CAp 2003

Identification d’automates probabilistes résiduels à la limite avec probabilité un.

François Denis, Yann Esposito.

Université de Provence,

Laboratoire d’Informatiqe Fondamentale (L.I.F.)

UMR CNRS 6166 Marseille

Automates Probabilistes (PFA)

Les *Automates Probabilistes (PFA)* sont des modèles qui possèdent exactement les mêmes propriétés que les Modèles de Markov Cachés (HMM) largements utilisés dans tous les domaines de reconnaissance de formes comme la bioinformatique, la reconnaissance vocale. . .

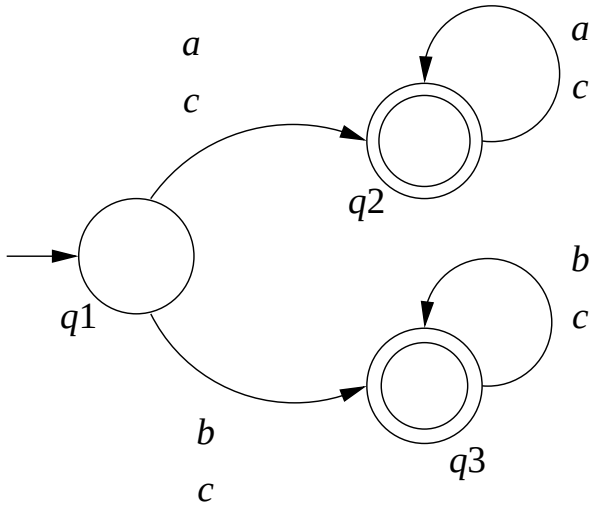
Ils modélisent des Langages stochastiques, c'est-à-dire des distributions de probabilités sur les mots :

$$P : \Sigma^* \to [0,1], \quad \sum_{w \in \Sigma^*} P(w) = 1$$

Par exemple, on peut considérer que l'alphabet Σ est l'ensemble des nucléotides $\{a, c, g, t\}$, un langage stochastique sur Σ sera une distribution de probabilités sur les séquences.

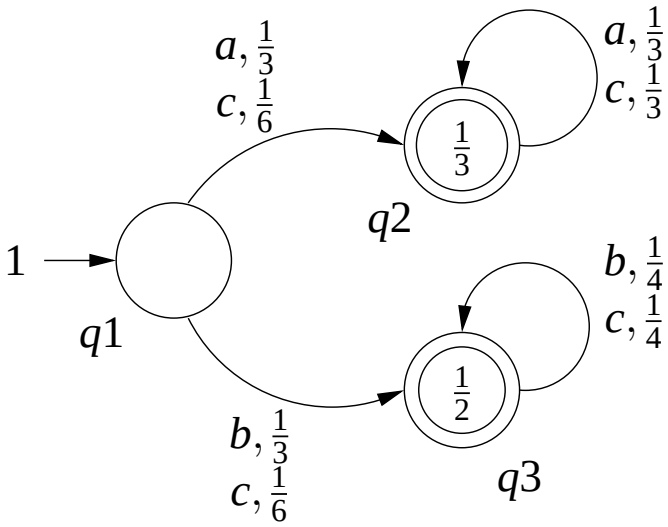
Automates Probabilistes (PFA)

Un Automate Fini Probabiliste (PFA) possède **une structure de NFA** (Un ensemble d'états, des états initiaux, terminaux et des transitions)



Automates Probabilistes (PFA)

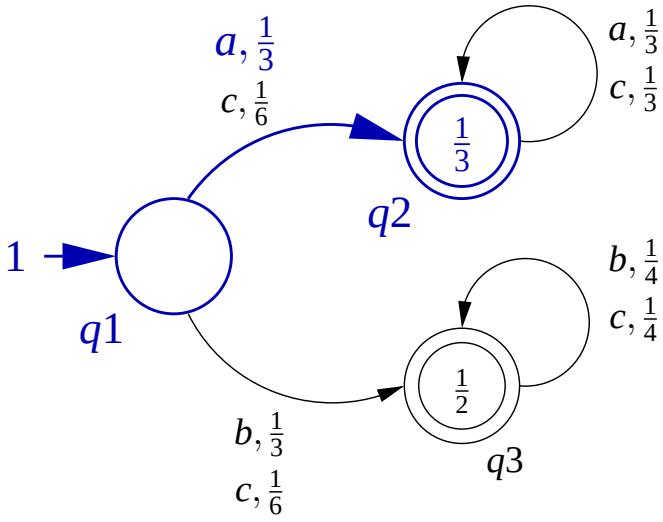
Un Automate Fini Probabiliste (PFA) possède une structure de NFA (Un ensemble d'états, des états initiaux, terminaux et des transitions) **avec des paramètres réels + des contraintes techniques qui imposent de renvoyer un langage stochastique.**



- Probabilité que $q1$ soit initial : 1.
- Probabilité que $q2$ soit terminal $1/3$.
- Probabilité d'emprunter la transition $(q1, a, q2)$: $1/3$.

Automates Probabilistes (PFA)

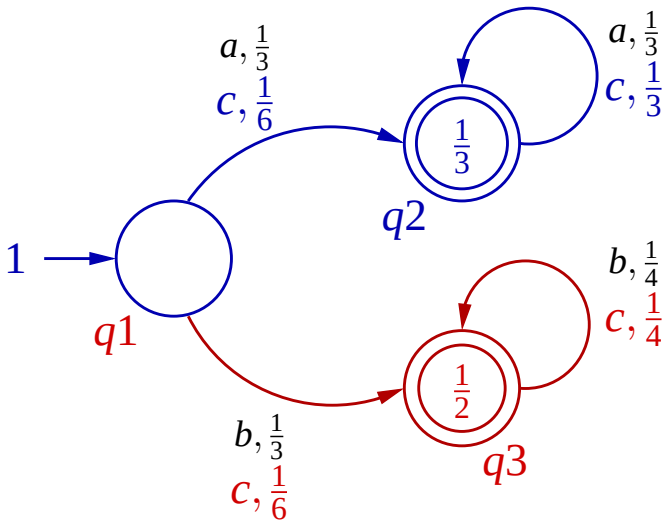
Un Automate Fini Probabiliste (PFA) possède une structure de NFA (Un ensemble d'états, des états initiaux, terminaux et des transitions) avec des paramètres réels + des contraintes techniques qui imposent de renvoyer un langage stochastique.



$$P(a) = 1 \times \frac{1}{3} \times \frac{1}{3}$$

$$P(cc) = 1 \times \frac{1}{6} \cdot \frac{1}{3} \cdot \frac{1}{3} + 1 \times \frac{1}{6} \cdot \frac{1}{4} \cdot \frac{1}{2}$$

Automates Probabilistes (PFA)



$$P(a) = 1 \times \frac{1}{3} \times \frac{1}{3}$$

$$P(cc) = 1 \times \frac{1}{6} \cdot \frac{1}{3} \cdot \frac{1}{3} + 1 \times \frac{1}{6} \cdot \frac{1}{4} \cdot \frac{1}{2}$$

Définition : Un *Automate Probabiliste Déterministe (PDFA)* est un PFA dont la structure est déterministe (DFA).

Inférence

Étant donné un échantillon fini généré par une cible inconnue, trouver un PFA A tel que P_A soit le plus proche possible de la cible.

Données + Structure $\longrightarrow$ Paramètres :	il existe des méthodes reconnues efficaces en pratique : Baulm Welch, Viterbi.
Données $\longrightarrow$ Structure + Paramètres :	Sujet ouvert de recherche.

Théorème : Quand la cible est un PDFA, alors elle peut être identifiée si l'échantillon est assez grand (identification à la limite avec probabilité 1). [Carrasco Oncina 94, de la Higuera Thollard 00].

Identification à la limite avec probabilité 1

Soit A un PDFA à paramètres rationnels qui génère P_A. Soit un échantillon infini de mots généré avec la distribution de probabilité P_A : $w_1, \ldots, w_n, \ldots$

Il existe un algorithme polynomial qui prenant en entré les n premiers mots de l'échantillon renvoie un PDFA hypothèse H_n : $\mathcal{A}(w_1, \ldots, w_n) \longrightarrow H_n$.

Alors, avec une probabilité de 1, il existe un rang après lequel l'algorithme renvoie toujours la même hypothèse H_N qui est équivalente à la cible.

$$P\left(\exists N, \forall n > N, H_n = H_N \text{ et } P_{H_N} = P_A\right) = 1$$

Algorithmes : `Alergia`, `rlips` (Carrasco Oncina 94, 96), `MDI` (Thollard, Dupont 00).

Langages résiduels stochastiques

Malheureusement les PDFAs sont beaucoups moins expressifs que les PFAs.

D'où l'intéret d'introduire une classe plus expressive : les Automates Résiduels Probabilistes (PRFA).

Langage résiduel de $L \subset \Sigma^*$ par rapport au mot u : $u^{-1}L = \{v \mid uv \in L\}$

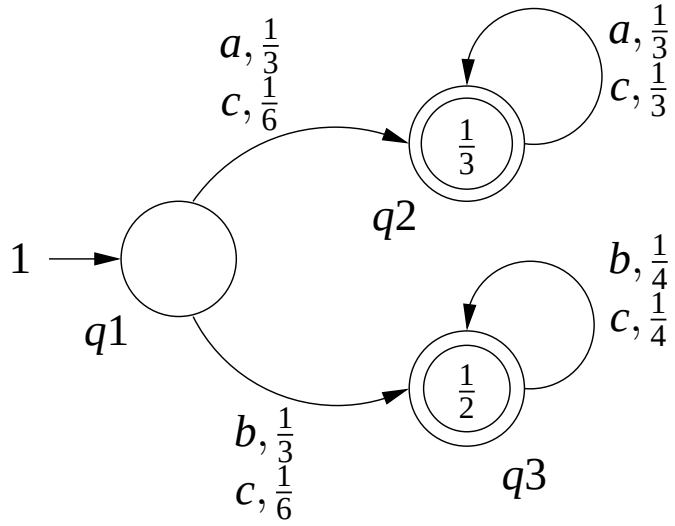
Langage résiduel stochastique de P par rapport au mot u est un langage stochastique défini par :

$$u^{-1}P(w) = P(w|u) = \frac{P(uw)}{P(u\Sigma^*)}$$

Automates Résiduels Probabilistes (PRFA)

Le langage stochastique généré par un état est le langage généré en commançant par cet état.

Définition : Un PRFA A est un PFA dont chaque état génère un langage résiduel de P_A.



$$P_{q3}(b) = \frac{1}{4} \times \frac{1}{2}$$

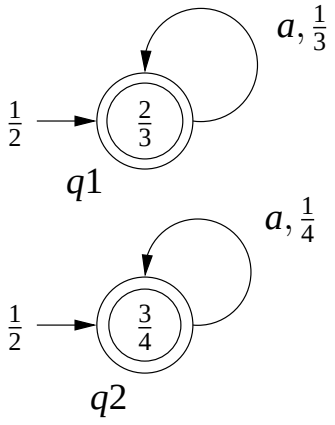
$$P_{q1} = \varepsilon^{-1}P,\ P_{q2} = a^{-1}P,\ P_{q3} = b^{-1}P$$

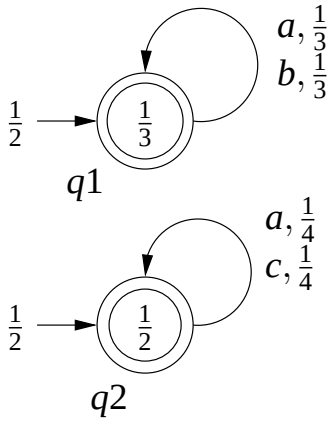
Quelques propriétés des PRFAs

Les PDFAs sont des PRFAs.

Tout PRFA A possède une forme normale unique $\mathrm{Fnorm}(A)$.

Les PRFAs sont plus expressifs que les PDFAs mais moins que les PFAs.





Exemple d'algorithme d'inférence de PDFA

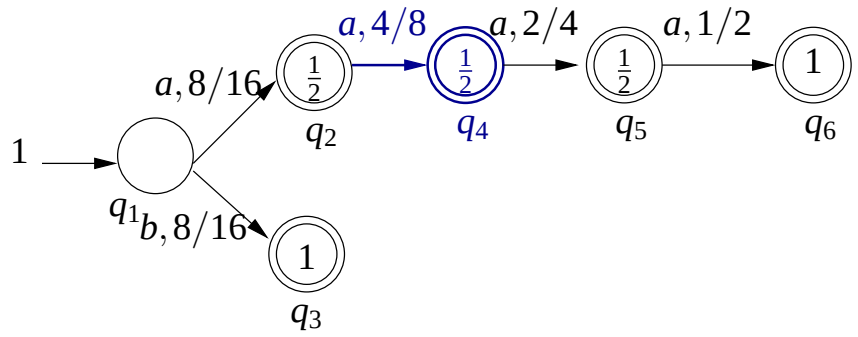
Entrée : Un échantillon de mots $\{a, c, aa, caa, bcbcc, a, a, a, cbb, \ldots, bc\}$
Sortie : Un PDFA

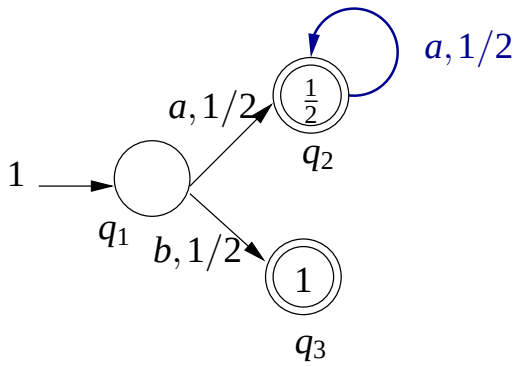
1. Construire un arbre préfixe à partir de l'échantillon.
2. Pour tous les couples d'états (qi, qj), si il semblent générer le même langage stochastique alors les fusionner.

Exemple d'inférence de PDFA

Echantillon avec le nombre de répétition entre parenthèses :

$$\{a(4), aa(2), aaa(1), aaaa(1), b(8)\}$$
$$aa^{-1}P = a^{-1}P \quad ?$$
$$P_{q_4} \simeq P_{q_2} \quad \text{? test statistique } \{\varepsilon(2), a(1), aa(1)\} \simeq \{\varepsilon(4), a(2), aa(1), aaa(1)\}$$





Remarque : tous les états correspondent à un résiduel du langage.

Extension de la notion de fusion d'états

PDFA : $wa^{-1}P = u^{-1}P$:

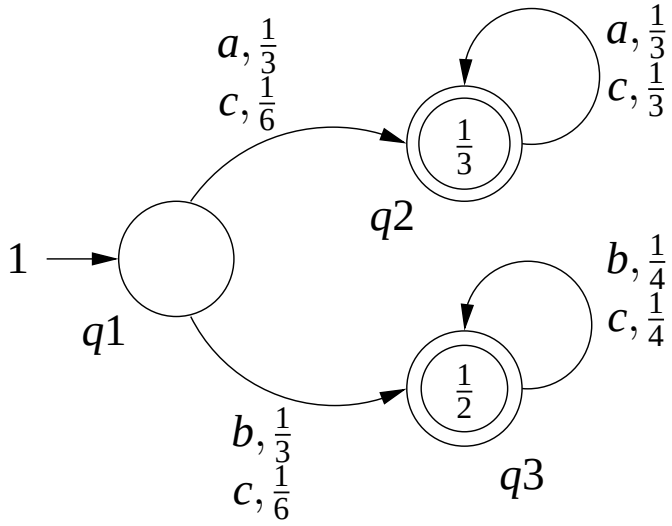
$$w^{-1}P \xrightarrow{a} u^{-1}P \qquad \left(w^{-1}P(a\Sigma^*)\right)$$

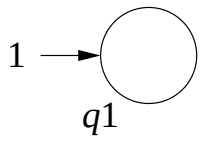
PRFA : $wa^{-1}P = \sum_{i=1}^{n} x_i \cdot w_i^{-1}P$: ($\forall i \in \{1,\ldots,n\}, x_i \geq 0$, $\sum_{i=1}^{n} x_i = 1$)

$$\begin{array}{ccccc} w^{-1}P & \xrightarrow{a} & w_1^{-1}P & & \left(x_1 \cdot w^{-1}P(a\Sigma^*)\right) \\ \vdots & \vdots & \vdots & & \\ w^{-1}P & \xrightarrow{a} & w_n^{-1}P & & \left(x_n \cdot w^{-1}P(a\Sigma^*)\right) \end{array}$$

L'algorithme d'inférence de PRFA : DEES

Entrée : $\{a, c, aa, caa, bcbcc, a, a, a, cbb, \ldots, bc\}$

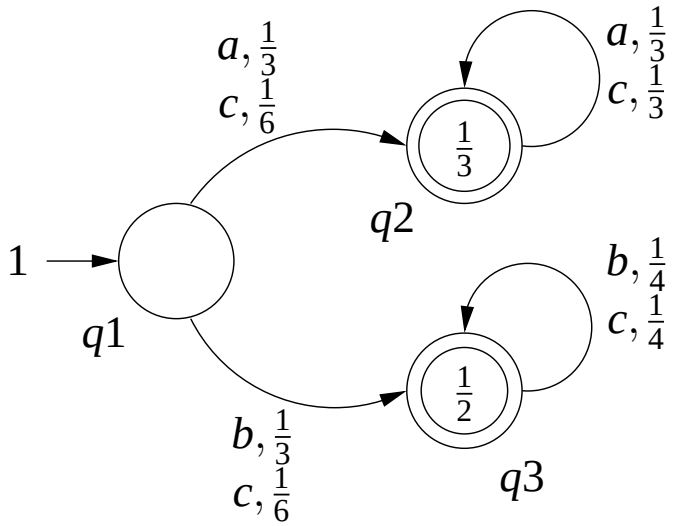


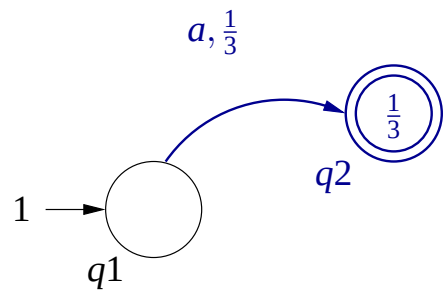


DEES : Etape 1

$$a^{-1}P = \varepsilon^{-1}P \quad ?$$

$$P_{q2} \simeq P_{q1} \quad ? \text{ non.}$$

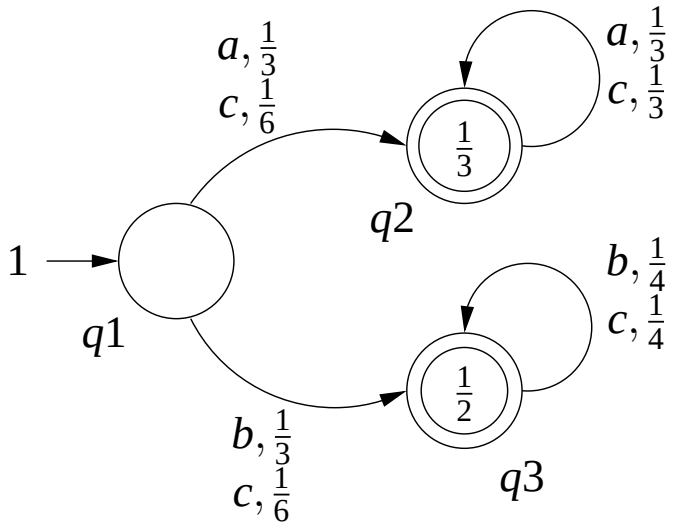


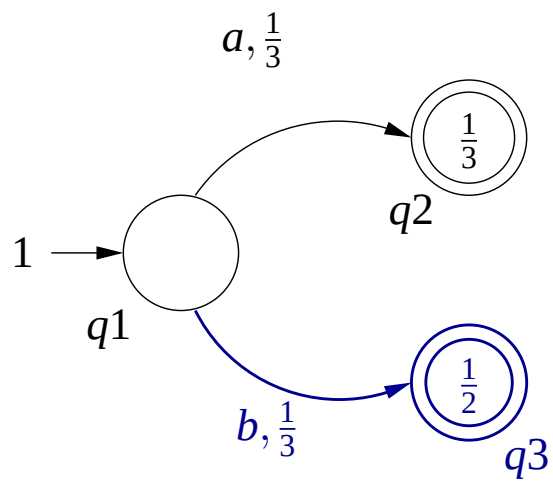


DEES : Etape 2

$$b^{-1}P = x_1 \cdot \varepsilon^{-1}P + x_2 \cdot a^{-1}P \quad ?$$

$$P_{q3} \simeq x_1 \cdot P_{q1} + x_2 \cdot P_{q2} \quad ? \text{ non.}$$

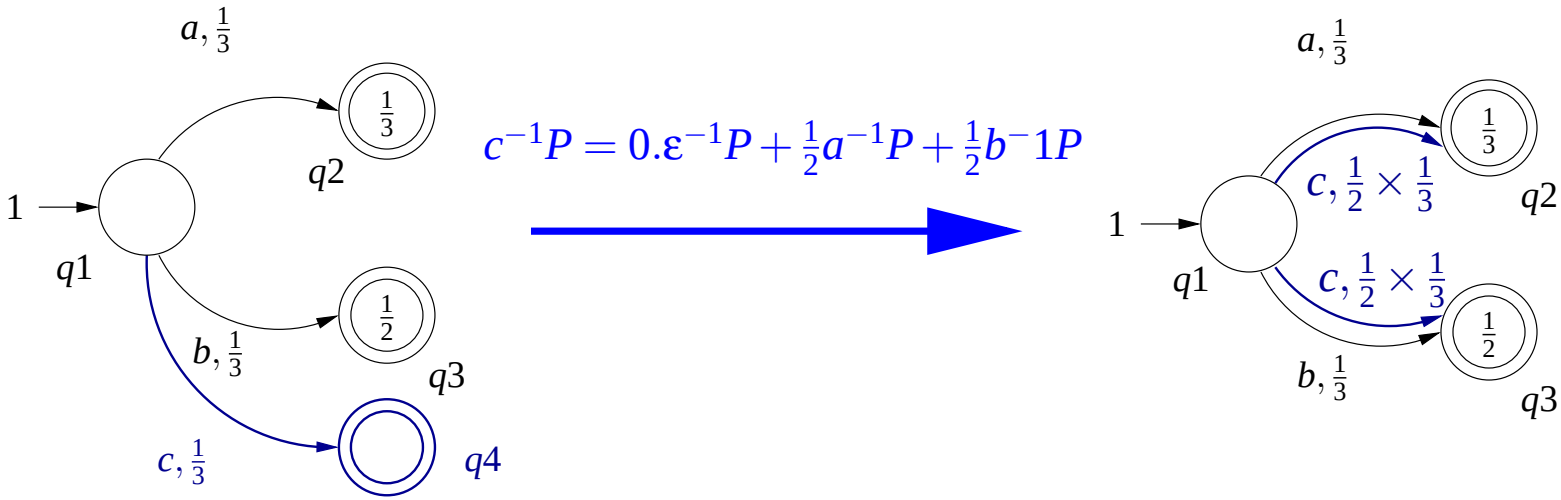




DEES : Étape 3

$$c^{-1}P = x_1 \cdot \varepsilon^{-1}P + x_2 \cdot a^{-1}P + x_3 \cdot b^{-1}P \quad ?$$

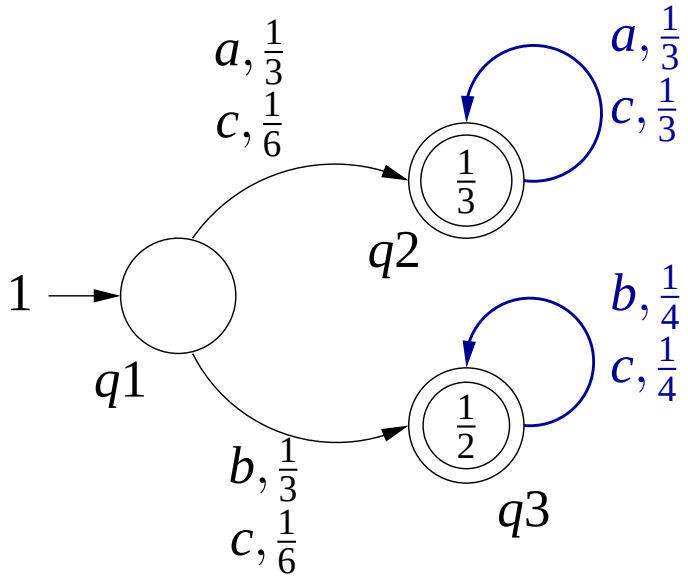
$$P_{q_4} \simeq x_1 \cdot P_{q1} + x_2 \cdot P_{q2} + x_3 \cdot P_{q3} \quad \text{? oui.}$$



DEES : Étape 4

$$aa^{-1}P \simeq P_{q5} \simeq P_{q2}, \qquad ac^{-1}P \simeq P_{q7} \simeq P_{q2},$$

$$bb^{-1}P \simeq P_{q12} \simeq P_{q3}, \qquad bc^{-1}P \simeq P_{q13} \simeq P_{q3}.$$



Identification de la structure

Théorème : Si la cible est un PRFA A, alors avec une probabilité 1, après un certain rang, l'algorithme $\mathtt{DEES}$ va renvoyer un PRFA H dont la structure est identique à celle de $\mathrm{Fnorm}(A)$ et dont les paramètres convergent vers les paramètres de $\mathrm{Fnorm}(A)$.

Identification des valeurs exactes des paramètres

Théorème : Si le langage cible P_A peut être généré par un PRFA dont tous les paramètres de sa forme normale sont des rationnels, alors il est identifiable à la limite avec probabilité 1.

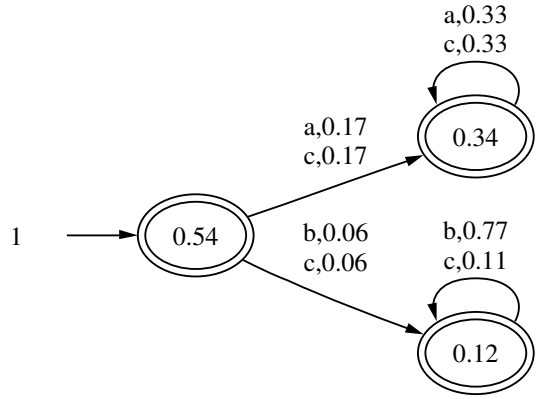
La valeur des paramètres convergent vers la valeur des paramètres de la cible.

Si $\alpha_n \to \alpha \in \mathbb{Q}$ suffisament rapidement (vitesse de convergence connue), alors, on peut construire à partir de α_n une suite β_n telle que :
$\beta_n \to \alpha$ et telle qu'elle soit constante à partir d'un certain rang.

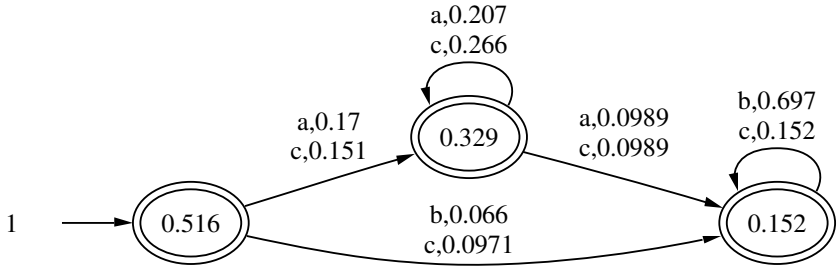
Conclusion

- Extension du résultat des PDFAs aux PRFAs.
- Le gain d’expressivité nous fait espérer avoir de meilleurs résultats pratiques.
- Expérimentations

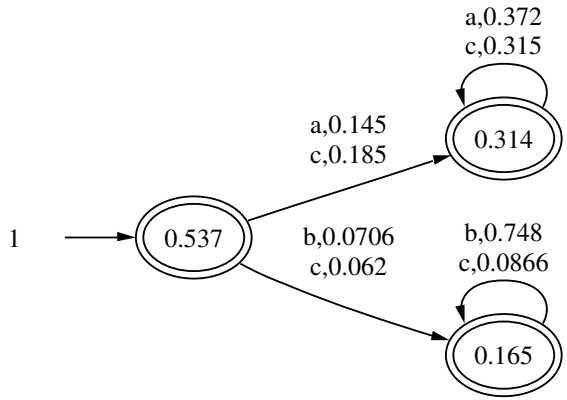
Cible



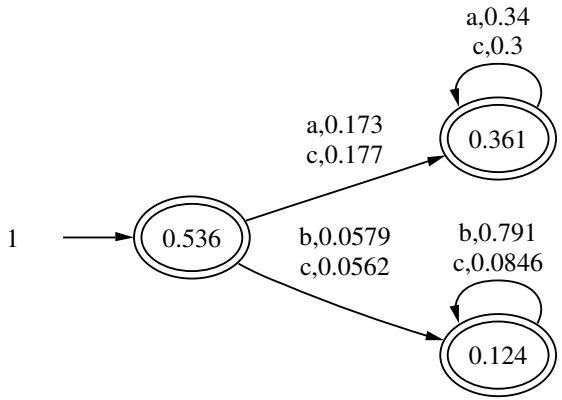
DEES (500)



DEES (2000)



DEES (10000)



Taille echantillon 500 :

	perplexité	divergence
Alergia	3.8025	0.866047
DEES	3.10735	0.17089

Taille echantillon 2000 :

	perplexité	divergence
Alergia	3.77281	0.700383
DEES	3.20014	0.127714

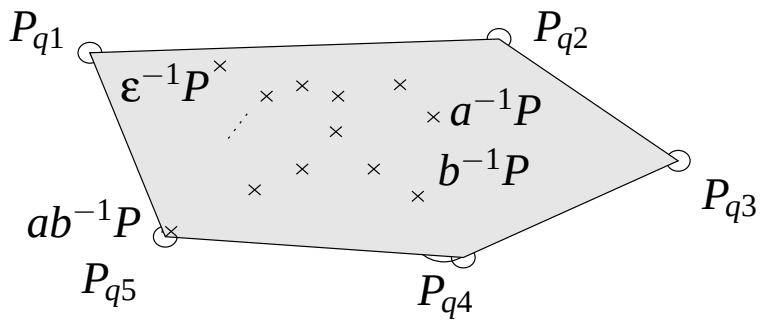
Taille echantillon 10000 :

	perplexité	divergence
Alergia	3.89598	0.887994
DEES	3.06058	0.0525914

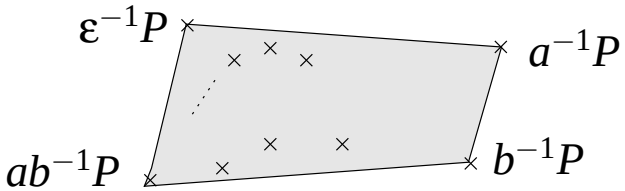
Différence d'expressivité : $\mathcal{L}_{PDFA} \subsetneq \mathcal{L}_{PRFA} \subsetneq \mathcal{L}_{PFA}$

$$\operatorname{conv}\{P_1, \ldots, P_n\} = \left\{\sum_{i=1}^{n} \alpha_i P_i \mid \alpha_i > 0 \text{ et } \sum_{i=1}^{n} \alpha_i = 1\right\}$$

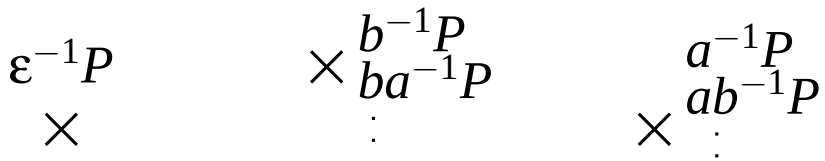
PFA : Il existe un nombre fini de langages stochastiques $(P_{q1}, \ldots, P_{qn})$ dont la partie convexe contient tous les résiduels de P et de $P_{q1}, \ldots, P_{qn}$.



PRFA : Tous les résiduels appartiennent à la partie convexe d'un nombre fini de résiduels.



PDFA : Nombre fini de résiduels.



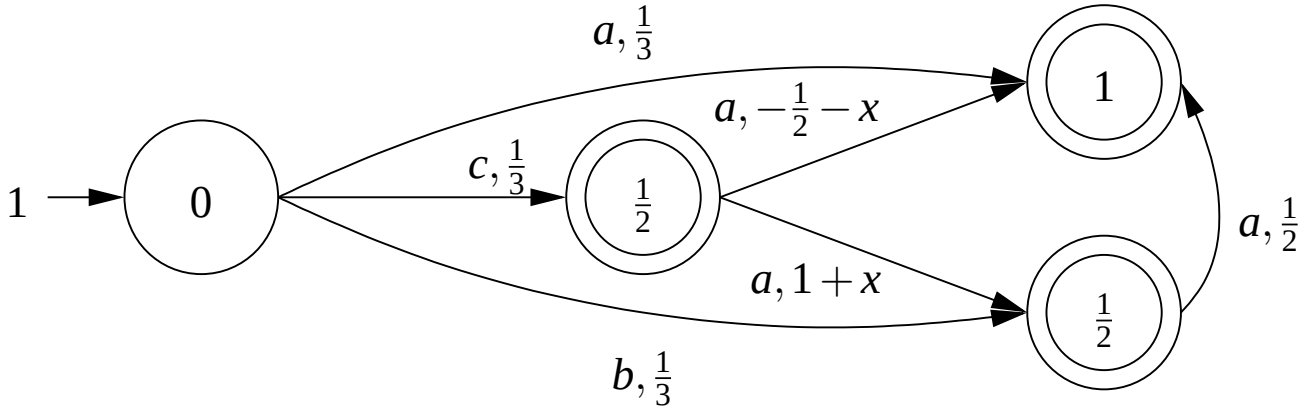
Apprentissage exact

MA (Automates à multiplicité) : PFA avec paramètres non bornés.

Apprenables à partir de **requetes** [Bergadano, Varrichio 94] :

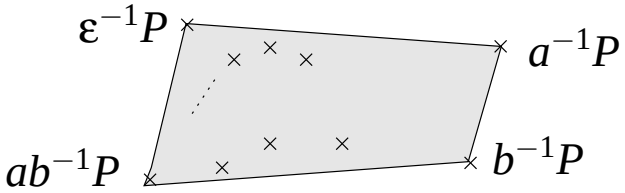
1. un mot u $\longrightarrow P(u)$.
2. un MA générant $\widehat{P} \longrightarrow \widehat{P} = P$ ou un mot w tel que $P(w) \neq \widehat{P}(w)$.

Extension directe au cas approché impossible



langage stochastiques ssi $x = 0$.

Il est indécidable de savoir si un MA génère un langage stochastique.

PRFA ?

Un PRFA : une fois réduit, pour chaque état il existe un mot qui une fois lu n'atteint que cet état.