Des arbres dans les urnes de Pólya, et réciproquement

Brigitte Chauvin
Résumé par Cécile Mailler

Séminaire de Combinatoire Philippe Flajolet
Institut Henri Poincaré, Séance du 6 décembre 2012

Résumé

Dans cet exposé, Brigitte Chauvin introduit le modèle des urnes de Pólya. L'exposé est centré sur l'étude des urnes à deux couleurs, et plus précisément des "grandes" urnes. L'oratrice montre comment utiliser pleinement la structure arborescente de ces urnes pour mieux les étudier. Elle précise que ces travaux sont en collaboration avec Nicolas Pouyanne, Quansheng Liu et Cécile Mailler. Outre les références citées dans ce résumé, Brigitte Chauvin conseille la lecture des notes de cours suivantes, rédigées à l'occasion de l'école d'été ADAMA 2012 :
- cours de Nicolas Pouyanne : `http://pouyanne.perso.math.cnrs.fr/mahdia2012.pdf`
- cours de Brigitte Chauvin : `http://chauvin.perso.math.cnrs.fr/coursMahdiaBC.pdf`

1 Introduction

Une urne de Pólya est décrite par une composition initiale $(\alpha, \beta)^t$, et par une matrice de remplacement

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Cela signifie que l'urne contient initialement α boules noires et β boules rouges ; et qu'à chaque étape, on tire uniformément au hasard une boule dans l'urne, on regarde sa couleur, on la remet dans l'urne, et on rajoute a boules noires et b boules rouges si elle était noire, ou c boules noires et d boules rouges si elle était rouge.

Les urnes étudiées dans cet exposé vérifient la propriété de balance $a + b = c + d = S$. La valeur S, appelée **balance** de l'urne est valeur propre de la matrice de remplacement. La seconde valeur propre est notée m. On note v_1 le vecteur propre associé à S et v_2 le vecteur propre associé à m. On supposera aussi $a, b, c, d \geq 0$, même si cette hypothèse peut être affaiblie : il suffit juste de s'assurer que la composition de l'urne restera toujours positive.

On note $U(n) = (B_n, R_n)^t$ la composition de l'urne à l'étape n, c'est à dire le nombre de boules noires B_n et le nombre de boules rouges R_n qu'elle contient. Le but de l'étude d'une urne est de mieux comprendre ce vecteur composition au temps n. Le comportement d'une urne dépend en fait du rapport entre ses deux valeurs propres $\sigma = \frac{m}{S}$.

Théorème 1.1. *(cf. Janson [5] ou Pouyanne [8] par exemple)*
- *Si $\sigma < \frac{1}{2}$, on dit que l'urne est* **petite**, *et on a le théorème limite suivant :*

$$\frac{U(n) - nv_1}{\sqrt{n}} \to G$$

en loi, quand n tend vers l'infini, avec G un vecteur Gaussien centré, de variance connue et explicite.
- *Si $\sigma > \frac{1}{2}$, on dit que l'urne est* **grande**, *et on a le théorème limite suivant :*

$$U(n) = nv_1 + n^\sigma W_{\alpha,\beta} v_2 + o(n^\sigma) \tag{1}$$

en loi et dans tous les L^p, $p \geq 1$, quand n tend vers l'infini.

On s'intéresse notamment dans cet exposé au cas des grandes urnes, et plus précisément à la variable aléatoire $W_{\alpha,\beta}$: quel est son support ? admet-elle une densité ? quels sont ses moments ?

Dans cet exposé, l'oratrice présente tout d'abord le cas de l'urne "originelle" de Pólya, qui est une urne à $d \geq 1$ couleurs, urne qui sera utile dans la suite. Elle présente ensuite les différentes méthodes qui permettent d'étudier les urnes : la combinatoire analytique, les méthodes "arborescentes", et finalement le plongement en temps continu.

2 L'urne "originelle"

L'urne originelle est une urne à $d \geq 2$ couleurs, représentée par la matrice de remplacement SI_d (ou I_d est la matrice identité en dimension d), et de composition initiale $(\alpha_1, \ldots, \alpha_d)$. Cette urne vérifie le théorème limite suivant (Pólya [7] ou Gouet [4] ou Johnson et Kotz [6]) :

$$\frac{U(n)}{nS} \to V$$

presque sûrement et dans tous les L^p, $p \geq 1$, quand n tend vers l'infini ; où V est un vecteur aléatoire de Dirichlet de paramètres $\left(\frac{\alpha_1}{S}, \ldots, \frac{\alpha_d}{S}\right)$.

3 L'étude par combinatoire analytique

Les urnes de Pólya ont été étudiées via des méthodes probabilistes, mais aussi, plus récemment par combinatoire analytique (voir les travaux de Flajolet [3] et [2]). Dans ce domaine, on représente la composition de l'urne par un mot $B^p R^q$ si l'urne contient p boules rouges et q boules noires. Si l'on tire une boule noire, on "transforme" un B en $B^{a+1}R^b$ et si l'on tire une boule rouge, on "transforme" un R en $B^c R^{d+1}$. La composition de l'urne au temps n est donnée par un mot W_n.

Définition 3.1. *On appelle* **histoire** *de longueur n menant de $\binom{u_0}{v_0}$ à $\binom{u}{v}$ une suite de mots $W_0 = B^{u_0}R^{v_0}, W_1, \ldots, W_n$ produits via des transformations $B \rightsquigarrow B^{a+1}R^b$ ou $R \rightsquigarrow B^c R^{d+1}$, avec W_n qui contient u symboles B et v symboles N. On note*

$$H_n \begin{pmatrix} u_0 & u \\ v_0 & v \end{pmatrix}$$

le nombre de ces histoires.

Enfin, on introduit la fonction génératrice des ces histoires :

$$H\left(x, y, z \middle| \begin{matrix} u_0 \\ v_0 \end{matrix}\right) = \sum_{u,v,n \in \mathbb{N}} H_n \begin{pmatrix} u_0 & u \\ v_0 & v \end{pmatrix} x^u y^v \frac{z^n}{n!}.$$

Cette représentation sous forme de mots mène au résultat suivant :

Théorème 3.2 (Flajolet, Dumas, Puyhaubert, 2006)**.** *Soient X et Y les solutions du problème de Cauchy*

$$\begin{cases} X' = X^{a+1}Y^b \\ Y' = X^c Y^{d+1} \\ X(0) = x, Y(0) = y \end{cases}$$

où x et y sont tels que $xy \neq 0$, alors, pour toute configuration initiale (α, β)

$$H\left(x, y, z \middle| \begin{matrix} \alpha \\ \beta \end{matrix}\right) = X(z)^\alpha Y(z)^\beta.$$

4 Structure arborescente

4.1 Équations de dislocation

L'idée de cette sous partie est de se ramener de l'étude de $W_{\alpha,\beta}$ (cf. Equation (1)) pour α et β quelconques, à l'étude de deux variables aléatoires W_{10} et W_{01}.

Un processus d'urne de Pólya peut être vu comme une forêt. Chaque boule de la composition initiale de l'urne est vue comme la racine d'un arbre. Lorsqu'une boule est tirée au hasard, elle devient un nœud interne et donne naissance à $S+1$ fils. Par exemple, si la boule tirée au hasard était noire, elle donne naissance à $a+1$ fils noirs et b fils rouges. Et la nouvelle composition de l'urne est donnée par les couleurs des feuilles de la forêt. A tout instant, cette forêt est constituée de $\alpha+\beta$ sous-arbres associés aux boules initiales.

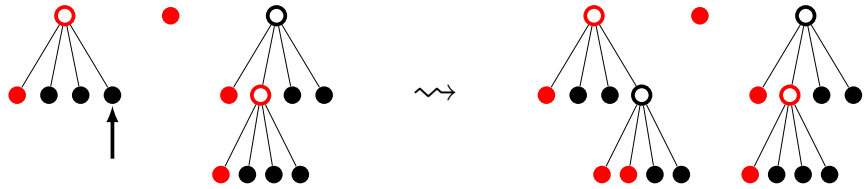

FIGURE 1 – *Structure arborescente de l'urne $a=2$, $b=1$, $c=3$ et $d=0$ de composition initiale $\alpha=1$ et $\beta=2$.*

Si l'on note $D_k(n)$ le nombre de feuilles à l'étape n dans l'arbre associé à la $k^{\text{ème}}$ boule initiale, comme à chaque tirage, on ajoute S boules dans l'arbre considéré, $\frac{D_k(n)-1}{S}$ est le nombre de tirages qui ont été effectués dans le sous-arbre k jusqu'à l'étape n. Ainsi,

$$U_{\alpha,\beta}(n) = \sum_{p=1}^{\alpha+1} U_{10}^{(p)}\left(\frac{D_p(n)-1}{S}\right) + \sum_{p=\alpha+2}^{\alpha+\beta} U_{01}^{(p)}\left(\frac{D_p(n)-1}{S}\right).$$

En remarquant que $(D_1(n),\dots,D_{\alpha+\beta}(n))$ est la composition d'une urne de Pólya originelle à l'étape n, on obtient, à la limite,

$$W_{\alpha,\beta} \overset{(loi)}{=} \sum_{p=1}^{\alpha}(V_p)^{\sigma} W_{10}^{(p)} + \sum_{p=1}^{\alpha+\beta}(V_p)^{\sigma} W_{01}^{(p)}$$

où $(V_1,\dots,V_{\alpha+\beta})$ est un vecteur aléatoire de loi de Dirichlet de paramètres $\left(\frac{1}{S},\dots,\frac{1}{S}\right)$.

4.2 Équations de points fixe

Si l'urne est composée initialement d'une seule boule noire, la première étape du processus est déterministe, et la composition de l'urne à l'étape 1 est $a+1$ boules noires et b boules rouges. En considérant ces $S+1$ boules comme les racines d'une forêt, et en raisonnant de même que précédemment, on obtient le système suivant :

$$\begin{cases} W_{10} \overset{(loi)}{=} \sum_{p=1}^{a+1} V_p^{\sigma} W_{10}^{(p)} + \sum_{p=a+2}^{S+1} V_p^{\sigma} W_{01}^{(p)} \\ W_{01} \overset{(loi)}{=} \sum_{p=1}^{c} V_p^{\sigma} W_{10}^{(p)} + \sum_{p=c+1}^{S+1} V_p^{\sigma} W_{01}^{(p)} \end{cases}.$$

Il est possible de démontrer par méthode de contraction que ce système admet, à moyenne fixée, une unique solution de carré intégrable. Ce système permet ensuite de démontrer l'existence d'une densité pour W_{01} et W_{10}, et estimer les moments de ces lois. Ce dernier objectif peut aussi être complété par une autre approche : celle du plongement en temps continu.

5 Plongement en temps continu

5.1 Définition et connexion

Les urnes de Pólya peuvent être "plongées en temps continu". Le processus d'urne en temps continu est le suivant. A l'instant initial, il y a α boules noires et β boules rouges dans l'urne. Chacune de ces boules est équipée d'une horloge qui sonnera au bout d'un temps aléatoire de loi exponentielle de paramètre 1, et ce indépendamment des autres. Lorsque l'horloge d'une boule sonne, elle se divise en $a+1$ boules noires et b boules rouges si elle était noire, ou en c boules noires et $d+1$ boules rouges si elle était rouge. On note $U^{CT}_{\alpha,\beta}(t)$ la composition de l'urne au temps t.

Si l'on note τ_n la date à laquelle une $n^{\text{ème}}$ horloge sonne, on sait que

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

et de plus, $(\tau_n)_{n\geq 0}$ est indépendant de $(U^{CT}(\tau_n))_{n\geq 0}$.

On a aussi un théorème limite en temps continu :

$$U^{CT}_{\alpha,\beta}(t) = n\xi v_1(1+o(1)) + n^\sigma W^{CT}_{\alpha,\beta} v_2(1+o(1))$$

presque sûrement et dans tous les L^p, $p \geq 1$. La variable aléatoire ξ est connue : elle suit une loi $\text{Gamma}\left(\frac{\alpha+\beta}{S}\right)$.

Les lois de $W_{\alpha,\beta}$ et $W^{CT}_{\alpha,\beta}$ sont reliées par une **martingale connexion** :

$$W^{CT}_{\alpha,\beta} \overset{(loi)}{=} \xi^\sigma \cdot W_{\alpha,\beta},$$

avec ξ de loi $\text{Gamma}\left(\frac{\alpha+\beta}{S}\right)$, et ξ et $W_{\alpha,\beta}$ indépendantes.

En conséquence, si nous avons des informations sur W^{CT}, nous en déduirons des informations sur W, et réciproquement. Il faut donc choisir le plus "facile à étudier".

5.2 Équations de dislocation et de point fixe

Tout comme dans le cas discret, on peut établir des équations de dislocation pour se ramener à l'étude de W^{CT}_{10} et W^{CT}_{01} :

$$W^{CT}_{\alpha,\beta} \overset{(loi)}{=} V^\sigma \left(\sum_{p=1}^{\alpha} X^{CT}_p + \sum_{p=\alpha+1}^{\alpha+\beta} Y^{CT}_p \right)$$

avec X^{CT}_p des copies indépendantes de $X = W^{CT}_{10}$ et Y^{CT}_p des copies indépendantes de $Y = W^{CT}_{01}$, et $V = U^S$ en loi, avec U uniforme sur $[0,1]$.

De plus, nous avons aussi un système de point fixe donné par les équations :

$$\begin{cases} X \overset{(loi)}{=} V^\sigma \left(\sum_{k=1}^{a+1} X_p + \sum_{k=a+2}^{S+1} Y_p \right) \\ Y \overset{(loi)}{=} V^\sigma \left(\sum_{k=1}^{c} X_p + \sum_{k=c+1}^{S+1} Y_p \right) \end{cases} \tag{2}$$

où $V = U^S$ en loi, avec U uniforme sur $[0,1]$. Tout comme dans le cas discret, on peut montrer que, à moyenne fixée, la solution du système de point fixe ci-dessus est unique. Ce système est plus simple qu'en cas discret, et permet en particulier de démontrer, par récurrence, que

Théorème 5.1. *Il existe une constante strictement positive C_2 telle que pour toute constante C_1, pour tout $p \geq 1$,*

$$C_1^p \leq \frac{\mathbb{E}|X|^p}{p!} \leq C_2^p (\log p)^p.$$

En particulier, la fonction génératrice des moments a un rayon de convergence nul. L'oratrice insiste sur le fait que l'ordre de grandeur de ces moments est encore inconnu.

5.3 Étude des transformées de Fourier

Si l'on pose

$$\begin{cases} \mathcal{F}(x) = \mathbb{E}\mathsf{e}^{ixX} \\ \mathcal{G}(x) = \mathbb{E}\mathsf{e}^{ixY} \end{cases}$$

En utilisant le système (2), on peut montrer que $\mathcal{F}$ et $\mathcal{G}$ vérifient le système d'équations différentielles suivant (cf. [1]) :

$$\begin{cases} \mathcal{F}(x) + mx\mathcal{F}'(x) = \mathcal{F}(x)^{a+1}\mathcal{G}^b(x) \\ \\ \mathcal{G}(x) + mx\mathcal{G}'(x) = \mathcal{F}(x)^c\mathcal{G}^{d+1}(x) \end{cases},$$

qui, via un changement de variable qui permet de le résoudre et de trouver les expressions explicites de $\mathcal{F}$ et $\mathcal{G}$, est équivalent au système suivant :

$$\begin{cases} f' = f^{a+1}g^b \\ \\ g' = f^c g^{d+1} \end{cases}.$$

Ce dernier système est le même que celui obtenu dans un cadre a priori totalement différent dans le Théorème 3.2. Pourquoi le système différentiel vérifié par les transformées de Fourier des deux variables limites en temps continu est égal à celui des fonctions génératrices des histoires dans le cadre d'une étude par combinatoire analytique ? Ce lien, s'il existe, reste encore mystérieux...

Références

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

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

[3] P. Flajolet, J. Gabarró, and H. Pekari. Analytic urns. *The Annals of Probability*, 33(3) :1200–1233, 2005.

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

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

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

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

[8] N. Pouyanne. Classification of large Pólya-Eggenberger urns with regard to their asymptotics. *Discrete Mathematics and Theoretical Computer Science, AD*, pages 275–286, 2005.