Processuss de Pólya

Brigitte CHAUVIN
Monastir, 26 Juin 2006

1 Définition

Des descriptions peuvent être trouvées dans [6] et [8]. Contrairement à un autre modèle dit lui aussi modèle d'urnes, on considère ici une seule urne qui contient des boules de différentes couleurs. Le modèle historique a été introduit par Eggenberger et Pólya en 1923 : l'urne contient des boules de deux couleurs B et W. On tire au hasard une boule dans l'urne (les boules sont supposées être tirées chacune avec la même probabilité) et si on a tiré une boule blanche (respectivement noire), on la remet dans l'urne puis on rajoute s boules blanches (respectivement s boules noires) dans l'urne. Appelons X_n le nombre de boules blanches dans l'urne après n tirages ; alors X_n est une chaîne de Markov mais non homogène.

De nombreuses généralisations de ce modèle sont possibles et ont donné lieu à une abondante littérature jusqu'à maintenant :

- à chaque tirage, on rajoute des boules de la même couleur que celle que l'on a tirée mais aussi d'autres couleurs ; l'opération est décrite par une matrice dite de remplacement R (appelée parfois aussi matrice d'addition) ;

- on peut retrancher des boules ;

- il y a plus de deux couleurs ;

- le nombre de boules que l'on rajoute varie avec le temps (ce qui donne un processus non homogène en temps) et c'est une variable aléatoire.

Nous allons dans la suite considérer les trois premières généralisations et appeler processus de Pólya une chaîne de Markov $(Y_n)_{n\geq 1}$ à valeurs dans $\mathbb{R}^k$ ou $\mathbb{C}^k$ (k est un entier fixé qui est le nombre de couleurs) dans laquelle la i-ème coordonnée $Y_n^{(i)}$, $i = 1, 2, \dots k$ représente le nombre de boules de couleur i à l'instant n. Etant donnée une matrice de remplacement $R = (r_{ij})$ à k lignes et k colonnes et à coefficients entiers relatifs, les règles d'évolution entre l'instant n et l'instant $n+1$ sont les suivantes :

- on tire au hasard une boule dans l'urne, les boules étant équiprobables.

- on note la couleur de la boule ainsi tirée et on la remet dans l'urne.

- si la boule tirée était de couleur i, on rajoute r_{ij} boules de couleur j.

Comme on autorise des coefficients négatifs de la matrice de remplacement, la question se pose de la viabilité de l'urne. Dans le modèle étudié ici, on supposera que seuls les coefficients diagonaux de R peuvent être négatifs. Une condition suffisante pour assurer dans ce cas la viabilité de l'urne est la condition arithmétique suivante qui lie les coefficients de R aux coefficients $Y_1^{(j)}$ de l'état initial ($n = 1$) : $\forall j \in \{1, 2, \dots, k\}$,

$$r_{jj} \geq 0 \quad \text{ou} \quad Y_1^{(j)}\mathbb{Z} + \sum_{i=1}^{k} r_{ij}\mathbb{Z} = r_{jj}\mathbb{Z}.$$

On peut faire la vérification par récurrence sur n.

Les questions qui se posent sont par exemple : quelle est la composition de l'urne, donnée par le vecteur Y_n au bout de n tirages ? Les couleurs se sont-elles mélangées ? Y a-t-il une couleur qui domine à la fin ? S'il y a une composition asymptotique, à quelle vitesse y a t-il convergence ?

2 Modélisation

Pour le premier modèle introduit par Pólya, la matrice de remplacement est $R = \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$. Dans le cas plus général considéré ici, la matrice de remplacement est $R = (r_{i,j})$. Nous ferons dans la suite l'hypothèse selon laquelle le nombre de boules rajoutées à chaque instant est constant, autrement dit que la somme des coefficients le long de chaque ligne de la matrice de remplacement est une constante S.

On appelle de telles urnes des urnes "balancées" ou des urnes "équilibrées" et la constante s'appelle la balance. Cette hypothèse permet de faire une étude du modèle comme chaîne de Markov en temps discret ; quand elle n'est pas satisfaite, d'autres méthodes sont utilisées comme le plongement en temps continu ([1, 5]) ou des méthodes analytiques ([4]). Pour simplifier les notations, nous supposerons que la balance est égale à 1, sauf dans l'exemple 3 où elle est nulle et où le traitement est très légèrement différent. Nous supposerons aussi que l'on démarre avec une boule de couleur 1 dans l'urne à l'instant 1. Du coup, le nombre de boules dans l'urne après le $(n-1)$-ème tirage est n (sinon ce serait nombre initial $+(n-1)$ fois la balance : $|Y_1|+(n-1)S$). Lorsque l'on modifie la condition initiale, les raisonnements qui suivent ne changent pas, mais les expressions obtenues dépendent de cette condition initiale.

Notre objet d'étude est donc le vecteur $Y_n = \begin{pmatrix} Y_n^{(1)} \\ \dots \\ Y_n^{(k)} \end{pmatrix}$ à valeurs dans $\mathbb{R}^k$. Son évolution est donnée par les règles ci-dessus qui se traduisent pour $j = 1, \dots k$ par :

$$Y_{n+1}^{(j)} = Y_n^{(j)} + \sum_{i=1}^{k} \mathbb{I}_{\{\text{tirage } i\}} r_{i,j},$$

c'est-à-dire que sur l'événement $\{\text{tirage } i\}$ qui est de probabilité $\frac{Y_n^{(i)}}{n}$, on a

$$Y_{n+1} = Y_n + \begin{pmatrix} r_{i,1} \\ \vdots \\ r_{i,k} \end{pmatrix}.$$

Ce dernier vecteur est la ligne i de la matrice R, c'est aussi la colonne i de la matrice $A := {}^tR$, on l'appelle Δ_i. Finalement on écrit

$$Y_{n+1} = Y_n + \Delta(n)$$

où les accroissements $\Delta(n)$ suivent la loi : pour tout $j = 1, 2, \dots, k$,

$$\mathbb{P}(\Delta(n) = \Delta_j / Y_n) = \frac{Y_n^{(j)}}{n} \tag{1}$$

où Δ_j est la j-ème colonne de la matrice $A = {}^tR$. Ainsi la suite (Y_n) est une marche aléatoire dont l'état initial est $Y_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ et dont les incréments suivent la loi donnée par (1) qui est une expression <u>linéaire</u> par rapport au vecteur Y_n. C'est ce qui va permettre une étude spectrale du processus. Par exemple, le passage de l'étape n à l'étape $n+1$ est donné par

$$\begin{aligned} \mathbb{E}^{\mathcal{F}_n}(Y_{n+1}) &= Y_n + \sum_{j=1}^{k} \frac{Y_n^{(j)}}{n} \Delta_j \\ &= (I + \frac{A}{n}) Y_n \end{aligned}$$

où A est l'endomorphisme (on désigne indifféremment par la même lettre un endomorphisme et sa matrice dans la base canonique) qui à un vecteur x de $\mathbb{R}^k$ associe $\sum_{j=1}^k x_j \Delta_j$. C'est bien le même $A = {}^tR$ que plus haut.

3 Etude en moyenne

En itérant l'expression ci-dessus, on trouve

$$\mathbb{E}Y_n = \gamma_n(A)Y_1 \tag{2}$$

où

$$\gamma_n(A) = \prod_{k=1}^{n-1}(I + \frac{A}{k})$$

est un polynôme en A. Supposons ici que la matrice A est diagonalisable. Quand elle ne l'est pas, elle se décompose suivant sa forme de Jordan et apparaissent des endomorphismes nilpotents introduisant des termes en $\log n$ dans l'asymptotique. L'étude est plus compliquée et peut être trouvée dans [9]. Nous en dirons un peu plus dans la section suivante.

Revenons au cas diagonalisable et appelons $Sp(A)$ l'ensemble des valeurs propres de A ; pour chaque valeur propre λ, appelons π_λ la projection sur le sous-espace propre associé à la valeur propre λ, parallèlement à la somme des autres espaces propres. Alors

$$A = \sum_{\lambda \in Sp(A)} \lambda \pi_\lambda$$

et par conséquent

$$\gamma_n(A) = \sum_{\lambda \in Sp(A)} \gamma_n(\lambda)\pi_\lambda.$$

Ainsi

$$\mathbb{E}Y_n = \gamma_n(A)Y_1 = \sum_{\lambda \in Sp(A)} \gamma_n(\lambda)\pi_\lambda Y_1.$$

Comme l'asymptotique des $\gamma_n(\lambda)$ est donnée par

$$\gamma_n(\lambda) = \prod_{k=1}^{n-1}(1 + \frac{\lambda}{k}) = \frac{n^\lambda}{\Gamma(1+\lambda)}(1 + O(\frac{1}{n})), \tag{3}$$

ceci entraîne que l'asymptotique de $\mathbb{E}Y_n$ sera donnée par la valeur propre de A de plus grande partie réelle.

Lemme 1 *La valeur propre de A de plus grande partie réelle est $\lambda_1 = 1$ et toutes les autres valeurs propres sont telles que $Re(\lambda) < 1$.*

Preuve

Comme la somme des coefficients d'une ligne de R est égale à 1, cela entraîne que ${}^t(1, 1, \ldots, 1)$ est vecteur propre de R pour la valeur propre 1. Comme A et $R = {}^tA$ ont mêmes valeurs propres, c'est que 1 est valeur propre de A.

Plaçons-nous dans un premier temps dans l'ensemble $\mathcal{M}$ des matrices à k lignes et k colonnes, à coefficients positifs ou nuls et dont la somme des lignes est égale à 1. Cet ensemble est stable par la multiplication des matrices et il est compact (pour la topologie des normes). Donc pour tout entier n, A^n est encore dans ce compact donc est bornée. Donc pour toute valeur propre λ de A, λ^n est bornée. Donc $|\lambda| \leq 1$, toutes les valeurs propres sont donc dans le disque de rayon 1 et 1 est la seule de partie réelle égale à 1, les autres sont de partie réelle strictement inférieure à 1.

Si les coefficients de la matrice de remplacement ne sont pas tous positifs, on s'y ramène en choisissant un réel $a > 0$ tel que $\frac{A+a}{1+a} \in \mathcal{M}$ (car seuls les termes diagonaux peuvent être négatifs). $\square$

L'application de ce lemme donne donc $\mathbb{E}Y_n \sim n\pi_1 Y_1$, autrement dit

Théorème 1 *Quand n tend vers $+\infty$,*

$$\frac{\mathbb{E}Y_n}{n} \longrightarrow \pi_1 Y_1.$$

En fait cette limite a aussi lieu presque sûrement, pas seulement en espérance, mais c'est un peu plus long à obtenir. Les mêmes méthodes fournissent aussi un développement asymptotique plus précis de Y_n, du type convergence de $Y_n - n\pi_1 Y_1$ renormalisé par une puissance de n (voir [2]).

4 Asymptotique de Y_n ; approche algébrique

4.1 Décomposition spectrale de A. Petites et grandes urnes

On effectue la réduction de Jordan de A dans $\mathbb{C}^k$: il existe une base de Jordan de vecteurs $(v_i)_{i=1,\dots,k}$ ou de formes linéaires $(u_i)_{i=1,\dots,k}$, parfois appelés vecteurs à gauche. On a déjà vu que $(1,\dots,1)$ est propre pour la valeur propre $\lambda = 1$. On appelle $u_1 = (1,\dots,1)$ et $u_i, i \geq 2$ les autres. La matrice A est réduite en blocs de Jordan et on définit

$$\sigma_2 = 1 \text{ si } 1 \text{ est valeur propre multiple}$$

$$\sigma_2 = \max\{Re(\lambda), \lambda \in Sp(A), \lambda \neq 1\} \text{ si } 1 \text{ est valeur propre simple.}$$

On a déjà vu que si 1 est valeur propre simple, $\sigma_2 < 1$.

Définition 1 *On dit que l'urne est grande ou que le processus est grand quand $1/2 < \sigma_2 \leq 1$. On dit que l'urne est petite ou que le processus est petit quand $\sigma_2 \leq 1/2$.*

On va s'intéresser essentiellement aux grandes urnes. Les blocs de Jordan les plus importants pour l'asymptotique de Y_n sont ceux pour lesquels $Re(\lambda) = \sigma_2$ et qui sont de taille maximale, disons $1 + \nu$. On les appelle blocs principaux. Ils conduisent à la classification suivante :

- on dit qu'une urne est réelle quand $Re(\lambda) = \sigma_2 \Longrightarrow \lambda$ réel.
- on dit qu'une urne est imaginaire quand $\exists \lambda \notin \mathbb{R}, \mathrm{Re}(\lambda) = \sigma_2$.
- on dit qu'une urne est principalement semi-simple si les blocs principaux sont de taille 1, c'est-à-dire correspondent à des valeurs propres simples. Sinon, on dit que l'urne est non principalement semi-simple. En abrégé pss et non pss.

4.2 Réduction de l'opérateur Φ. Polynômes réduits

On introduit l'opérateur Φ de transition du processus (Y_n) : pour tout $x \in \mathbb{R}^k$, pour toute fonction $f : \mathbb{C}^k \to E$ espace vectoriel,

$$\Phi(f)(x) = \sum_{i=1}^{k} x_i (f(x + \Delta_i) - f(x)).$$

Cet opérateur se décompose sur une base de polynômes appelés polynômes réduits, polynômes à k variables, notés $(Q_\alpha)_{\alpha \in \mathbb{N}^k}$ que l'on va définir plus bas. Ces polynômes vont permettre de calculer les moments de la loi limite de Y_n.

Notations dans $\mathbb{N}^k$:

On appelle $\delta_i = (0, \dots, 1, \dots, 0)$ le i-ème vecteur de la base canonique de $\mathbb{Z}^k$, de sorte que tout α de $\mathbb{N}^k$ s'écrit $\alpha = (\alpha_1, \dots, \alpha_k) = \sum_{i=1}^k \alpha_i \delta_i$. On note $|\alpha| = \sum_{i=1}^k \alpha_i$.

Rappelons que $(u_1, \dots, u_k)$ est une base de Jordan de formes linéaires et que $\lambda_1, \dots, \lambda_k$ sont les valeurs propres, et que la valeur propre λ_i est associée à u_i. On note $\lambda := (\lambda_1, \dots, \lambda_k)$ et on adopte les notations suivantes :

$$u^\alpha := \prod_{1 \le i \le k} u_i^{\alpha_i}$$

$$< \alpha, \lambda > = \sum_{i=1}^k \alpha_i \lambda_i.$$

Enfin on définit sur $\mathbb{N}^k$ l'ordre degré-alphabétique inverse : $\alpha < \beta$ si et seulement si : $|\alpha| < |\beta|$ ou bien

$$|\alpha| = |\beta| \text{ et } \exists i \in \{1, \dots, k\}, \alpha_i < \beta_i \text{ et } \alpha_j = \beta_j, \forall j > i.$$

Par exemple, $(1,0,0) < (0,1,0) < (0,0,1) < (2,0,0) < (1,1,0) < (1,0,1) < (0,2,0) < (0,1,1) < (0,0,2) < (3,0,0) < (2,1,0) < \dots$

Proposition 1 *Il existe une base de polynômes à k variables notés $(Q_\alpha)_{\alpha \in \mathbb{N}^k}$, appelés polynômes réduits par rapport à la base $(u_1, \dots, u_k)$ qui vérifient :*

(i) $Q_0 = 1$; $Q_{\delta_i} = u_i$

(ii) $\forall \alpha, Q_\alpha - u^\alpha \in Vect\{Q_\beta, \beta < \alpha, < \beta, \lambda > \neq < \alpha, \lambda >\}$

(iii) $\Phi(Q_\alpha) - < \alpha, \lambda > Q_\alpha \in Vect\{Q_\beta, \beta < \alpha, < \beta, \lambda > = < \alpha, \lambda >\}$

Les Q_α se calculent récursivement et le calcul peut être fait avec Maple.

Théorème 2 *Soit une grande urne. Soit $(u_1, \dots, u_k)$ une base de Jordan de formes linéaires et $(v_1, \dots, v_k)$ la base duale de Jordan de vecteurs. Soient $J_1, \dots, J_r$ les r blocs de Jordan principaux c'est-à-dire tels que $Re(\lambda(J_i)) = \sigma_2$ et de taille maximale $1 + \nu$, qui correspondent aux formes linéaires $u_{m_1}, \dots, u_{M_1}, \dots, u_{m_r}, \dots, u_{M_r}$. Alors*

(i) $\frac{u_{m_i}(Y_n)}{n^{\lambda(J_i)}}$ est une martingale qui converge presque sûrement et dans tous les L^p vers une variable aléatoire W_i, $i = 1, \dots, r$.

(ii) $Y_n - nv_1 = \frac{1}{\nu!}(\log n)^\nu \sum_{i=1}^r n^{\lambda(J_i)} W_i v_{M_i} + o(n^{\sigma_2}(\log n)^\nu)$

(iii) (moments) $\forall \alpha_1, \dots, \alpha_r$, soit $\alpha = \sum_{i=1}^r \alpha_i \delta_{m_i}$,

$$\mathbb{E}\left(\prod_{i=1}^r W_i^{\alpha_i}\right) = \frac{1}{\Gamma(1 + < \alpha, \lambda >)} Q_\alpha(Y_1).$$

Cas principalement semi-simple

Dans ce cas, $Re(\lambda_i) = \sigma_2$ entraîne u_i est une forme linéaire propre et v_i un vecteur propre. Autrement dit, la matrice de Jordan de A est composée de termes diagonaux, sans nilpotents, correspondant aux blocs principaux. Ceux-ci sont de taille 1, égaux à $\lambda_2, \dots, \lambda_r$. Notons que λ_2 peut être égal à 1 (revoir la définition de σ_2) et que les formes linéaires associées à 1 sont propres.

La proposition ci-dessus fournit une base de polynômes à k variables $(Q_\alpha)_{\alpha \in \mathbb{N}^k}$, que l'on peut calculer récursivement et implémenter sur Maple. L'opérateur Φ agit assez simplement sur les Q_α, il les multiplie grosso modo par $< \alpha, \lambda >$, car les Q_α sont des vecteurs caractéristiques de Φ, associés à la valeur propre $< \alpha, \lambda >$. Les moments du processus Y_n sont calculés en fonction des polynômes réduits et l'on arrive à démontrer que

$$\mathbb{E}(u^\alpha(Y_n)) = \frac{n^{<\alpha,\lambda>}}{\Gamma(1 + < \alpha, \lambda >)} Q_\alpha(Y_1) + o(n^{Re(<\alpha,\lambda>)}).$$

Dans ce cas principalement semi-simple, le théorème 2 devient :

Théorème 3 *Soit une grande urne, principalement semi-simple. Soit $(u_1, \ldots, u_k)$ une base de Jordan de formes linéaires telle que $u_2, \ldots, u_r$ sont les formes linéaires propres associées aux valeurs propres $\lambda_2, \ldots, \lambda_r$ telles que $Re(\lambda_i) = \sigma_2$ pour $2 \leq i \leq r$. Alors il existe des variables aléatoires $W_2, \ldots, W_r$ telles que :*

(i) $\frac{u_i(Y_n)}{n^{\lambda_i}}$ *est une martingale qui converge presque sûrement et dans tous les L^p vers W_i, $i = 2, \ldots, r$.*

(ii) $Y_n = nv_1 + \sum_{i=2}^{r} n^{\lambda_i} W_i v_i + o(n^{\sigma_2})$

(iii) (moments) $\forall \alpha_2, \ldots, \alpha_r$, *soit* $\alpha = \sum_{i=2}^{r} \alpha_i \delta_i = (0, \alpha_2, \ldots, \alpha_r, \ldots, 0)$,

$$\mathbb{E}\Big(\prod_{i=2}^{r} W_i^{\alpha_i}\Big) = \frac{1}{\Gamma(1+ < \alpha, \lambda >)} Q_\alpha(Y_1).$$

5 Exemples

De très nombreux exemples se trouvent dans [9, 10].

5.1 Exemple 1 : arbres m-aires de recherche

Comme cela a été remarqué depuis longtemps, par Mahmoud notamment, l'état Y_n des arbres de recherche m-aires (m est un entier fixé, $m \geq 2$), est une urne de Pólya pourvu que l'on considère que les boules sont les places disponibles dans chacun des nœuds de l'arbre et que les couleurs sont les types de nœuds ; on dit qu'un nœud est de type $i, i = 1, 2, \ldots, m$, quand il contient $i-1$ clés et i places (gaps). On convient de ne pas s'occuper des nœuds saturés c'est-à-dire de type m, car une relation déterministe lie les nœuds entre eux : la somme des clés vaut n à l'instant n.

Habituellement, le vecteur d'étude est X_n, il appartient à $\mathbb{R}^{m-1}$, ses coordonnées $X_n^{(i)}$ représentent le nombre de nœuds de type i dans l'arbre à n clés. Dans l'analogie ci-dessus, on a donc : $Y_n^{(i)} = iX_n^{(i)}$. En termes de matrices, si R est la matrice de remplacement dans le modèle de Pólya des Y (des places) on a

$$R = \begin{pmatrix} -1 & 2 & & & & & \\ & -2 & 3 & & & & \\ & & -3 & 4 & & & \\ & & & \ddots & & \ddots & \\ & & & & -(m-2) & & m-1 \\ m & & & & & & -(m-1) \end{pmatrix}$$

(quand rien n'est écrit le coefficient est nul). Cette matrice est semblable à la matrice M habituellement associée aux arbres m-aires de recherche, en faisant un changement de base qui consiste à passer du comptage des nœuds au comptage des places, de matrice de passage P

$$P = \begin{pmatrix} 1 & & & \\ & 2 & & \\ & & \ddots & \\ & & & m-1 \end{pmatrix}.$$

Lemme 2 *La matrice habituelle (cf. [7]) des arbres m-aires de recherche est*

$$M := \begin{pmatrix} -1 & & & & & m(m-1) \\ 1 & -2 & & & & \\ & 2 & -3 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & -(m-2) & \\ & & & & m-2 & -(m-1) \end{pmatrix} = P^{-1}\ {}^t(R)\ P.$$

Appliquons les résultats de la section précédente : comme la matrice A a $m-1$ racines simples distinctes, c'est que l'urne est principalement semi-simple. De plus, 1 est racine simple et il existe λ_2 et $\bar{\lambda_2}$ (conjuguée de λ_2) telles que $Re(\lambda_2) = \sigma_2$, donc l'urne est imaginaire. Appelons $\tau_2 = Im(\lambda_2)$. La classification en petites et grandes urnes donne :

$$\sigma_2 \leq \frac{1}{2} \Longleftrightarrow m \leq 26 \Longleftrightarrow \text{ c'est une petite urne}$$

$$\sigma_2 > \frac{1}{2} \Longleftrightarrow m \geq 27 \Longleftrightarrow \text{ c'est une grande urne}$$

pour laquelle le théorème 3 s'applique et donc :

$$Y_n = nv_1 + n^{\lambda_2} W_2 v_2 + n^{\bar{\lambda_2}} W_3 v_3 + o(n^{\sigma_2})$$

qui peut encore s'écrire (cf. [2])

$$Y_n = nv_1 + n^{\sigma_2} \rho [C \cos(\tau_2 \log n + \phi) + S \sin(\tau_2 \log n + \phi)] + o(n^{\sigma_2}),$$

où v_1, C et S se calculent explicitement, $\rho = |W_2|$ et $\phi = Arg(W_2) \in [0, 2\pi[$ sont aléatoires et on peut calculer leurs moments, l'expression se trouve dans [10].

5.2 Exemple 2 : arbres 2-3 ou arbres équilibrés

Les arbres 2-3 sont des arbres toujours équilibrés c'est-à-dire qu'à chaque étape, les feuilles de l'arbre sont toutes au même niveau. Les nœuds ont tous soit deux soit trois descendants. Il y a deux types de nœuds, les nœuds de type 2 ont deux descendants et les nœuds de type 3 ont trois descendants. Il y a donc aussi deux types de feuilles : les feuilles de type 2 sont feuilles d'un nœud de type 2 et les feuilles de type 3 sont feuilles d'un nœud de type 3.

La règle d'insertion dans l'arbre est la suivante : on insère sur les feuilles ; un nœud de type 2 se transforme en un nœud de type 3 ; un nœud de type 3 se transforme en deux nœuds de type 2, ce qui provoque éventuellement d'autres transformations en amont vers la racine.

La règle de transformation est illustrée dans la figure 1 ci-dessous.

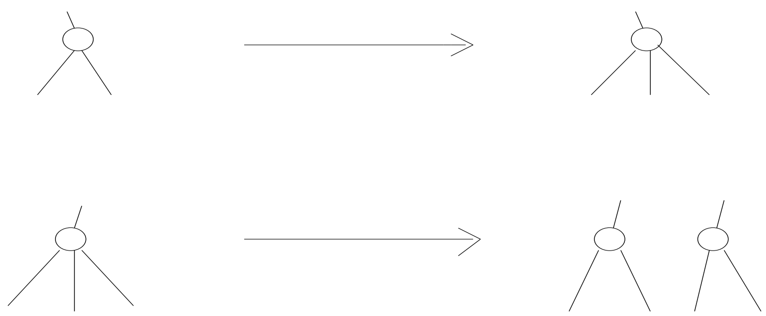

Figure 1: *règle d'insertion dans un arbre 2-3.*

Et la figure 2 montre un début possible de suite d'arbres 2-3.

Ces arbres sont difficiles à étudier, par exemple on ne sait pas encore prouver que l'asymptotique de leur hauteur est Cte $\log n$ avec la constante égale à 1. Plus facilement, on sait compter les feuilles de différents types : appelons Y_n le vecteur à deux coordonnées comptant les feuilles de type 2 et 3 à l'instant n. C'est une urne de Pólya dont la matrice de remplacement est

$$R = \begin{pmatrix} -2 & 3 \\ 4 & -3 \end{pmatrix}$$

C'est l'exemple type de l'article [4] qui utilise d'autres méthodes. L'urne est équilibrée, de balance égale à 1. La matrice $A = {}^tR$ est $A = \begin{pmatrix} -2 & 4 \\ 3 & -3 \end{pmatrix}$ ses valeurs propres sont 1 et -6

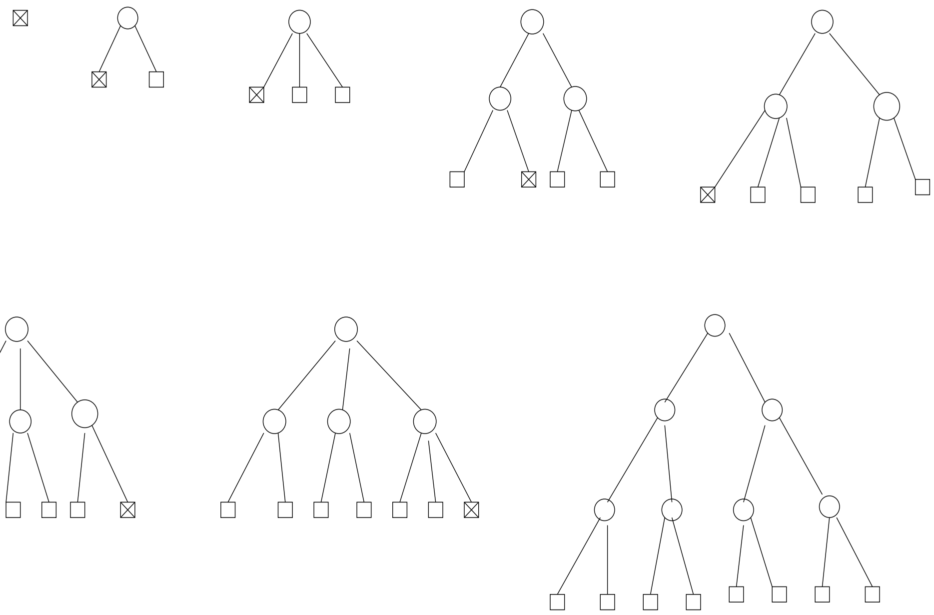

Figure 2: *exemple de début de suite d'arbres 2-3. La feuille sur laquelle a lieu l'insertion est marquée d'une croix.*

et d'après le théorème général, $\frac{\mathbb{E}Y_n}{n} \longrightarrow \pi_1 Y_1$. Dans cet exemple, on part de $Y_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ et on peut faire le calcul de $\pi_1 Y_1$ comme suit. Ce vecteur est d'une part sur la droite $Ker(A-I)$ de vecteurs propres invariants par A, ce qui donne une première équation entre les coordonnées de $\pi_1 Y_1$: $-3x+4y=0$; d'autre part, la somme des coordonnées de Y_n est égale à $n+1$ ce qui entraine que la somme des coordonnées de $\pi_1 Y_1$ est égale à 1 : $x+y=1$. On trouve $\frac{\mathbb{E}Y_n}{n} \longrightarrow \begin{pmatrix} 4/7 \\ 3/7 \end{pmatrix}$.

Remarque : dans la terminologie plus haut, il s'agit d'une petite urne et donc la section 4 ne s'applique pas. On peut montrer que le terme du second ordre dans le développement asymptotique converge en distribution vers une loi gaussienne par les méthodes de type Janson ([5]).

5.3 Exemple 3 : hachage, une urne de balance 0

Cet exemple est tiré de la thèse de Marianne Durand [3].

Dans une table de hachage, on répartit des éléments d'un certain ensemble $\mathcal{E}$ dans une table qui contient m cases numérotées de 0 à $m-1$ grâce à une fonction de hachage h. Le nombre de clés à répartir est N. Dans le hachage avec pagination, chaque case de la table a une capacité b qui est un entier fixé, $b \geq 1$.

Les collisions sont possibles (une clé est envoyée par h dans une case déjà pleine) et il y a plusieurs façons de les régler. Dans le "hachage à essai aléatoire" (random probing hashing), pour chaque élément $e \in \mathcal{E}$, $h(e)$ donne une suite de positions $\in \{0,1,\ldots,m-1\}$ possibles où e peut être inséré. Ces positions possibles sont tirées uniformément dans $\{0,1,\ldots,m-1\}$ avec répétition possible. Voici un exemple de fonction de hachage avec essai aléatoire pour ranger $N=8$ éléments dans une table à $m=4$ cases de capacité $b=2$.

Exemple

$h(a) = \{3, 0, 2, 1, \dots\}$
$h(b) = \{2, 0, 2, 1, \dots\}$
$h(c) = \{3, 0, 2, 2, \dots\}$
$h(d) = \{3, 1, 3, 1, \dots\}$
$h(e) = \{0, 2, 1, 2, \dots\}$
$h(f) = \{1, 1, 3, 1, \dots\}$
$h(g) = \{0, 2, 0, 1, \dots\}$
$h(h) = \{1, 1, 0, 3, 2, \dots\}$,

ce qui donne un placement dans la table comme suit (les lettres devenues majuscules représentent les collisions) :

$$\begin{array}{cccc} & H & & H \\ H & H & & D \\ & & & \\ g & f & h & c \\ e & d & b & a \end{array}$$

Les indicateurs recherchés dans ce problème sont notamment le nombre de collisions et l'état de la table qui est donné par un vecteur $x = (x_0, x_1, \dots, x_b)$ où pour tout $i = 0, 1, \dots, b$, x_i est le nombre de cases de la table contenant i éléments. Chaque case peut ainsi être dans l'un des états $0, 1, \dots, b$. L'ensemble est modélisé par une urne de Pólya dans laquelle les boules sont les cases, il y en a m, et les couleurs sont les états de chaque case, il y en a $b+1$. Le vecteur d'étude est X_n à valeurs dans $\mathbb{R}^{b+1}$ dont les coordonnées $X_n^{(i)}$ représentent le nombre de cases dans l'état i. Afin de tenir compte des collisions, il faut prendre comme unité de temps le nombre de *tentatives* d'insertions des clés dans la table et non le nombre de clés effectivement insérées. Ce dernier nombre est égal à $\sum_{i=1}^{b} iX_n^{(i)}$, ce qui donne par différence avec n le nombre de collisions.

L'étude de X_n se fait de manière analogue au cas général. A chaque tentative d'insertion, une case dans l'état i passe dans l'état $i+1$ si elle n'est pas saturée et donc le nombre de cases dans l'état i diminue de 1 tandis que le nombre de cases dans l'état $i+1$ augmente de 1. S'il y a collision, il ne se passe rien du point de vue du vecteur X_n. Ainsi la matrice de remplacement est

$$R = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots \\ 0 & -1 & 1 & 0 & \dots \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

C'est une urne de balance 0. Le nombre de boules est dans ce cas constant égal à m. Et la probabilité pour que ce soit une case de type j qui soit modifiée dans le passage de n à $n+1$ est égale à $\frac{X_n^{(j)}}{m}$. D'où :

$$\mathbb{E}^{\mathcal{F}_n}(X_{n+1}) = X_n + \sum_{j=0}^{b} \frac{X_n^{(j)}}{m} \Delta_j$$

où les vecteurs colonnes Δ_j sont les colonnes de la matrice A qui est la transposée de R. On obtient ainsi

$$\mathbb{E}^{\mathcal{F}_n}(X_{n+1}) = (I + \frac{A}{m})X_n$$

et en itérant

$$\mathbb{E}X_n = (I + \frac{A}{m})^n X_0$$

ce qui donne l'asymptotique en moyenne de X_n en calculant $(I + \frac{A}{m})^n$ qui est $0(1)$ (faire le calcul en mettant $(I + \frac{A}{m})$ sous forme de Jordan).

References

[1] K. B. Athreya, S. Karlin Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.* **39** (1968), 1801–1817.

[2] B. Chauvin, N. Pouyanne m-ary search trees when $m \geq 27$: a strong asymptotics for the space requirements. *Random Structures and Algorithms* **24** (2004), 133–154.

[3] M. Durand, Combinatoire analytique et algorithmique des ensembles de données. Thèse de l'Ecole Polytechnique (2004).

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

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

[6] Johnson , Kotz. Urn models and their applications. Wiley, (1977)

[7] H.M. Mahmoud Evolution of random search trees. Wiley, New-York, (1992)

[8] H.M. Mahmoud Urn models and connections to random trees: A review. *Journal of the Iranian Statistical Society,* **2**, 53–114.(2003)

[9] N. Pouyanne An algebraic approach of Pólya processes (2005). Soumis Ann.IHP Disponible à l'adresse http://www.math.uvsq.fr/∼pouyanne/

[10] N. Pouyanne Classification of large Pólya-Eggenberger urns with regard to their asymptotics. *Discrete Mathematics and Theoretical Computer Science,* **AD** (2005), 177–245.

[11] V. Puyhaubert Modèles d'urnes et phénomènes de seuils en combinatoire analytique. Thèse de l'Ecole Polytechnique (2005).