Arbres, Booléens et urnes de Pólya

Brigitte CHAUVIN
http://chauvin.perso.math.cnrs.fr/

ADAMA 12, Mahdia
Ecole d'automne en Analyse d'Algorithmes et Modèles Aléatoires
17-21 octobre 2012

Contents

1 Arbres et booléens

Introduction

Dans le cours de Danièle Gardy, deux types d'arbres, Catalan et GW, ont permis de représenter les expressions booléennes. On va ici considérer un autre type d'arbres, qui poussent comme un arbre binaire de recherche. On rappelle quelques notations : l'ensemble de tous les arbres binaires étiquetés est appelé $\mathcal{E}_k$. L'application Φ

$$\begin{aligned}\Phi \ : \ & \mathcal{E}_k \to \mathcal{F}_k \\ & t \mapsto \Phi(t) = f \text{ si et seulement si } t \text{ calcule } f.\end{aligned}$$

permet de transporter une loi de probabilité sur les arbres vers une loi de probabilité sur les fonctions booléennes $\mathcal{F}_k$.

Le plan de cette partie :

- definition d'un processus d'arbres qui poussent comme l'abr ;
- un processus d'arbres étiquetés pour représenter des fonctions booléennes ;
- le résultat ;
- deux méthodes possibles pour démontrer le résultat.

1.1 Le modèle des abr pour les expressions booléennes

NB : la taille d'un arbre binaire est son nombre de nœuds internes.

Par définition, le *processus abr* $(\mathcal{T}_n)_{n\in\mathbb{N}}$ est une suite d'arbres binaires définis récursivement par

- $\mathcal{T}_0$ est réduit à un sommet.
- Sachant $\mathcal{T}_n$, une feuille est choisie uniformément au hasard et elle est remplacée par un sommet interne et deux feuilles. Le nouvel arbre est $\mathcal{T}_{n+1}$.

L'arbre aléatoire $\mathcal{T}_n$ est appelé abr de taille n.

Exercice. Soit $n \geq 1$. Montrer que les sous-arbres de $\mathcal{T}_n$, l'abr de taille n, sont eux-mêmes des abr et que la probabilité que le sous-arbre gauche soit de taille $k \in \{0, \ldots, n-1\}$ est égale à $\dfrac{1}{n}$.

Le modèle Et/Ou.

On étiquette $\mathcal{T}_n$ en mettant indépendamment sur chaque feuille un littéral uniformément choisi parmi $\{x_1, \bar{x}_1, \ldots, x_k, \bar{x}_k\}$ et sur chaque nœud interne un connecteur $\wedge$ ou $\vee$ choisi

indépendamment avec probabilité $\frac{1}{2}$. L'arbre ainsi étiqueté est noté $\mathcal{T}_{n,k}$ et sa loi est appelée $\mathbb{P}_{n,k}$. La loi image par Φ est $p_{n,k}$: $\forall f \in \mathcal{F}_k$,

$$p_{n,k}(f) = \mathbb{P}(\mathcal{T}_{n,k} \text{ calcule } f).$$

Nous allons répondre aux questions : y a-t-il une limite de $p_{n,k}$ quand $n \to +\infty$? Si oui, comment décrire cette nouvelle loi sur $\mathcal{F}_k$?

Le théorème qui suit montre que la limite p_k des $p_{n,k}$ quand $n \to +\infty$ existe. Elle est surprenamment simple puisqu'elle ne charge que les fonctions constantes *Vrai* et *Faux*.

Théorème 1 *(loi limite pour le modèle Et/Ou)*

$$p_{n,k} \underset{n\to\infty}{\longrightarrow} p_k = \frac{1}{2}\delta_{Vrai} + \frac{1}{2}\delta_{Faux}$$

où pour toute fonction $f \in \mathcal{F}_k$, δ_f est la loi de probabilité définie par $\delta_f(f) = 1$.
De plus, $\|p_{n,k} - p_k\|_\infty = \mathcal{O}\left(\frac{1}{\ln n}\right)$, quand $n \to +\infty$[1].

Ce résultat peut être obtenu par deux méthodes différentes, soit en utilisant des fonctions génératrices et de la combinatoire analytique (voir dans [1]), soit par une méthode probabiliste, qui est détaillée maintenant.

1.2 La méthode du plongement en temps continu appliquée à l'abr.

Au lieu de pousser à des instants discrets, on considère maintenant que l'arbre pousse à des instants aléatoires, en temps continu: chaque feuille, indépendamment des autres, pousse à un instant aléatoire, de loi exponentielle de paramètre égal à 1. Le processus d'arbres ainsi défini s'appelle un *arbre de Yule*, nous le notons $(\mathcal{Y}_t)_{t\geq 0}$.

L'arbre de Yule est lié à l'abr : appelons $n(t)$ le nombre de nœuds internes de $\mathcal{Y}_t$. Comme le mécanisme de pousse est le même (à cause des lois exponentielles indépendantes, la première feuille qui pousse est choisie uniformément parmi les feuilles présentes à l'instant t), on a

$$(\mathcal{Y}_t)_{t\geq 0} \overset{\mathcal{L}}{=} (\mathcal{T}_{n(t)})_{t\geq 0}.$$

De manière duale, si on appelle τ_n le premier instant où il y a n feuilles dans un arbre de Yule : $\tau_n := \inf\{t \geq 0, n(t) = n\}$, alors les arbres $\mathcal{Y}_{\tau_n}$ et $\mathcal{T}_n$ ont même loi, et les processus aussi :

$$(\mathcal{Y}_{\tau_n})_{n\geq 0} \overset{\mathcal{L}}{=} (\mathcal{T}_n)_{n\geq 0}.$$

De plus, comme tous ces processus peuvent être considérés sur le même espace de probabilité, les égalités ci-dessus sont en fait des égalités presque sûres. Le gain important obtenu par ce

[1] Dans ce cours, $\|\cdot\|_\infty$ est définie sur l'ensemble des mesures signées sur $\mathcal{F}_k$ par : pour toute loi p, $\|p\|_\infty = \sup_{f\in\mathcal{F}_k} p(f)$.

plongement réside dans le fait que en tout nœud de l'arbre de Yule, les sous-arbres gauche et droit de ce nœud sont *indépendants* alors que ce n'était pas le cas dans l'arbre discret.

Comme pour l'arbre discret, on étiquette $\mathcal{Y}_t$ en mettant indépendamment sur chaque feuille un littéral uniformément choisi parmi $\{x_1, \bar{x}_1, \ldots, x_k, \bar{x}_k\}$ et sur chaque nœud interne un connecteur $\wedge$ ou $\vee$ choisi indépendamment avec probabilité $\frac{1}{2}$. L'arbre de Yule ainsi étiqueté est noté $\mathcal{Y}_{t,k}$ et sa loi est appelée $P_{t,k}$. La loi image par Φ est $\mathfrak{p}_{t,k} : \forall f \in \mathcal{F}_k$,

$$\mathfrak{p}_{t,k}(f) = \mathbb{P}(\mathcal{Y}_{t,k} \text{ calcule } f).$$

Preuve du théorème (grandes lignes).

Grâce aux liens entre arbre de Yule et arbre discret (qu'il faut néanmoins soigneusement écrire pour le passage de $\mathfrak{p}_{t,k}$ à $p_{n,k}$), il suffit de montrer que $\mathfrak{p}_{t,k} \to p_k = \frac{1}{2}\delta_{Vrai} + \frac{1}{2}\delta_{Faux}$ quand $t \to +\infty$.

Pour montrer que la limite de $\mathfrak{p}_{t,k}$ ne charge que les constantes, on montre que la probabilité pour avoir deux images différentes par une fonction booléenne donnée tend vers 0 quand $t \to +\infty$. Autrement dit on va montrer que pour $a \neq b \in \{0,1\}^k$ fixés, pour $\alpha \neq \beta \in \{0,1\}$ fixés, pour $f \in \mathcal{F}_k$,

$$\mathfrak{p}_{t,k}\left(f(a) = \alpha, f(b) = \beta\right) \underset{t\to\infty}{\longrightarrow} 0.$$

Pour des raisons de symétrie, il suffit de considérer $\alpha = 1$ et $\beta = 0$. Appelons pour simplifier $P_t := \mathfrak{p}_{t,k}\left(f(a) = 1, f(b) = 0\right)$. En conditionnant par rapport au premier instant de saut, qui est exponentiel de paramètre 1 et parce que les sous-arbres obtenus à cet instant sont indépendants, on obtient une équation différentielle sur P_t du type

$$e^t P_t = \frac{c}{2k} + \int_0^t (P_s - P_s^2) e^s ds$$

où c est une constante dépendant de a et b. Cette équation se résout en $P_t = \frac{1}{t+t_0}$, ce qui donne le résultat. □

1.3 Une urne de Pólya cachée dans un arbre d'implication

Dans ce paragraphe, on considère un processus d'abr $(\mathcal{T}_n)_{n\in\mathbb{N}}$ étiquetés différemment : on étiquette $\mathcal{T}_n$ en mettant indépendamment sur chaque feuille un littéral *positif* uniformément choisi parmi $\{x_1, \ldots, x_k\}$ et sur chaque nœud interne un connecteur $\to$ qui exprime l'implication. Pour simplifier, on garde les mêmes notations que pour le modèle Et/Ou : l'abr de taille n ainsi étiqueté est noté $\mathcal{T}_{n,k}$ et sa loi est appelée $\mathbb{P}_{n,k}$. La loi image par Φ est $p_{n,k} : \forall f \in \mathcal{F}_k$,

$$p_{n,k}(f) = \mathbb{P}(\mathcal{T}_{n,k} \text{ calcule } f).$$

Dans ce modèle, tout arbre peut se représenter comme un "peigne" car la branche droite est une succession d'implications jusqu'à la feuille finale α, appelée le but, et les dents du peigne sont eux-mêmes des arbres et s'appellent les *prémisses*.

On s'intéresse aux *tautologies*, c'est-à-dire aux arbres qui calculent la fonction *Vrai*. Sous les modèles de Catalan et de Galton-Watson, il a été prouvé dans [5] que asymptotiquement en k, toutes les tautologies sont simples. On va voir si c'est encore le cas pour les abr.

Un *arbre tautologique simple* est un arbre dans lequel au moins une prémisse est une feuille et cette feuille a même étiquette que le but. L'ensemble des arbres tautologiques simples de taille n est noté $\mathcal{ST}_{n,k}$. Le théorème suivant indique que le comportement est très différent pour les abr, puisque asymptotiquement en k, aucune tautologie n'est simple.

Théorème 2

$$\mathbb{P}_{n,k}(\mathcal{ST}_{n,k}) \xrightarrow{n\to+\infty} 1 - e^{-\frac{1}{k}} \sim \frac{1}{k} \text{ quand } k \to +\infty.$$

Preuve.

Appelons X_n le nombre de prémisses réduites à une feuille dans un arbre $\mathcal{T}_{n,k}$. On les appelle les "bonnes prémisses". L'idée de la preuve consiste à conditionner par ce nombre. La loi du nombre de bonnes prémisses est calculable exactement, grâce à une modélisation par une urne :

Les boules de l'urne sont les feuilles de l'arbre, elles sont de trois types. Les feuilles de type 2 sont les bonnes prémisses. Les feuilles de type 1 sont la feuille but. Les feuilles de type 3 sont les feuilles des prémisses qui ne sont pas bonnes. L'insertion a bien lieu uniformément sur les feuilles/boules. La matrice de remplacement de cette urne est

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

et X_n est le nombre de boules de type 1 dans l'urne à l'instant n. Grâce aux méthodes de combinatoire analytique développées par Flajolet par exemple dans [4] et détaillées dans le cours de Nicolas Pouyanne (ADAMA 12, Mahdia), Morcrette [9] a démontré que pour tout $n \geq 1$, pour tout $m \leq n$,

$$\mathbb{P}_{n,k}(X_n = m) = \frac{1}{m!} \sum_{j=0}^{n-m} \frac{(-1)^j}{j!},$$

ce qui montre en passant que le nombre de bonnes prémisses converge en loi lorsque $n \to +\infty$ vers une loi de Poisson de paramètre 1.

On calcule maintenant $\mathbb{P}_{n,k}(\mathcal{ST}_{n,k})$ en conditionnant par le nombre de bonnes prémisses. On remarque que la probabilité pour que l'une des bonnes prémisses soit étiquetée comme le but est égale à $(1 - (1 - \frac{1}{k})^m)$.

$$\mathbb{P}_{n,k}(\mathcal{ST}_{n,k}) = \sum_{m=1}^{n} \mathbb{P}_{n,k}(X_n = m) \left(1 - (1 - \frac{1}{k})^m\right).$$

Posons $c = (1 - \frac{1}{k})$.

$$\mathbb{P}_{n,k}(\mathcal{ST}_{n,k}) = \sum_{m=1}^{n} \frac{1 - c^m}{m!} \sum_{j=0}^{n-m} \frac{(-1)^j}{j!}.$$

Le théorème de Fubini s'applique pour cette double somme. Lorsque $n \to +\infty$, la deuxième somme tend vers $\frac{1}{e}$ et la première tend vers $e - e^c$. D'où finalement $\mathbb{P}_{n,k}(\mathcal{ST}_{n,k}) \to 1 - e^{-\frac{1}{k}}$.□

2 Urnes de Pólya en temps discret

[Travail en cours, en commun avec Quansheng Liu, Cécile Mailler et Nicolas Pouyanne]
Rappel, déjà vu dans le cours de Nicolas Pouyanne :

définition d'une urne de matrice de remplacement $R = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Le vecteur composition s'appelle $U^{DT}(n)$; partant de α boules rouges et β boules noires il s'appelle $U^{DT}_{(\alpha,\beta)}(n)$. Dans la décomposition spectrale apparaissent m, S et σ. On rappelle que
v_1 et v_2 sont deux vecteurs propres de ${}^tR$ associés respectivement à S et m :

$$v_1 = \frac{S}{(b+c)} \begin{pmatrix} c \\ b \end{pmatrix} \qquad\qquad v_2 = \frac{S}{(b+c)} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{1}$$

et u_1 et u_2 sont les deux formes linéaires duales (appelées parfois vecteurs propres à gauche)

$$u_1(x,y) = \frac{1}{S}(x+y) \qquad\qquad u_2(x,y) = \frac{1}{S}(bx - cy). \tag{2}$$

2.1 Parenthèse : l'urne originelle de Pólya

2.1.1 Le résultat

Etudiée par Pólya [10] en dimension 2, il s'agit de l'urne dont la matrice de remplacement est diagonale, égale à SI_d, où S est un entier, $S \geq 1$ et I_d est la matrice identité en dimension d, $d \geq 2$. Appelons P_n le vecteur composition de cette urne à l'instant n. Supposons que la composition initiale de l'urne est ${}^t(\alpha_1, \dots, \alpha_d)$ où $(\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d \setminus \{0\}$. Alors presque sûrement et dans tous les L^t, $t \geq 1$, on a

$$\frac{P_n}{nS} \underset{n\to\infty}{\longrightarrow} V$$

où V est un vecteur de taille d qui suit une loi de Dirichlet de paramètres $(\frac{\alpha_1}{S}, \dots, \frac{\alpha_d}{S})$.
Ce résultat peut être trouvé dans le livre de Johnson et Kotz [8], section 6.3.3 page 376. La preuve de ce résultat peut être déduite de Gouet [6] et Pouyanne [11] mais on en donne une preuve autonome dans la Section 2.1.3.

2.1.2 Loi de Dirichlet

On rappelle ici ce qu'est une loi de Dirichlet. Soit $d \geq 2$ un entier. Soit Σ le simplexe de dimension $(d-1)$ dans $\mathbb{R}^d$:

$$\Sigma = \left\{ (x_1, \dots x_d) \in [0,1]^d, \ \sum_{k=1}^{d} x_k = 1 \right\}.$$

Soit $d\Sigma$ la mesure positive sur le simplexe Σ définie par

$$\begin{aligned} &f(x_1, \dots, x_d) d\Sigma(x_1, \dots, x_d) \\ &\quad = f(x_1, \dots, x_{d-1}, 1 - \textstyle\sum_{k=1}^{d-1} x_k) \mathbb{1}_{\{x \in [0,1]^{d-1}, \ \sum_{k=1}^{d-1} x_k \leq 1\}} dx_1 \dots dx_{d-1}. \end{aligned}$$

On a la généralisation suivante de la fonction Beta : soient $(\nu_1, \dots, \nu_d)$ des réels positifs, alors

$$\int_\Sigma \left[\prod_{k=1}^{d} x_k^{\nu_k - 1} \right] d\Sigma (x_1, \dots, x_d) \ = \ \frac{\Gamma(\nu_1) \dots \Gamma(\nu_d)}{\Gamma(\nu_1 + \dots + \nu_d)}. \tag{3}$$

Définition 1 *On appelle* loi de Dirichlet *de paramètres* $(\nu_1, \dots, \nu_d)$*, notée Dirichlet* $(\nu_1, \dots, \nu_d)$*, la loi de probabilité dont la densité est donnée par*

$$\frac{\Gamma(\nu_1 + \dots + \nu_d)}{\Gamma(\nu_1) \dots \Gamma(\nu_d)} \left[\prod_{k=1}^{d} x_k^{\nu_k - 1} \right] d\Sigma (x_1, \dots, x_d).$$

On peut calculer les moments d'une loi de Dirichlet : soit $p = (p_1, \dots, p_d) \in \mathbb{N}^d$, alors le moment (joint) d'ordre p de $D = (D_1, \dots, D_d)$ est

$$\mathbb{E}(D^p) = \mathbb{E}(D_1^{p_1} \dots D_d^{p_d}) = \frac{\Gamma(\nu)}{\Gamma(\nu + |p|)} \prod_{k=1}^{d} \frac{\Gamma(\nu_k + p_k)}{\Gamma(\nu_k)}$$

où $\nu = \sum_{k=1}^{d} \nu_k$ et $|p| = \sum_{k=1}^{d} p_k$.

Enfin, les lois marginales d'une loi de Dirichlet $D = (D_1, \dots, D_d)$: la k-ème marginale D_k suit une loi $Beta(\nu_k, \nu - \nu_k)$ c'est-à-dire admet la densité suivante sur l'intervalle $[0,1]$:

$$\frac{1}{B(\nu_k, \nu - \nu_k)} t^{\nu_k - 1} (1-t)^{\nu - \nu_k - 1} \mathbb{1}_{[0,1]} dt.$$

2.1.3 Preuve du résultat

La preuve repose sur une convergence de martingale et sur la caractérisation de la limite par ses moments.

Soit $\alpha = \sum_{k=1}^{d} \alpha_k \geq 1$ le nombre de boules dans l'urne à l'instant initial. En prenant l'espérance conditionnelle sachant le passé avant n, on a

$$\mathbb{E}\left(P_{n+1}|\mathcal{F}_n\right) = \frac{\alpha + (n+1)S}{\alpha + nS} P_n$$

de sorte que $\left(\frac{P_n}{\alpha+nS}\right)_{n\geq 0}$ est une martingale à valeurs dans $[0,1]^d$, de moyenne ${}^t(\alpha_0/\alpha, \ldots, \alpha_d/\alpha)$ et qui converge vers une limite V. Cherchons les moments de V.
Pour toute fonction f définie sur $\mathbb{R}^d$,

$$\mathbb{E}\left(f(P_{n+1})|\mathcal{F}_n\right) = \left(I + \frac{\Phi}{\alpha + nS}\right)(f)(P_n)$$

où

$$\Phi(f)(v) = \sum_{k=1}^{d} v_k \left[f(v + Se_k) - f(v)\right]$$

(e_k est le k-ème vecteur de la base canonique de $\mathbb{R}^d$ et $v = \sum_{k=1}^{d} v_k e_k$).
Un petit calcul montre que pour $p = (p_1, \ldots, p_d) \in \mathbb{N}^d$ et $|p| = \sum_{k=1}^{d} p_k$, les fonctions définies sur $\mathbb{R}^d$ par

$$\Gamma_p(v) = \prod_{k=1}^{d} \frac{\Gamma\left(\frac{v_k}{S} + p_k\right)}{\Gamma\left(\frac{v_k}{S}\right)},$$

sont des fonctions propres pour l'opérateur Φ, associées à la valeur propre $|p|S$. Par conséquent, pour tout $p \in \mathbb{N}^d$,

$$\mathbb{E}\left(\Gamma_p(P_n)\right) = \frac{\Gamma\left(\frac{\alpha}{S} + n + |p|\right)}{\Gamma\left(\frac{\alpha}{S} + n\right)} \frac{\Gamma\left(\frac{\alpha}{S}\right)}{\Gamma\left(\frac{\alpha}{S} + |p|\right)} \Gamma_p(P_0)$$

et avec la formule de Stirling :

$$\mathbb{E}\left(\Gamma_p(P_n)\right) = n^{|p|} \frac{\Gamma\left(\frac{\alpha}{S}\right)}{\Gamma\left(\frac{\alpha}{S} + |p|\right)} \Gamma_p(P_0) \left(1 + O\left(\frac{1}{n}\right)\right).$$

Par ailleurs les polynômes $X^p = X_1^{p_1} \ldots X_d^{p_d}$ se développent dans la base $(\Gamma_p)_{p\in\mathbb{N}^d}$:

$$X^p = S^{|p|}\Gamma_p + \sum_{\substack{k\in\mathbb{N}^d \\ |k|\leq|p|-1}} a_{p,k}\Gamma_k(X)$$

où les $a_{p,k}$ sont des rationnels. Par conséquent

$$\mathbb{E}\left(\frac{P_n}{\alpha + nS}\right)^p = \frac{\Gamma\left(\frac{\alpha}{S}\right)}{\Gamma\left(\frac{\alpha}{S} + |p|\right)} \Gamma_p(P_0) \left(1 + O\left(\frac{1}{n}\right)\right).$$

et en passant à la limite quand n tend vers l'infini, pour tout $p \in \mathbb{N}^d$:

$$\mathbb{E}\left(V^p\right) = \frac{\Gamma\left(\frac{\alpha}{S}\right)}{\Gamma\left(\frac{\alpha}{S} + |p|\right)} \prod_{k=1}^{d} \frac{\Gamma\left(\frac{\alpha_k}{S} + p_k\right)}{\Gamma\left(\frac{\alpha_k}{S}\right)}. \tag{4}$$

ce qui prouve la convergence de la martingale dans L^t pour tout $t \geq 1$ et caractérise sa limite car une loi de Dirichlet est caractérisée par ses moments. ■

2.2 Asymptotique d'une urne de Pólya en temps discret

L'asymptotique fait apparaître une transition de phase, suivant la position de σ par rapport à $\frac{1}{2}$. Quand $\sigma \leq \frac{1}{2}$, il y a un théorème central limite, c'est-à-dire une limite en loi Gaussienne, et on dit que l'urne est "petite". Quand $\sigma > \frac{1}{2}$, on dit que l'urne est "grande" et

Théorème 3 *(Janson [7], Pouyanne[11]) Quand n tend vers $+\infty$,*

$$U^{DT}_{(\alpha,\beta)}(n) = nv_1 + n^\sigma W^{DT}_{(\alpha,\beta)} v_2 + o(n^\sigma) \tag{5}$$

où la convergence indiquée par le o est presque sûre et dans tous les $L^p, p \geq 1$ et où $W^{DT}_{(\alpha,\beta)}$ est définie par

$$W^{DT}_{(\alpha,\beta)} := \lim_{n\to+\infty} \frac{1}{n^\sigma} u_2 \left(U^{DT}_{(\alpha,\beta)}(n)\right). \tag{6}$$

Cette mystérieuse nouvelle loi, celle $W^{DT}_{(\alpha,\beta)}$, va maintenant être étudiée de plusieurs façons :

- on se ramène à l'étude de $X := W^{DT}_{(1,0)}$ et $Y := W^{DT}_{(0,1)}$ en mettant en exergue la structure arborescente de l'urne ;

- on écrit des équations de "convolution" sur le vecteur composition de l'urne en tirant parti encore de cette structure arborescente, et on en déduit que X et Y sont solution d'une équation de point fixe en distribution. Des propriétés de X et Y s'en déduisent.

- on plonge l'urne en temps continu, ce qui permet de gagner de l'indépendance en considérant $U(t)$, analogue continu de U_n. C'est un processus de branchement et son asymptotique fait apparaître une variable aléatoire W^{CT}. Elle est liée à W^{DT} par une connexion explicite.

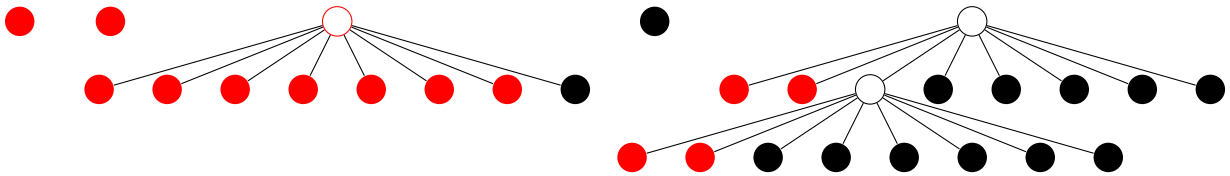
- pour étudier W^{CT}, on écrit des équations de "convolution" sur $U(t)$ et on en déduit que $X := W^{CT}_{(1,0)}$ et $Y := W^{CT}_{(0,1)}$ sont solution d'une équation de point fixe en distribution. Ce ne sont pas les mêmes X et Y, ni la même équation.

2.3 Structure arborescente de l'urne de Pólya discrète

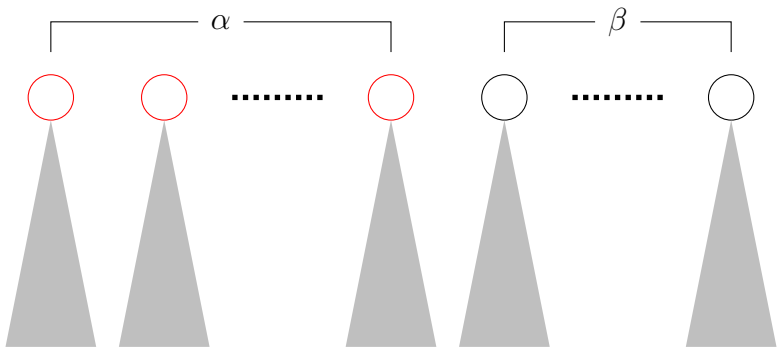
Convenons que les boules dans l'urne à l'instant n sont les feuilles d'un arbre T_n, de sorte que "tirer au hasard uniformément une boule dans l'urne" revient à tirer au hasard uniformément une feuille de l'arbre T_n. Ainsi, les enfants d'une feuille u de T_n sont les boules créées quand

on a tiré cette boule u. On appelle sous-arbre d'une boule u l'ensemble constitué de u et de ses descendants.

Par exemple, prenons comme matrice de remplacement $R = \begin{pmatrix} 6 & 1 \\ 2 & 5 \end{pmatrix}$, (c'est une grande urne) et partons de $\alpha = 3$ boules rouges et $\beta = 2$ boules noires.



Numérotons les boules initiales rouges de 1 à α et les boules initiales noires de $\alpha+1$ à $\alpha+\beta$. On a une forêt issue de ces boules initiales.



Appelons $D_k(n)$ le nombre de feuilles à l'instant n du k-ième sous-arbre, c'est-à-dire la taille du k-ième sous-arbre. Alors le nombre de tirages qui ont eu lieu dans ce k-ième sous-arbre est égal à $\frac{D_k(n)-1}{S}$. C'est aussi le temps à l'intérieur de ce k-ième sous-arbre. Et comme les tirages de boules sont uniformes à tout instant, on a

$$U_{(\alpha,\beta)}(n) \overset{\mathcal{L}}{=} \sum_{k=1}^{\alpha} U_{(1,0)}^{(k)}\left(\tfrac{D_k(n)-1}{S}\right) + \sum_{k=1+\alpha}^{\alpha+\beta} U_{(0,1)}^{(k)}\left(\tfrac{D_k(n)-1}{S}\right) \tag{7}$$

où $U_{(1,0)}^{(k)}$ et $U_{(0,1)}^{(k)}$ sont des copies indépendantes de $U_{(1,0)}$ et $U_{(0,1)}$.

Maintenant remarquons qu'à chaque tirage $D_k(n)$ augmente de S et encore mieux : $(D_1(n), \ldots, D_{\alpha+\beta}(n))$ est exactement le vecteur composition d'une urne originelle de Pólya ayant comme matrice de remplacement $SI_{\alpha+\beta}$ et partant du vecteur initial ${}^t(1, \ldots, 1)$. Par conséquent, prenons l'équation (7), appliquons la seconde projection u_2, divisons par n^σ, passons à la limite quand $n \to \infty$ (en remarquant que $D_k(n)$ aussi tend vers $+\infty$), et

utilisons (5) et le résultat de la Section 2.1. On obtient :

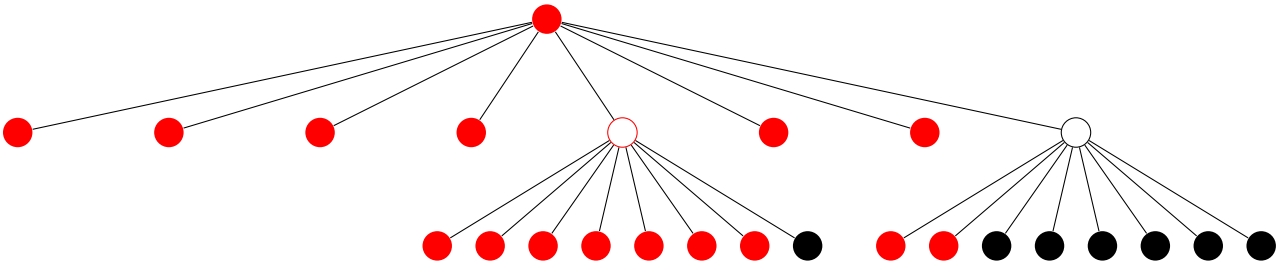
$$W_{(\alpha,\beta)} \overset{\mathcal{L}}{=} \sum_{k=1}^{\alpha} (Z_k)^{\sigma} W_{(1,0)}^{(k)} + \sum_{k=\alpha+1}^{\alpha+\beta} (Z_k)^{\sigma} W_{(0,1)}^{(k)}$$

où

(i) $Z = (Z_1, \ldots, Z_{\alpha+\beta})$ est un vecteur de loi de Dirichlet de paramètres $(\frac{1}{S}, \ldots, \frac{1}{S})$;

(ii) les $W_{(1,0)}^{(k)}$ et les $W_{(0,1)}^{(k)}$ sont des copies indépendantes de $W_{(1,0)}$ et $W_{(0,1)}$, et indépendantes de Z.

On s'est ainsi ramenés à l'étude de $W_{(1,0)}$ et $W_{(0,1)}$. Concentrons-nous donc maintenant sur l'étude de $U_{(1,0)}(n)$. A l'instant 1, la composition de l'urne est déterministe : il y a $(a+1)$ boules rouges et b boules noires. Il y a toujours la structure arborescente. Dans l'exemple ci-dessus, pour $R = \begin{pmatrix} 6 & 1 \\ 2 & 5 \end{pmatrix}$,



et si on appelle $J_k(n)$ le nombre de feuilles du k-ième sous-arbre issu de la boule numéro k présente à l'instant 1 dans l'urne, on a comme précédemment :

$$U_{(1,0)}(n) \overset{\mathcal{L}}{=} \sum_{k=1}^{a+1} U_{(1,0)}^{(k)}\left(\tfrac{J_k(n)-1}{S}\right) + \sum_{k=a+2}^{S+1} U_{(0,1)}^{(k)}\left(\tfrac{J_k(n)-1}{S}\right) \tag{8}$$

où $U_{(1,0)}^{(k)}$ et $U_{(0,1)}^{(k)}$ sont des copies indépendantes de $U_{(1,0)}$ et $U_{(0,1)}$. Par un raisonnement analogue à celui mené ci-dessus, on obtient pour X et Y qui sont des notations simplifiées :

$$\begin{cases} X := W_{(1,0)}^{DT} = \lim\limits_{n\to+\infty} u_2\left(\dfrac{U_{(1,0)}(n)}{n^{\sigma}}\right) \\ \\ Y := W_{(0,1)}^{DT} = \lim\limits_{n\to+\infty} u_2\left(\dfrac{U_{(0,1)}(n)}{n^{\sigma}}\right) \end{cases} \tag{9}$$

le système d'équations de point fixe suivant :

$$\begin{cases} X \overset{\mathcal{L}}{=} \sum_{k=1}^{a+1} (V_k)^\sigma X^{(k)} + \sum_{k=a+2}^{S+1} (V_k)^\sigma Y^{(k)} \\ \\ Y \overset{\mathcal{L}}{=} \sum_{k=1}^{c} (V_k)^\sigma X^{(k)} + \sum_{k=c+1}^{S+1} (V_k)^\sigma Y^{(k)} \end{cases} \tag{10}$$

où

(i) $V = (V_1, \ldots, V_{S+1})$ est un vecteur aléatoire de loi de Dirichlet de paramètres $(\frac{1}{S}, \ldots, \frac{1}{S})$;

(ii) les $X^{(k)}$ et les $Y^{(k)}$ sont des copies de X and Y, indépendantes les unes des autres et de V.

Remarque : chaque V_k a même loi que U^S, U de loi uniforme sur $[0, 1]$. Autrement dit, $(V_k)^\sigma$ a même loi que U^m.

Ce système d'équations de point fixe donne des renseignements sur X et Y : caractérisation par méthode de contraction, existence de densité par méthode d'analyse de Fourier à la Liu, ordre de grandeur des moments (voir [2]).

3 Urnes de Pólya en temps continu

3.1 Le plongement en temps continu des urnes de Pólya

Suivant une méthode classique, on va considérer maintenant l'analogue continu de l'urne de Pólya discrète. Définissons pour cela le processus de branchement à temps continu $(U^{CT}_{(\alpha,\beta)}(t))_{t\geq 0}$ de la façon suivante :

- il part de la même condition initiale (α, β) ;

- chaque boule (ou particule, ou individu) est munie d'une loi de durée de vie (une horloge) exponentielle de paramètre 1, et toutes ces lois sont indépendantes ;

- quand une horloge rouge sonne, a boules rouges et b boules noires sont ajoutées dans l'urne ;

- quand une horloge noire sonne, c boules rouges et d boules noires sont ajoutées dans l'urne.

Le mécanisme de remplacement est donc le même que dans le processus discret.

Les instants de saut du processus continu sont

$$0 = \tau_0 < \tau_1 < \cdots < \tau_n < \cdots$$

Le principe du plongement dit que que le processus continu arrêté au temps τ_n est le même que le processus discret.

$$\left(U^{CT}(\tau_n)\right)_{n\geq 0} = \left(U^{DT}(n)\right)_{n\geq 0}.$$

Et de manière duale, si on appelle

$$n(t) := \inf\{n \geq 0, \tau_n \geq t\}.$$

le nombre de tirages effectués avant l'instant t, alors le processus continu est le même que le processus discret à l'instant discret $n(t)$.

$$\left(U^{CT}(t)\right)_{t\geq 0} = \left(U^{DT}\left(n(t)\right)\right)_{t\geq 0}$$

Le gain du plongement, comme pour les abr, c'est que dans le processus continu, les sous-arbres de n'importe quel nœud sont *indépendants*, donc on va pouvoir appliquer la propriété de branchement. C'est ce qui va produire les équations de convolution de la Section 3.3.

3.2 Asymptotique d'une urne de Pólya en temps continu

Pour une urne de Pólya plongée en temps continu, c'est-à-dire pour le processus de branchement $(U^{CT}(t))_{t\geq 0}$, il apparaît la même transition de phase qu'en temps discret :
quand $\sigma \leq \frac{1}{2}$, il y a un théorème central limite, c'est-à-dire une limite en loi Gaussienne ;
quand $\sigma > \frac{1}{2}$, on a le théorème suivant.

Théorème 4 *(Janson [7] ou [3])*
Quand t tend vers l'infini,

$$U^{CT}(t) = e^{St}\xi v_1\left(1 + o(1)\right) + e^{mt}W^{CT}v_2\left(1 + o(1)\right), \tag{11}$$

où ξ et W^{CT} sont des variables aléatoires à valeurs réelles et le petit o signifie une convergence presque sûre et dans tous les L^p, $p \geq 1$.
De plus, ξ suit une loi Gamma(u/S) *où $u = \alpha + \beta$ est le nombre de boules à l'instant initial.*

3.3 Les équations de point fixe en temps continu.

Grâce au plongement en temps continu, la propriété de branchement s'applique au processus $\left(U^{CT}(t)\right)_{t\geq 0}$.
Première conséquence : le processus $(U^{CT}_{(\alpha,\beta)}(t), t \geq 0)$ partant de α boules rouges et β boules noires est la somme de α copies de $U^{CT}_{(1,0)}(t)$ (le processus partant de une boule rouge) et de β copies de $U^{CT}_{(0,1)}(t)$ (le processus partant de une boule noire). Nous sommes donc ramenés à étudier ces deux processus. Nous simplifions les notations de la façon suivante :

$$X := \lim_{t\to+\infty} e^{-mt}u_2(U^{CT}_{(1,0)}(t)) \qquad \text{and} \qquad Y := \lim_{t\to+\infty} e^{-mt}u_2(U^{CT}_{(0,1)}(t)). \tag{12}$$

Seconde conséquence : notons $\tau = \tau_1$ le premier instant de saut de chacun des deux processus $U^{CT}_{(1,0)}(t)$ et $U^{CT}_{(0,1)}(t)$; comme on part d'une seule boule, τ est de loi $\mathcal{E}xp(1)$. Alors pour tout

instant $t > \tau$, le processus $U^{CT}_{(1,0)}(t)$ à l'instant t a même loi que la somme de $(a+1)$ copies de $U^{CT}_{(1,0)}(t-\tau)$ et b copies de $U^{CT}_{(0,1)}(t-\tau)$ (et l'analogue pour $U^{CT}_{(0,1)}(t)$). Ce qui donne les équations de "dislocation" suivantes.

$$\forall t > \tau, \quad \begin{cases} U^{CT}_{(1,0)}(t) \stackrel{\mathcal{L}}{=} [a+1]U^{CT}_{(1,0)}(t-\tau) + [b]U^{CT}_{(0,1)}(t-\tau) \\ \\ U^{CT}_{(0,1)}(t) \stackrel{\mathcal{L}}{=} [c]U^{CT}_{(1,0)}(t-\tau) + [d+1]U^{CT}_{(0,1)}(t-\tau), \end{cases} \tag{13}$$

où la notation $[n]X$ signifie la somme de n copies de variables aléatoires indépendantes de même loi que X. Il suffit alors de renormaliser et de passer à la limite quand t tend vers l'infini en utilisant l'asymptotique (11) pour obtenir le système d'équations de point fixe suivant.

$$\begin{cases} X \stackrel{\mathcal{L}}{=} U^m \Big(\displaystyle\sum_{k=1}^{a+1} X^{(k)} + \sum_{k=a+2}^{S+1} Y^{(k)} \Big) \\ \\ Y \stackrel{\mathcal{L}}{=} U^m \Big(\displaystyle\sum_{k=1}^{c} X^{(k)} + \sum_{k=c+1}^{S+1} Y^{(k)} \Big), \end{cases} \tag{14}$$

où U suit une loi uniforme sur $[0,1]$ et où les $X^{(k)}$ et les $Y^{(k)}$ sont des copies de X et de Y respectivement, et sont indépendantes entre elles et indépendantes de U.

References

[1] B. Chauvin, D. Gardy, and C. Mailler. The growing tree distribution on boolean functions. *SIAM Proceedings Analco11*, pages 45–56, 2011.

[2] B. Chauvin, C. Mailler, and N. Pouyanne. More about the limit distributions for large Pólya urns. *Preprint*, 2012.

[3] B. Chauvin, N. Pouyanne, and R. Sahnoun. Limit distributions for large Pólya urns. *Annals Applied Prob.*, 21(1):1–32, 2011.

[4] P. Flajolet, P. Dumas, and V. Puyhaubert. Some exactly solvable models of urn process theory. *DMTCS Proceedings*, AG:59–118, 2006.

[5] H. Fournier, D. Gardy, A. Genitrini, and B. Gittenberger. The fraction of large random trees representing a given boolean function in implicational logic. *Random Structures and Algorithms*, 20(7):875–887, 2012.

[6] R. Gouet. Strong convergence of proportions in a multicolor Pólya urn. *J. Appl. Prob.*, 34:426–435, 1997.

[7] S. Janson. Functional limit theorem for multitype branching processes and generalized Pólya urns. *Stochastic Processes and their Applications*, 110:177–245, 2004.

[8] N.L. Johnson and S. Kotz. *Urn Models and Their Application*. Wiley, 1977.

[9] B. Morcrette. Combinatoire analytique et modèles d'urnes. Mémoire de master MPRI. 2010.

[10] G. Pólya. Sur quelques points de la théorie des probabilités. *Ann. Inst. Henri Poincaré*, 1:117–161, 1931.

[11] N. Pouyanne. An algebraic approach to Pólya processes. *Ann. Inst. Henri Poincaré*, 44(2):293–323, 2008.