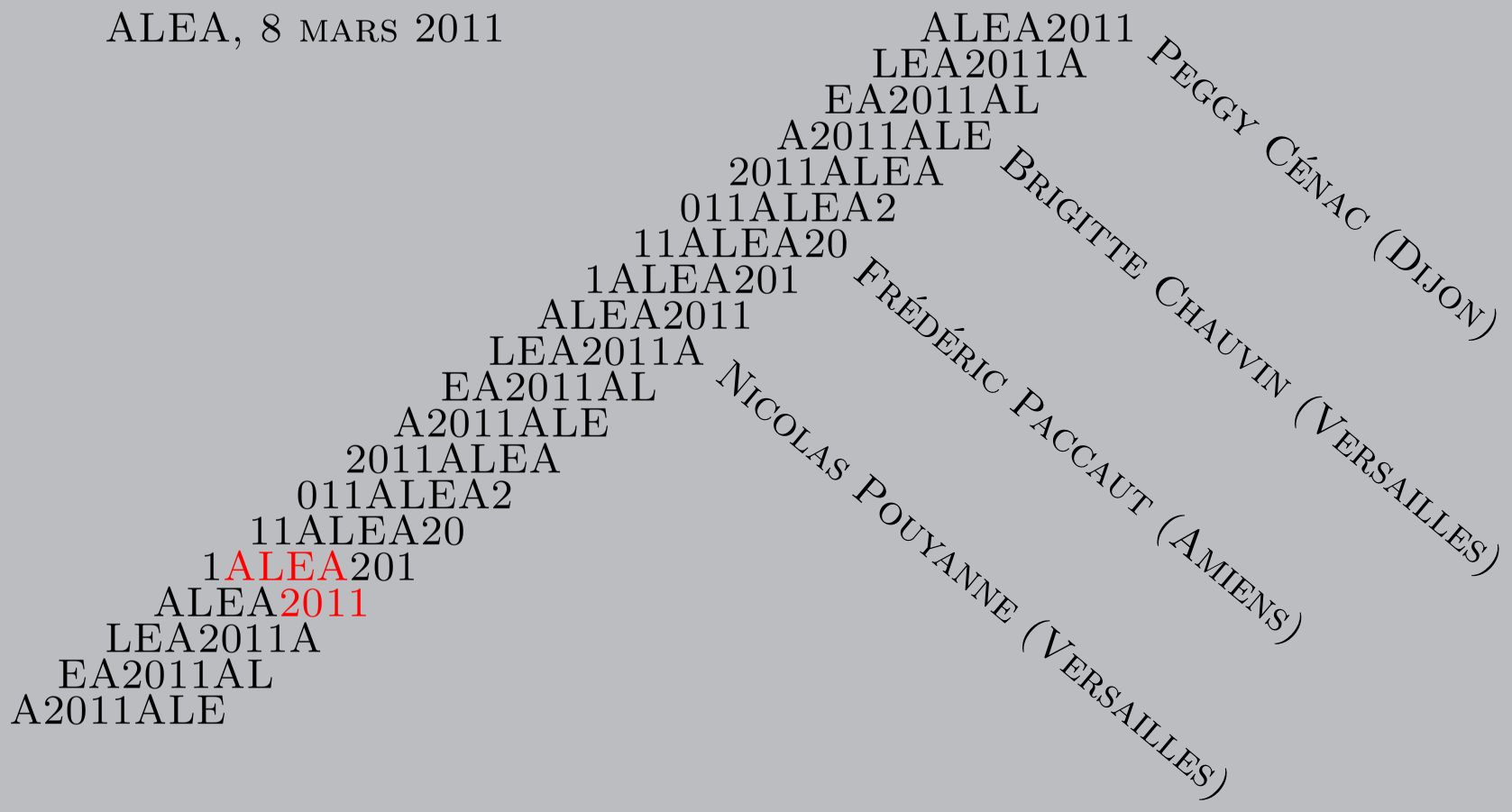
Le peigne riemannien mélange peu

Le peigne riemannien mélange peu



VLMC pour "Variable Length Markov Chains"

Alphabet fini $\mathcal{A} = \{0, 1\}$.
VLMC : suite aléatoire $(U_n)_{n\in\mathbb{N}}$ de mots infinis à gauche.

On part de $U_0 = \ldots X_{-2}X_{-1}X_0$ tiré selon une probabilité initiale.
A chaque pas de temps, on ajoute aléatoirement une lettre à droite :

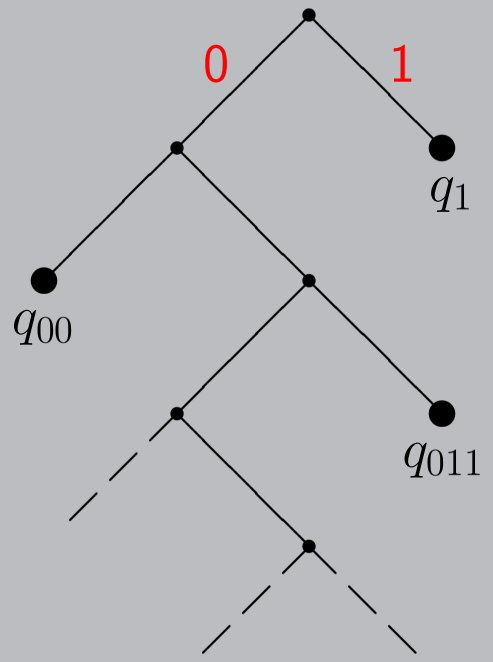
$$\forall n \geq 0,\ U_{n+1} = U_n X_{n+1} = \ldots X_0 X_1 \ldots X_n X_{n+1}.$$

Le processus est défini de façon markovienne :
$\forall \alpha \in \mathcal{A},\ \mathbf{P}(X_{n+1} = \alpha | U_n)$ est une fonction de U_n.

Pour mériter le vocable VLMC, on demande que $\mathbf{P}(X_{n+1} = \alpha | U_n)$ ne dépende que d'un suffixe de longueur finie (mais en général non bornée) de U_n.

VLMC pour "Variable Length Markov Chains" (2)

Pour définir le processus markovien, on se donne un arbre de contextes probabilisé (vocabulaire vient de Rissanen, algorithme de compression) :



Un arbre saturé (frère et sœur)
à feuilles appelées contextes,
chacune munie d'une probabilité sur $\mathcal{A}$.
[Branches infinies dénombrables.]

Fonction préfixe :
$\overleftarrow{\text{pref}}(\ldots 010011100110) = 011.$

Définition de la chaîne de Markov :

$$\mathbf{P}(U_{n+1} = U_n\alpha | U_n) = q_{\overleftarrow{\text{pref}}(U_n)}(\alpha).$$

Source probabiliste

La donnée d'un arbre des contextes probabilisé et d'une distribution initiale sur les mots infinis à gauche définit une suite aléatoire $(U_n)_{n\in\mathbb{N}}$ de mots infinis à gauche.

On a noté $U_n = \ldots X_0X_1\ldots X_{n-1}X_n$.

La suite $(X_n)_{n\in\mathbb{N}}$ des lettres finales définit une source probabiliste dont les probabilités des préfixes sont, avec les notations usuelles,

$$p_w = \mathbf{P}(X_0X_1\ldots X_{|w|-1} = w)$$

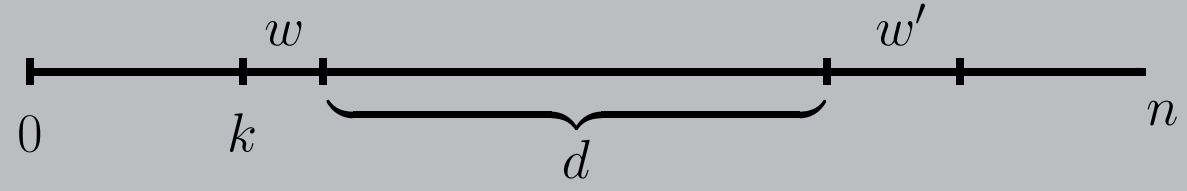
pour tout mot fini w.

Les p_w sont fonction des données de l'arbre de contextes probabilisé (forme de l'arbre et probabilités q_c) et de la distribution de U_0.

Mille questions (existence de mesures stationnaires et convergence, arbres finis et automates, processus des contextes utilisés, version dynamique, algorithmes du texte, *etc*).

Propriétés de mélange d'un processus

La question : dans la suite aléatoire $(X_n)_{n\in\mathbb{N}}$, mesurer l'indépendance de l'occurrence de deux mots finis w et w' lorsque d lettres les séparent.



Il s'agit, pour chaque $k \geq 0$, de comparer

$$\mathbf{P}(X_k X_{k+1} \cdots \in w\mathcal{A}^d w' \dots)$$

et

$$\mathbf{P}(X_k \dots X_{k+|w|-1} = w) \times \mathbf{P}(X_{k+|w|+d} \dots X_{k+|w|+d+|w'|-1} = w').$$

On considère en général un $\sup_{k\geq 0}$, qui disparaît lorsque la suite $(X_n)_n$ est stationnaire.

Propriétés de mélange d'un processus (2)

Ainsi, pour tous les mots finis w et w', on considérera la fonction

$$\Phi(d,w,w') = \frac{\left(\sum_{|v|=d} p_{wvw'}\right) - p_w p_{w'}}{p_w p_{w'}}.$$

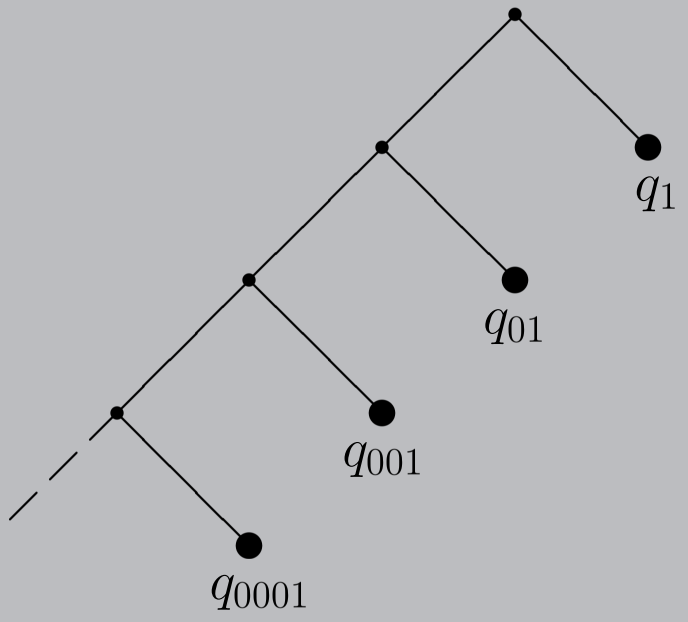
[$\Phi \equiv 0$ lorsque le processus $(X_n)_n$ est une suite de variables i.i.d.]

La propriété de mélange du processus est la vitesse de décroissance vers 0 de la fonction

$$\Phi(d) = \sup_{w,w' \in \mathcal{W}} \Phi(d,w,w').$$

Les cas favorables : Φ décroît géométriquement vers 0.

Le peigne infini riemannien stationnaire



Les probabilités aux contextes :

$$q_{0^n1}(0) = \left(\frac{n+1}{n+2}\right)^{\alpha}, \ \alpha > 1.$$

Il existe une unique mesure stationnaire π sur les mots infinis à gauche [CCPP2010].

On considère la source $(X_n)_{n\in\mathbb{N}}$ définie par
- cet arbre de contextes probabilisé ;
- la distribution initiale $U_0 \sim \pi$.

Mélange du peigne infini riemannien stationnaire

Si w et w' sont des mots finis, le calcul montre que

$$\Phi(d, w1, 1w') = \zeta(\alpha) u_{d+1} - 1$$

où

$$\sum_{n \geq 0} u_n z^n = \frac{z}{(1-z)\,\mathrm{Li}_\alpha(z)},$$

dont on déduit que

$$\boxed{\left|\Phi(d, w1, 1w')\right| \underset{d\to\infty}{\sim} \frac{1}{\alpha - 1}\left(\frac{1}{d}\right)^{\alpha-1}}.$$

[Calculs du même acabit pour les mots finis ne contenant pas de 1.]

Le mélange $\Phi(d)$ n'est pas exponentiel mais en puissance de d.
Le peigne riemannien privilégie les runs de 0.
La source est intermittente (vision dynamique).

Trie du peigne infini riemannien stationnaire

On procède au tirage aléatoire d'une suite de mots infinis indépendants produits par un peigne riemannien stationnaire.
On insère cette suite de mots dans un trie.

La hauteur du trie de n clefs est proportionnelle à $\log n$, p.s. ([Pittel]).

Calcul explicite de la série de Dirichlet qui intervient dans le calcul de l'asymptotique des paramètres principaux ([ClémFlajVall...]) :

$$\sum_{w\in\mathcal{W}} p_w^s = \frac{1}{\zeta(\alpha)^s}\left[\sum_{n\geq 1}\zeta(n,\alpha)^s + \frac{\zeta(\alpha s)^2}{\sum_{n\geq 1}\frac{1}{n^{\alpha s}}\left[1-\left(\frac{n}{n+1}\right)^{\alpha}\right]^s}\right]$$

où ζ est la fonction de Riemann et $\zeta(n,s)$ le reste de sa série.
L'asymptotique en moyenne des paramètres du trie dépend des singularités de cette série.

...à suivre.

Trie des suffixes du peigne infini riemannien stationnaire

On procède au tirage aléatoire d'un mot infini $X_0X_1X_2\ldots$ produit par un peigne riemannien stationnaire.

On insère dans un trie la suite des suffixes

$$\begin{aligned} X_0X_1X_2X_3X_4\ldots \\ X_1X_2X_3X_4\ldots \\ X_2X_3X_4\ldots \\ etc\ldots \end{aligned}$$

La hauteur du trie après insertion du $n^{\text{ième}}$ suffixe n'est pas en $\log n$ mais en $n^{1/\alpha}$.

Simulations par M. Guesdon (INRIA Rocquencourt).

Trie des suffixes du peigne infini stationnaire : simulations

Un cas de mélange exponentiel [CCPP] : le trie des suffixes est équilibré, sa hauteur en $\log n$.

Dessin1

Le cas du peigne riemannien : la source est intermittente et privilégie les runs de 0. La branche gauche du trie des suffixes croît en $n^{1/\alpha}$.

Dessin2

Trie des suffixes : méthode d'analyse presque sûre

Méthodes probabilistes utilisant la fonction génératrice de la deuxième apparition d'un motif donné. Les calculs de ces fonctions s'apparentent à ceux du mélange.

Si s est un mot infini, *i.e.* une branche infinie de l'arbre complet, on note

• $X_n(s)$ la longueur du plus grand préfixe de s inséré dans le trie des suffixes au temps n.

• $T_k(s)$ le premier moment où le préfixe de longueur k de s est inséré dans le trie des suffixes.

On joue sur la dualité suivante : $\mathbf{P}(X_n(s) \geq k) = \mathbf{P}(T_k(s) \leq n)$.

Pour l'instant, on sait démontrer que, presque sûrement, la hauteur est minorée par $\zeta(\alpha) n^{\frac{1}{\alpha}}$...

...à suivre.