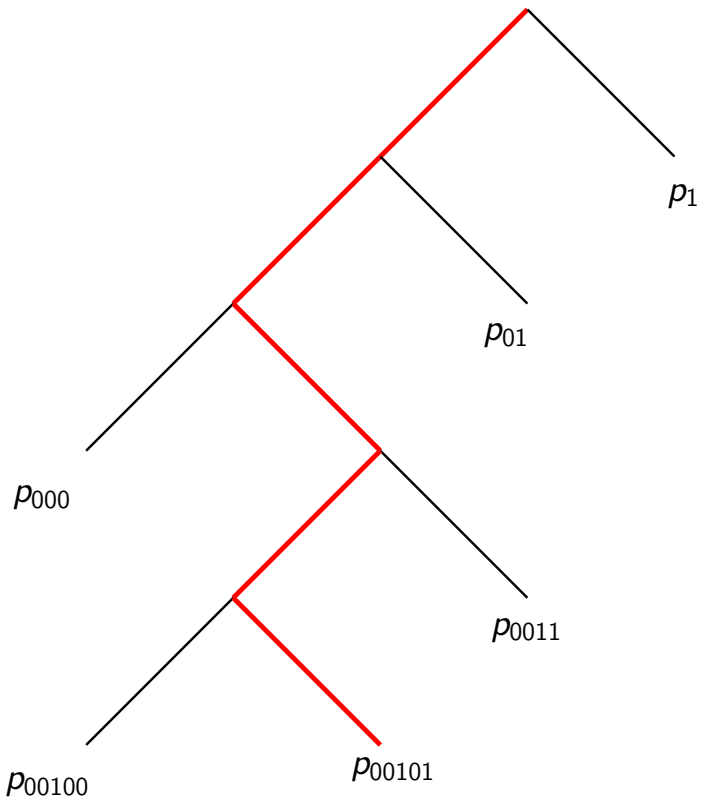
Soit $\mathcal{S} = \{0,1\}^{\mathbb{Z}_-}$ *La VOMC associée à l'arbre des contextes probabilisé* $(\mathcal{T}, (p_u)_{u\in\partial\overline{\mathcal{T}}})$ *est une chaîne de Markov* $(S_n)_{n\in\mathbb{N}}$ *d'ordre* 1 *sur* $\mathcal{S}$ *définie par les formules de transition*

$$\forall n \geqslant 0,\ \forall \alpha \in \{0,1\},\ \mathbf{P}\left(S_{n+1} = S_n\alpha | S_n\right) = p_{\mathrm{Suff}(S_n)}\left(\alpha\right).$$

suffixe du mot 11010100



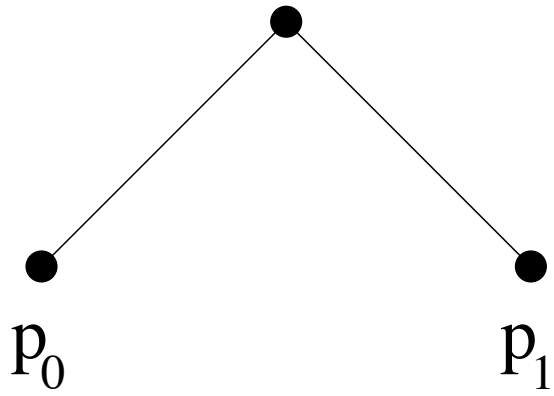
suffixe du mot 11010100



Processus de la dernière lettre

On note X_n la lettre la plus à droite de S_n.
$(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov d'ordre d si l'arbre des contextes est fini de hauteur d.
$(X_n)_{n\in\mathbb{N}}$ n'est pas Markov si l'arbre est infini.

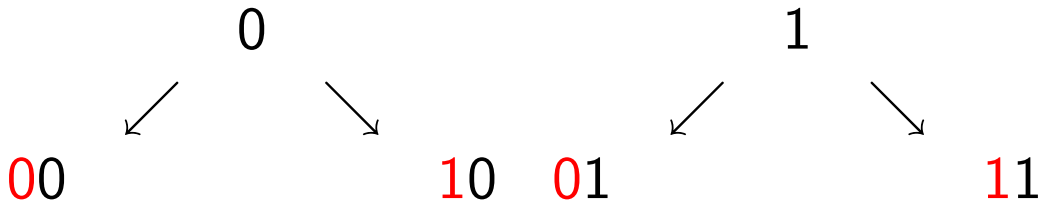
Une mesure stationnaire pour la VOMC est une mesure sur $\mathcal{S}$ qui rend le processus $(S_n)_{n\in\mathbb{N}}$ stationnaire.
Est-ce que toute VOMC admet une unique mesure stationnaire ?

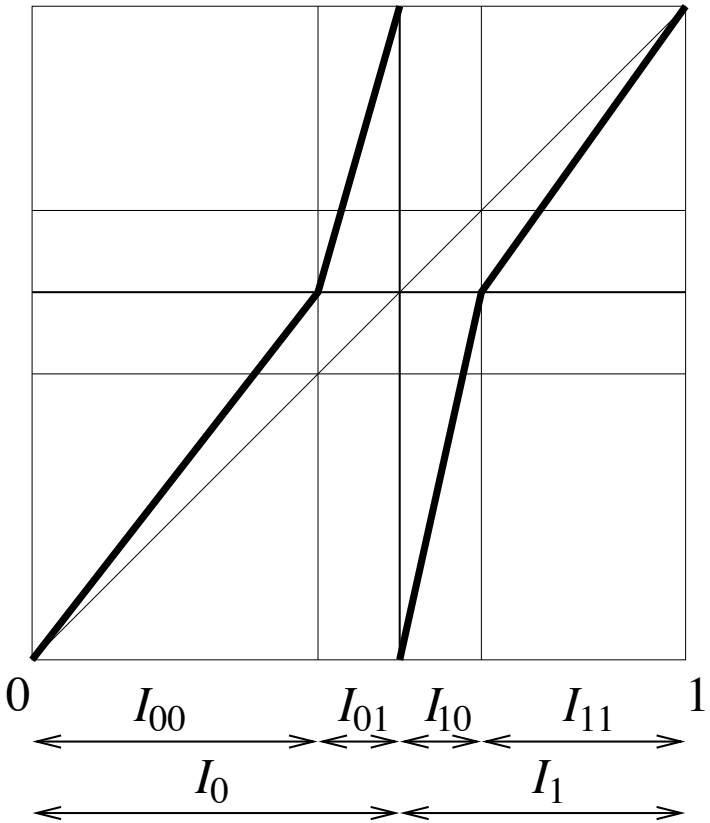


w mot fini, $\mathcal{C}_w$ ensemble des mots infinis finissant par w.
Si $w = \alpha_{-k}\alpha_{-k+1}\dots\alpha_0$ et $\mathbf{P}_{St}$ est la mesure stationnaire alors

$$\begin{aligned}\mathbf{P}_{St}(\mathcal{C}_w) &= \mathbf{P}_{St}(\mathcal{C}_{\alpha_{-k}})p_{\alpha_{-k}}(\alpha_{-k+1})p_{\alpha_{-k+1}}(\alpha_{-k+2})\cdots p_{\alpha_{-1}}(\alpha_0)\\ &= \mathbf{P}_{St}(\mathcal{C}_{\alpha_{-k}})\prod_{0\le j\le k-1} p_{\alpha_{-(j+1)}}(\alpha_{-j})\end{aligned}$$

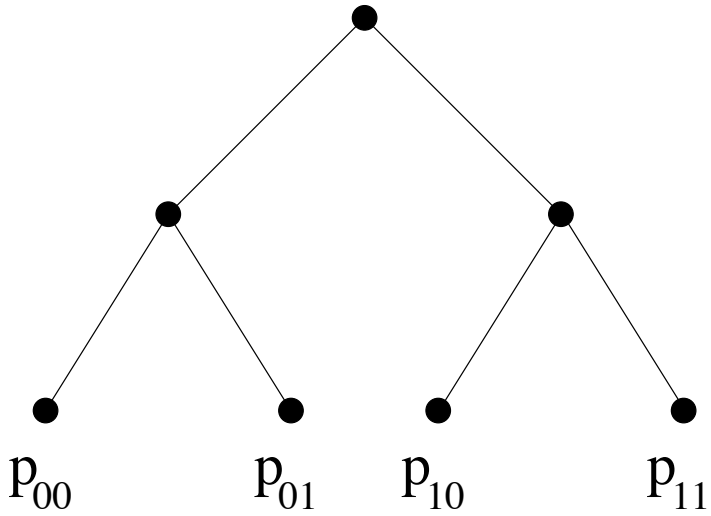
Système dynamique : deux contextes donnent quatre intervalles



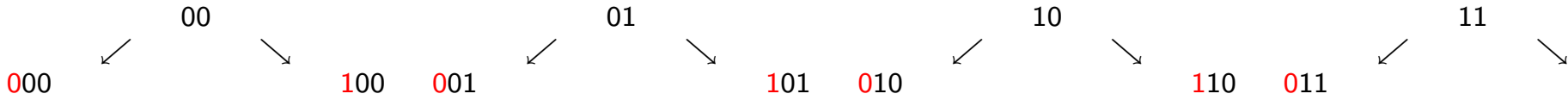


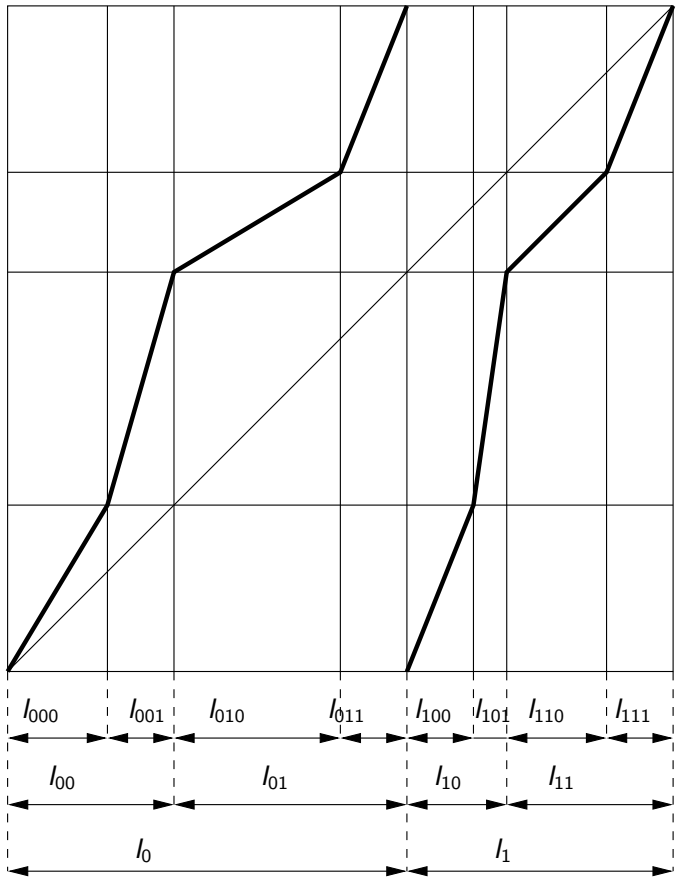
$$|I_0| = \mathbf{P}_{St}(\mathcal{C}_0)$$
$$|I_1| = \mathbf{P}_{St}(\mathcal{C}_1)$$
$$|I_{00}| = \mathbf{P}_{St}(\mathcal{C}_{00})$$
$$|I_{01}| = \mathbf{P}_{St}(\mathcal{C}_{01})$$
$$|I_{10}| = \mathbf{P}_{St}(\mathcal{C}_{10})$$
$$|I_{11}| = \mathbf{P}_{St}(\mathcal{C}_{11})$$

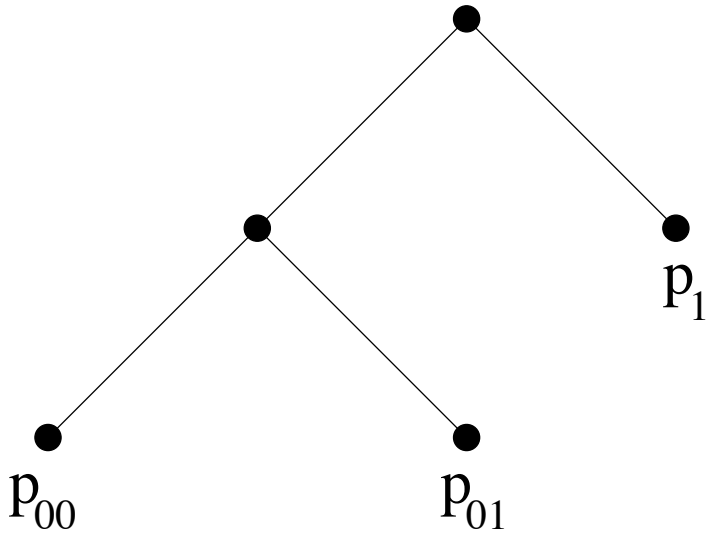
Chaîne de Markov d'ordre 2



Chaîne de Markov d'ordre 2



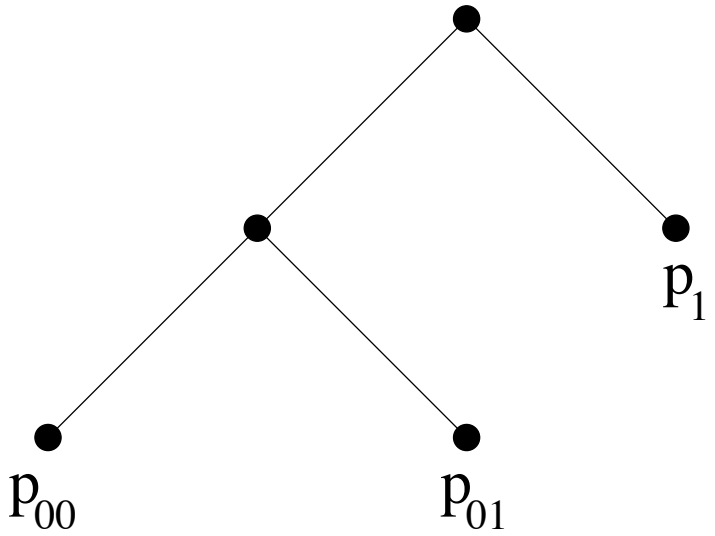




si $w = \alpha_1\alpha_2 \dots \alpha_N$, alors

$$\mathbf{P}_{St}(\mathcal{C}_w) = \prod_{k=0}^{N-1} p_{\mathrm{Suff}(\alpha_1 \dots \alpha_k)}(\alpha_{k+1})$$

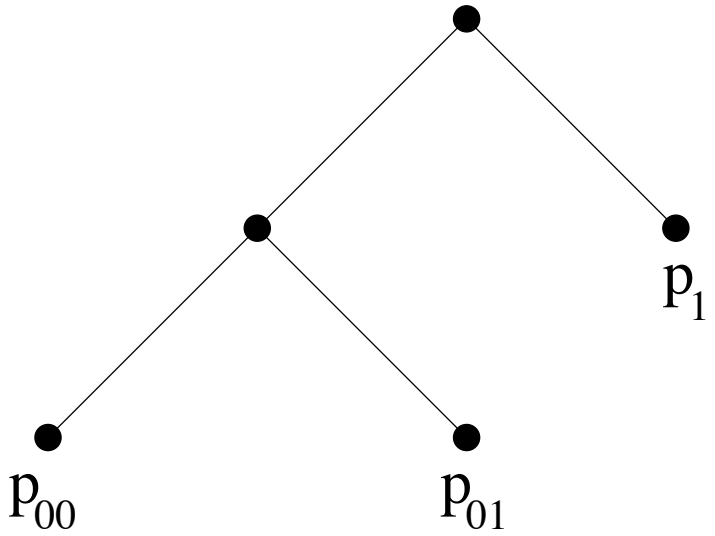
$$\mathbf{P}_{St}(110100) = p_{01}(0)\,\mathbf{P}_{St}(11010)$$



si $w = \alpha_1\alpha_2 \dots \alpha_N$, alors

$$\mathbf{P}_{St}\left(\mathcal{C}_w\right) = \prod_{k=0}^{N-1} p_{\mathrm{Suff}(\alpha_1 \dots \alpha_k)}(\alpha_{k+1})$$

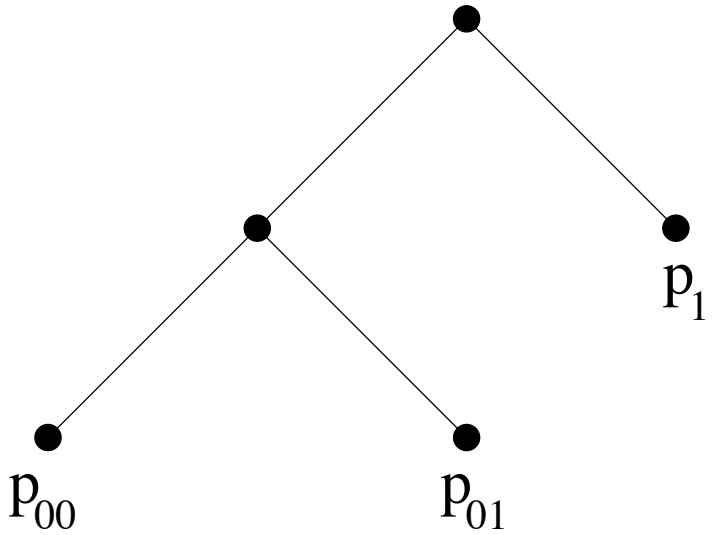
$$\mathbf{P}_{St}(110100) = p_{01}(0)p_1(0)\,\mathbf{P}_{St}(1101)$$



si $w = \alpha_1\alpha_2 \dots \alpha_N$, alors

$$\mathbf{P}_{St}(\mathcal{C}_w) = \prod_{k=0}^{N-1} p_{\mathrm{Suff}(\alpha_1 \dots \alpha_k)}(\alpha_{k+1})$$

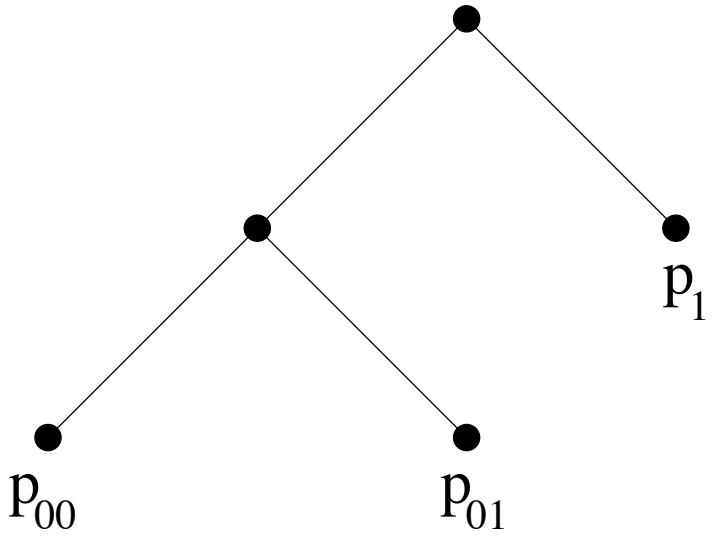
$$\mathbf{P}_{St}(110100) = p_{01}(0)p_1(0)p_{01}(1)\,\mathbf{P}_{St}(110)$$



si $w = \alpha_1\alpha_2 \dots \alpha_N$, alors

$$\mathbf{P}_{St}\left(\mathcal{C}_w\right) = \prod_{k=0}^{N-1} p_{\mathrm{Suff}(\alpha_1 \dots \alpha_k)}(\alpha_{k+1})$$

$$\mathbf{P}_{St}(110100) = p_{01}(0)p_1(0)p_{01}(1)p_1(0)\,\mathbf{P}_{St}(11)$$

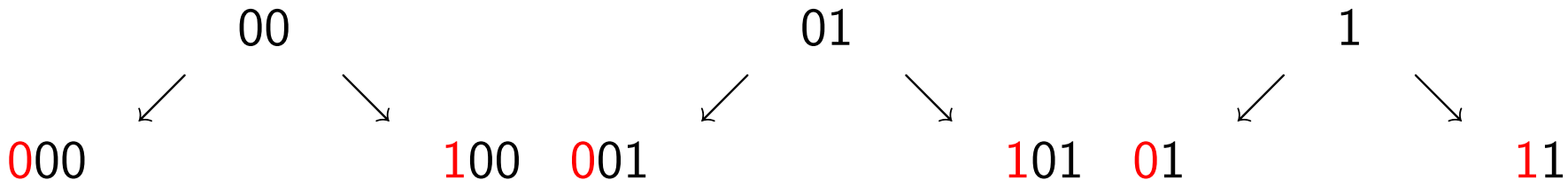


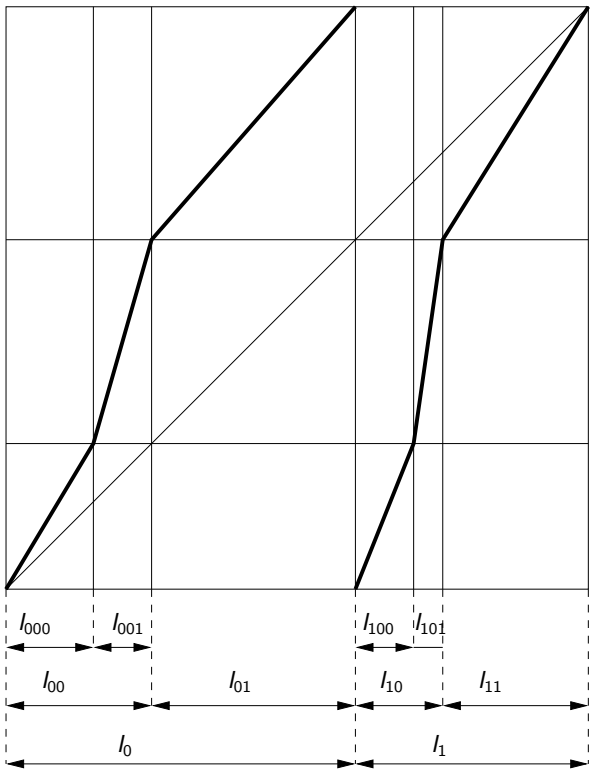
si $w = \alpha_1\alpha_2 \dots \alpha_N$, alors

$$\mathbf{P}_{St}(\mathcal{C}_w) = \prod_{k=0}^{N-1} p_{\mathrm{Suff}(\alpha_1 \dots \alpha_k)}(\alpha_{k+1})$$

$$\mathbf{P}_{St}(110100) = p_{01}(0)p_1(0)p_{01}(1)p_1(0)p_1(1)\,\mathbf{P}_{St}(1)$$

Système dynamique : trois contextes donnent six intervalles





$T : [0,1] \to [0,1]$ affine sur chaque intervalle.
$c : [0,1] \to \{0,1\}$ codage de T : $c(I_0) = 0$ et $c(I_1) = 1$.

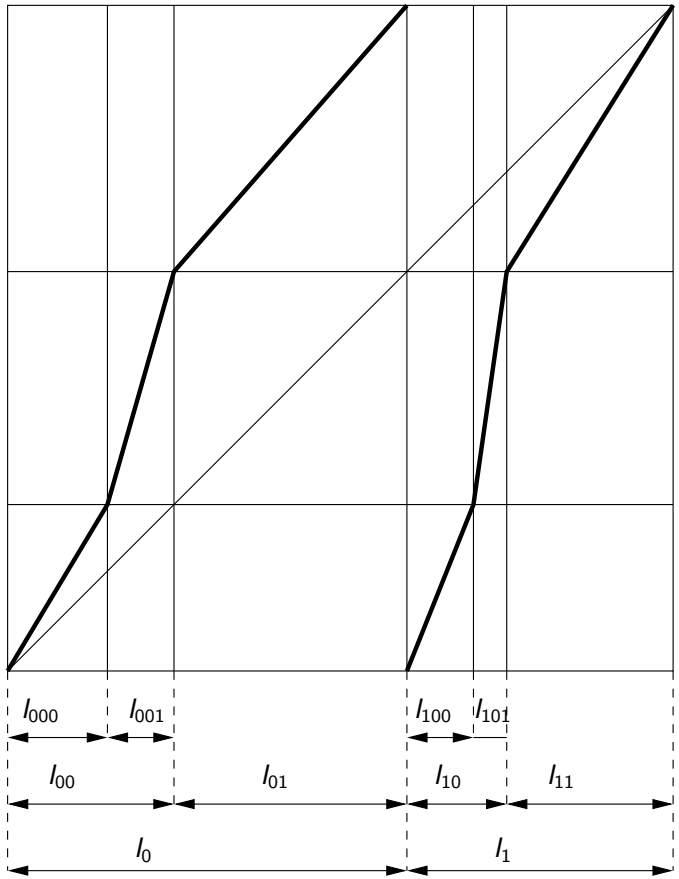
ξ de loi uniforme sur $[0,1]$.

Proposition

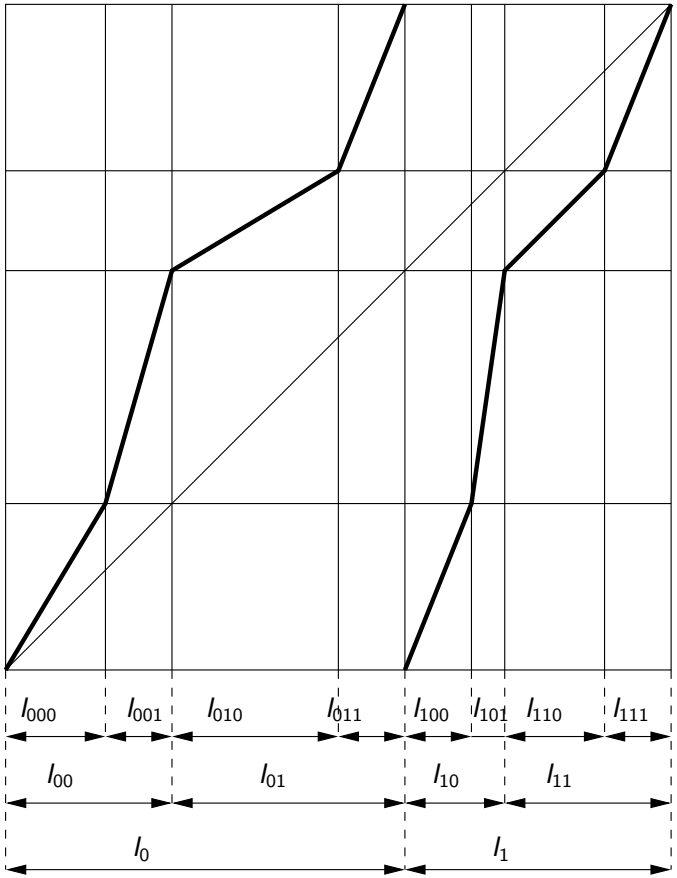
1. *La mesure de Lebesgue est T-invariante.*
2. *X_n et $c(T^n\xi)$ ont même loi.*

la source dynamique et la source issue de l'arbre des contextes probabilisé sont les mêmes.

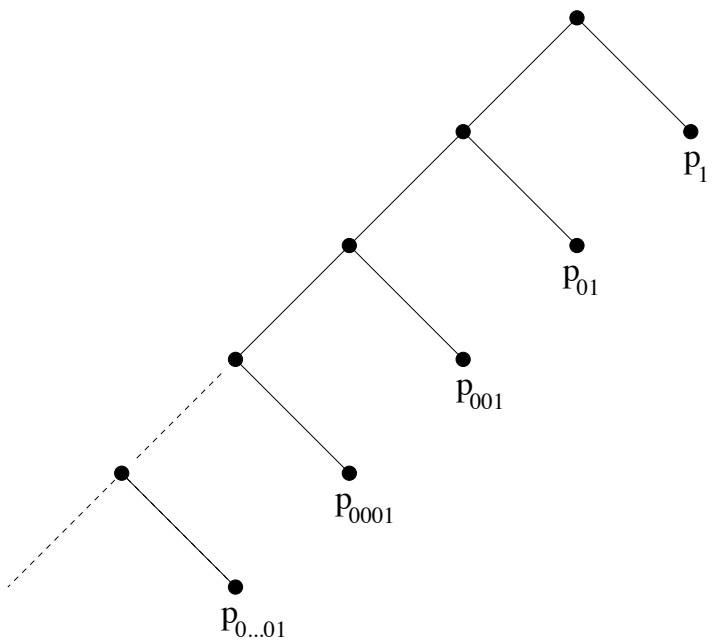
Le peigne a deux dents



Le peigne a deux dents



Le peigne infini



Proposition

Soit $(S_n)_{n\geqslant 0}$ la VOMC définie par le peigne infini probabilisé. Alors le processus de Markov $(S_n)_n$ admet une mesure de probabilité stationnaire sur $\mathcal{S}$ si et seulement si la série numérique suivante converge

$$\sum_n \left(\prod_{k=0}^{n} p_{0^k 1}(0) \right).$$

De plus, si $p_{0^\infty} \neq 0$, la mesure stationnaire est unique.

Fonction génératrice de la r-ième occurence d'un mot w

Soit $w = w_1 \dots w_k$, T_w^r la $r^{\text{ième}}$ occurence de w dans S et $\Phi_w^{(r)}$ sa fonction génératrice.

Proposition

Pour $|x| < 1$ *:*

$$\Phi_w^{(r)}(x) = \Phi_w^{(1)}(x) \left(\frac{S_w(x)}{1 + S_w(x)} \right)^{r-1},$$

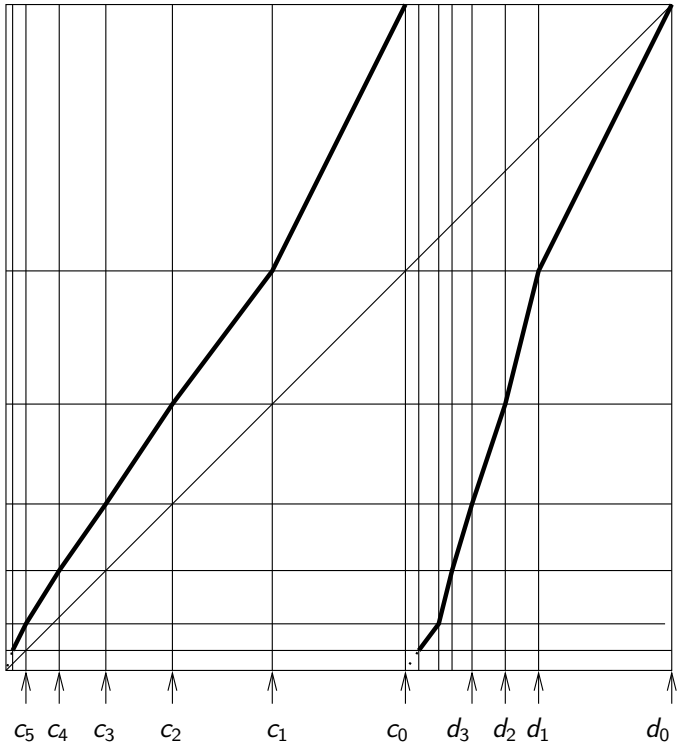
avec

Fonction génératrice de la r-ième occurence d'un mot w

$$\begin{aligned} S_w(x) &= \sum_{j=k}^{\infty} x^j \pi^{j+k-i}(10^{k-i}, w) \\ &+ \sum_{j=1}^{k-1} x^j \mathbb{1}_{\{w_{j+1}^k = w_1^{k-j}\}} x^j \pi^{j+k-i}(10^{k-i}, w_{k-j+1}^k), \\ \Phi_w^{(1)}(x) &= \frac{x^k \, \mathbf{P}_{St}(\mathcal{C}_w)}{(1-x)\,(1+S_w(x))}, \end{aligned}$$

et pour deux mots finis u et v et tout entier $n \geqslant |u| + |v| - 1$,

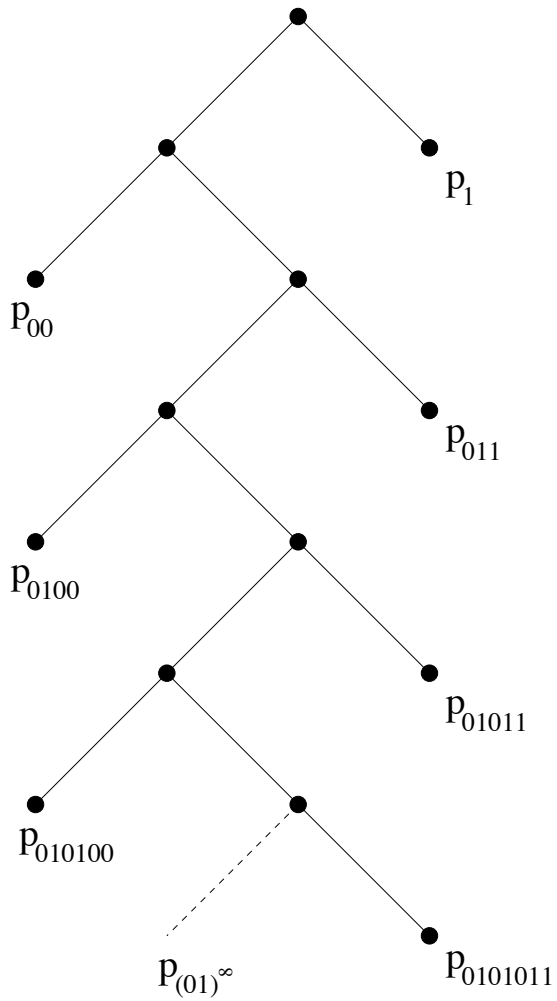
$$\pi^n(u, v) = \mathbf{P}\left(X_{n-|v|+1} \dots X_n = v | X_0 \dots X_{|u|-1} = u\right).$$



$$T'(O) = \lim_{n\to\infty} \frac{1}{p_{0^n1}(0)} \qquad T'(c_0) = \lim_{n\to\infty} \frac{1}{1 - p_{0^n1}(0)}$$

- si $(p_{0^n1}(0))_{n\in\mathbb{N}}$ converge vers $l < 1$, système uniformément dilatant, mélange exponentiel.
- si $l = 1$, 0 est point fixe indifférent, mélange polynômial.

Le bambou fleuri

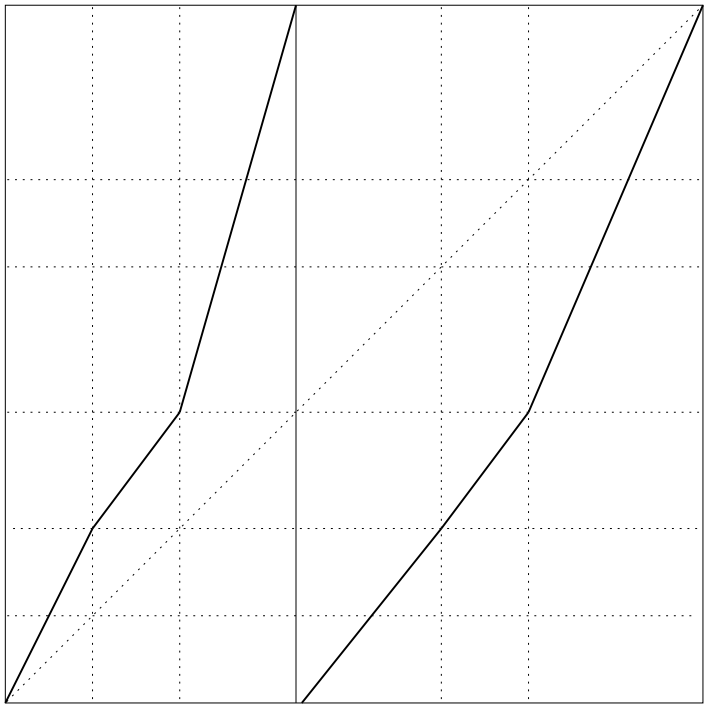


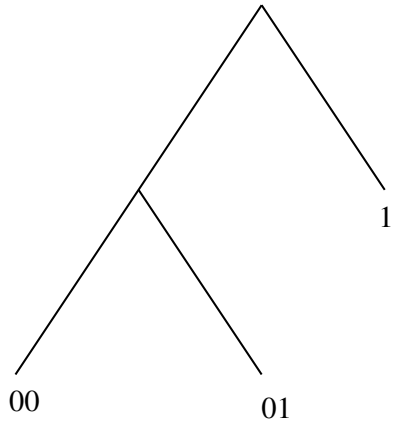
Proposition

Soit $(S_n)_{n\geqslant 0}$ la VOMC définie par le bambou fleuri probabilisé. Alors le processus de Markov $(S_n)_n$ admet une mesure de probabilité stationnaire sur $\mathcal{S}$ si et seulement si les deux séries numériques suivantes convergent.

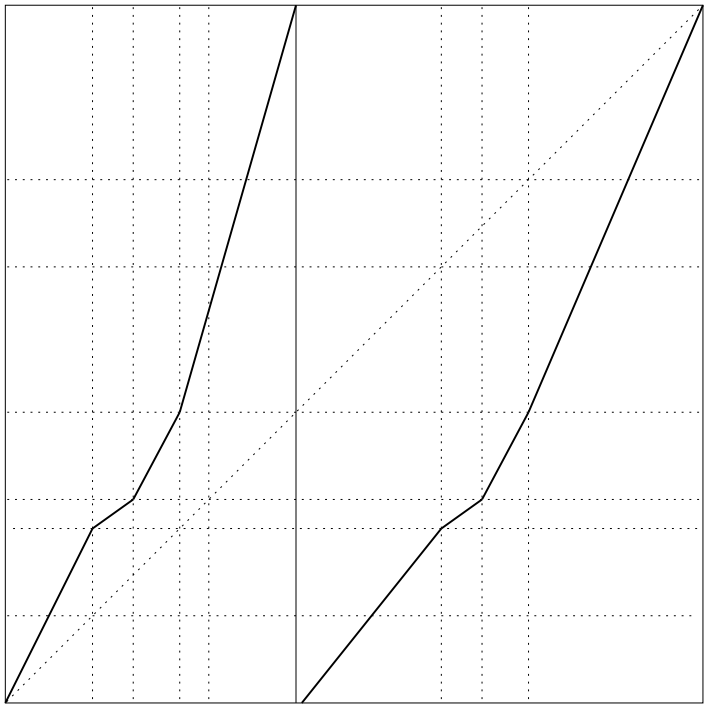
$$\begin{cases} S_1 = \sum_n \left(p_1(0)^n \prod_{k=0}^{n-1} p_{(01)^k 1}(1) \right) \\ S_{00} = \sum_n \left(p_1(0)^n \prod_{k=0}^{n-1} p_{(01)^k 0^2}(1) \right) \end{cases} \tag{3.1}$$

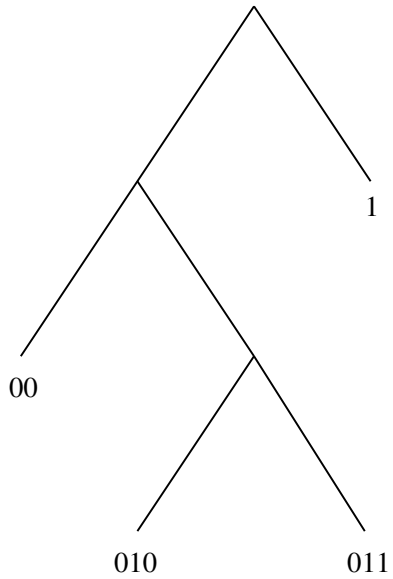
De plus, si $S_{00}(1 + p_1(0)) - S_1 \neq 0$, alors la mesure stationnaire est unique.

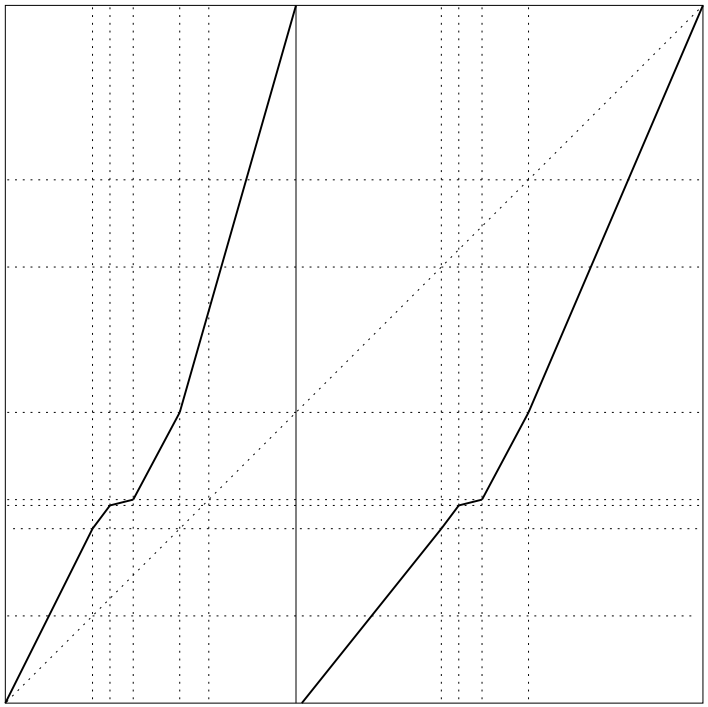


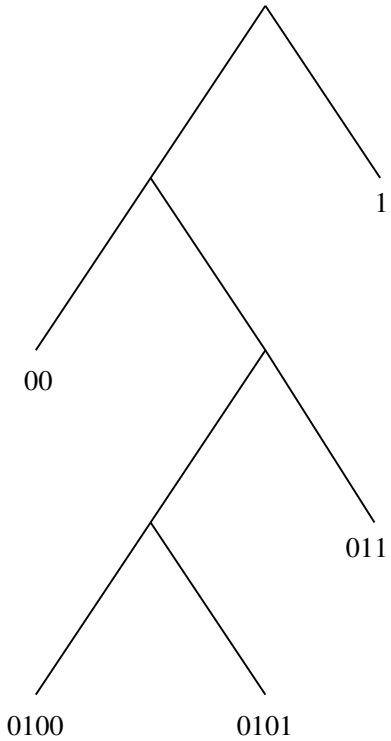

1
00
01

Le bambou fleuri








1
00
011
0100
0101

$\mathcal{T}$ arbre des contextes avec branche infinie de la forme $u = v^\infty$ ($|v| = k$)

Proposition

Il existe une mesure de probabilité stationnaire $\mathbf{P}_{St}$ *pour l'arbre si et seulement si les* k *séries suivantes convergent : pour* $i \in \{0, \dots, k-1\}$,

$$S_i = \sum_n \left(\prod_{j=1}^{k-1} p_{s(v,i,j)}(v_j) \right)^{n-1} \left(\prod_{m=1}^{n-1} p_{\mathrm{Suff}(\bar{v}_{i+1} v_i \dots v_1 \bar{v}^m)}(v_k) \right).$$

avec

$$s(v, i, j) = \mathrm{Suff}(\bar{v}_{i+1} v_i \dots v_1 \bar{v} v_k \dots v_{j+1})$$

Contextes minimaux

certains contextes jouent un rôle particulier : 1 pour le peigne infini, 1 et 00 pour le bambou.

Definition

Une feuille finie est un contexte minimal quand son code renversé $w = \alpha_1 \dots \alpha_k$ *est tel que* $\forall j \in \{1, \dots, k-1\}$, $\mathrm{Suff}(\alpha_1 \dots \alpha_j) \notin \partial\mathcal{T}$.

Questions

- Peut-on trouver des conditions d'existence et d'unicité de la mesure stationnaire pour des VOMC plus générale ?
- Comment montrer que le système dynamique associé est dilatant ?
- Quel type de mélange pour ce système ?
- Peut-on utiliser les séries de Dirichlet pour le trie des suffixes ?