

Combinatoire analytique et modèles d'urnes

Basile Morcrette,
encadré par Philippe Flajolet

Projet Algorithms, INRIA Rocquencourt.

6 septembre 2010



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



INRIA

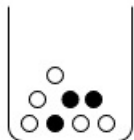
Des situations diverses et variées

- ▶ épidémies
- ▶ génétique des populations
- ▶ physique statistique : particules de gaz dans un compartiment
- ▶ **structures informatiques** : arbres, graphes

Plan

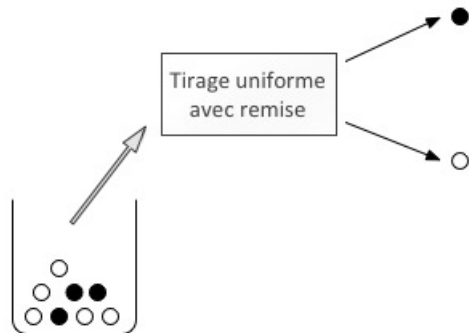
1. Modèles d'urnes
2. Urnes et branches d'arbres aléatoires
3. Urnes de croissance préférentielle
4. Urnes des k -arbres

Le modèle d'urne



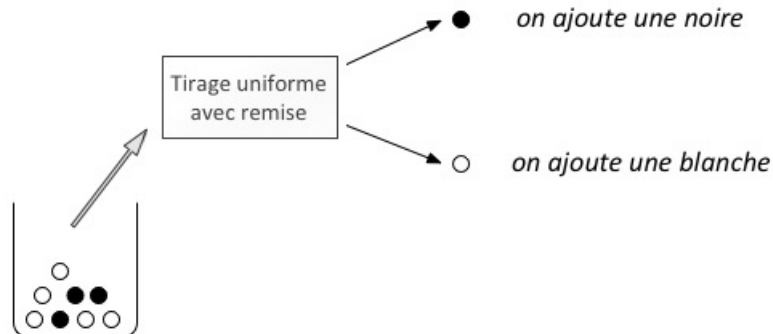
- ▶ une urne et des boules de couleurs différentes

Le modèle d'urne



- ▶ une urne et des boules de couleurs différentes

Le modèle d'urne



- ▶ une urne et des boules de couleurs différentes
- ▶ des règles fixées pour l'évolution de l'urne

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

Urne de Pólya

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \alpha, \delta \in \mathbb{Z}, \quad \beta, \gamma \in \mathbb{Z}_{\geq 0}.$$

Urne de Pólya

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \alpha, \delta \in \mathbb{Z}, \quad \beta, \gamma \in \mathbb{Z}_{\geq 0}.$$

Urne équilibrée : $\alpha + \beta = \gamma + \delta$ (nombre total de boules déterministe).

Urne de Pólya

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \alpha, \delta \in \mathbb{Z}, \quad \beta, \gamma \in \mathbb{Z}_{\geq 0}.$$

Urne équilibrée : $\alpha + \beta = \gamma + \delta$ (nombre total de boules déterministe).

Configuration initiale fixée (a_0, b_0) : a_0 boules \bullet (comptées par x)

b_0 boules \circ (comptées par y)

Urne de Pólya

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \quad \alpha, \delta \in \mathbb{Z}, \quad \beta, \gamma \in \mathbb{Z}_{\geq 0}.$$

Urne équilibrée : $\alpha + \beta = \gamma + \delta$ (nombre total de boules déterministe).

Configuration initiale fixée (a_0, b_0) : a_0 boules \bullet (comptées par x)

b_0 boules \circ (comptées par y)

Définition

Histoire de longueur n : une suite de n évolutions (n tirages).

$H_{n,a,b}$: nombre d'histoires de longueur n , débutant en (a_0, b_0) , et terminant en (a, b) .

Fonction génératrice des $H_{n,a,b}$:

$$H(x, y, z) = \sum_{n,a,b} H_{n,a,b} x^a y^b \frac{z^n}{n!}.$$

Compter les histoires

Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.

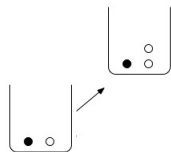


$$H(x, y, z) =$$

$$xyz$$

Compter les histoires

Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.

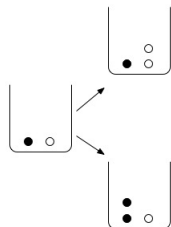


$$H(x, y, z) =$$

xyz

Compter les histoires

Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.



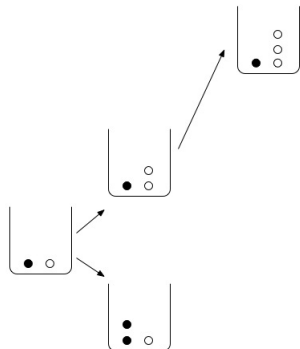
$$H(x, y, z) =$$

$$xyz$$

$$+ (xy^2 + x^2y) \frac{z^2}{2}$$

Compter les histoires

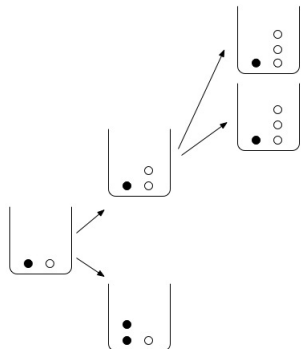
Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.



$$H(x, y, z) =$$
$$xyz$$
$$+ (xy^2 + x^2y)\frac{z^2}{2}$$

Compter les histoires

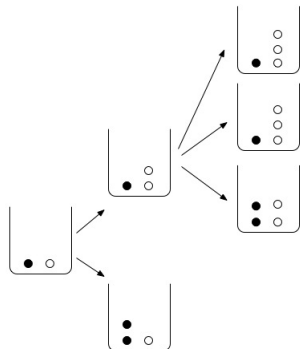
Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.



$$H(x, y, z) =$$
$$xyz$$
$$+ (xy^2 + x^2y)\frac{z^2}{2}$$

Compter les histoires

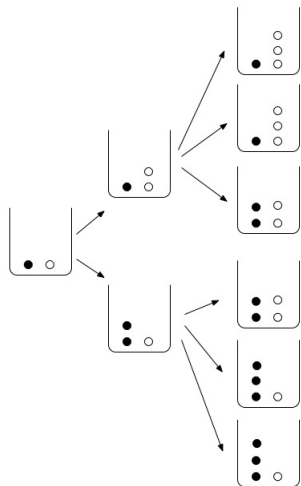
Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.



$$H(x, y, z) =$$
$$xyz$$
$$+ (xy^2 + x^2y)\frac{z^2}{2}$$

Compter les histoires

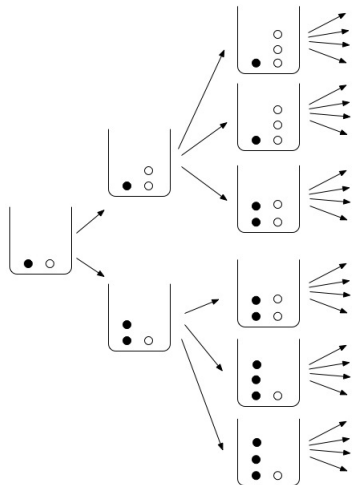
Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.



$$\begin{aligned} H(x, y, z) = & \\ & xyz \\ & + (xy^2 + x^2y) \frac{z^2}{2} \\ & + (2xy^3 + 2x^2y^2 + 2x^3y) \frac{z^3}{6} \end{aligned}$$

Compter les histoires

Prenons l'urne $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $(a_0, b_0) = (1, 1)$.



$$\begin{aligned}
 H(x, y, z) = & \\
 & xyz \\
 & + (xy^2 + x^2y) \frac{z^2}{2} \\
 & + (2xy^3 + 2x^2y^2 + 2x^3y) \frac{z^3}{6} \\
 & + \dots
 \end{aligned}$$

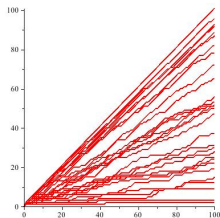
Des comportements très divers

Problème Comprendre la composition de l'urne après n étapes, lorsque n tend vers ∞ .

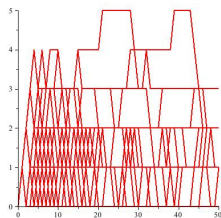
Des comportements très divers

Problème Comprendre la composition de l'urne après n étapes, lorsque n tend vers ∞ .

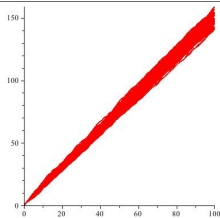
$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



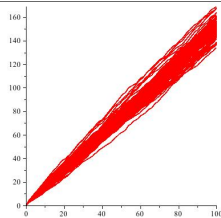
$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$



$$\begin{pmatrix} 3 & 1 \\ 2 & 2 \end{pmatrix}$$



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



Urnes équilibrées et système différentiel

Approche analytique Théorème [Flajolet–Dumas–Puyhaubert 2005]

$$\text{Urne } \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \text{ et } \begin{cases} (a_0, b_0) \\ \alpha + \beta = \gamma + \delta \end{cases} \implies \text{où } \begin{cases} \dot{X} = X^{\alpha+1} Y^{\beta} \\ \dot{Y} = X^{\gamma} Y^{\delta+1} \end{cases}$$

$H = X^{a_0} Y^{b_0}$

Urnes équilibrées et système différentiel

Approche analytique Théorème [Flajolet–Dumas–Puyhaubert 2005]

$$\text{Urne } \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \text{ et } \begin{cases} (a_0, b_0) \\ \alpha + \beta = \gamma + \delta \end{cases} \implies \text{où } \begin{cases} \dot{X} = X^{\alpha+1} Y^{\beta} \\ \dot{Y} = X^{\gamma} Y^{\delta+1} \end{cases}$$

$H = X^{a_0} Y^{b_0}$

Travaux réalisés :

- ▶ étude d'urnes en **grande dimension**
- ▶ étude d'urnes à **coefficients positifs** (= additives)

2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.

2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)

2. Urnes et branches d'arbre aléatoire

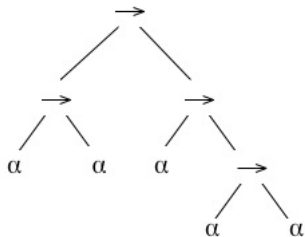
- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément

2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément
 - ▶ la remplacer par un noeud binaire et deux feuilles

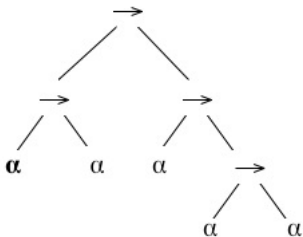
2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément
 - ▶ la remplacer par un noeud binaire et deux feuilles



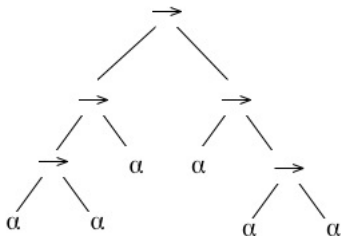
2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément
 - ▶ la remplacer par un noeud binaire et deux feuilles



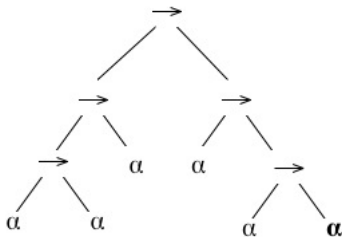
2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément
 - ▶ la remplacer par un noeud binaire et deux feuilles



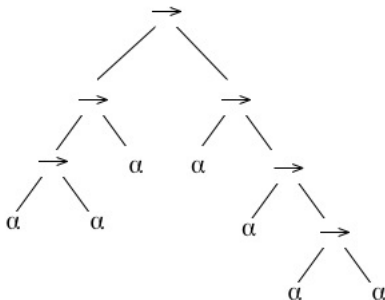
2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément
 - ▶ la remplacer par un noeud binaire et deux feuilles

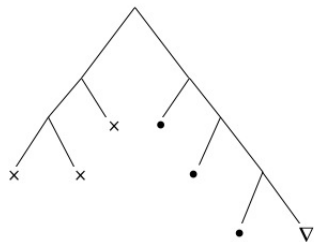


2. Urnes et branches d'arbre aléatoire

- ▶ **Motivation** : quantifier la proportion de tautologies parmi les formules logiques ayant comme seul connecteur l'implication.
- ▶ **Modèle probabiliste** : croissance uniforme aux feuilles (modèle ABR)
 - ▶ tirer une feuille au hasard uniformément
 - ▶ la remplacer par un noeud binaire et deux feuilles



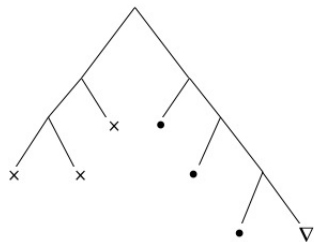
Modélisation par une urne 3×3



3 couleurs,
avec les règles :

▽	→	●▽
●	→	xx
x	→	xx

Modélisation par une urne 3×3



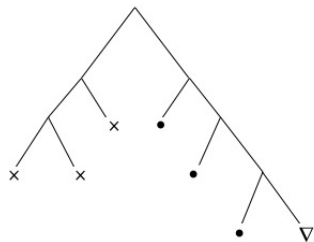
3 couleurs,
avec les règles :

$\nabla \rightarrow \bullet \nabla$
 $\bullet \rightarrow \times \times$
 $\times \rightarrow \times \times$

Urne
correspondante :

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

Modélisation par une urne 3×3



3 couleurs,
avec les règles :

$\nabla \rightarrow \bullet \nabla$
 $\bullet \rightarrow x x$
 $x \rightarrow x x$

Urne
correspondante :

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

Fonction génératrice des histoires

$$H(y, z) = \exp \left(\ln \left(\frac{1}{1-z} \right) + (y-1)z \right)$$

z compte la longueur de l'histoire,
 y compte le nombre de boules \bullet .

Loi de Poisson dans les sous-arbres

Soit $U_{k,n}$ le nombre de sous-arbres gauches de taille k directement accrochés à la branche droite d'un arbre aléatoire de taille n .

Théorème

- ▶ $U_{1,n}$ converge en loi, $U_{1,n} \xrightarrow[n \rightarrow \infty]{} U_1$,
- ▶ $U_1 \sim \mathcal{Poisson}(1)$, avec vitesse de convergence en $O\left(\frac{2^n}{n!}\right)$.

Loi de Poisson dans les sous-arbres

Soit $U_{k,n}$ le nombre de sous-arbres gauches de taille k directement accrochés à la branche droite d'un arbre aléatoire de taille n .

Théorème

- ▶ $U_{1,n}$ converge en loi, $U_{1,n} \xrightarrow[n \rightarrow \infty]{} U_1$,
- ▶ $U_1 \sim \mathcal{Poisson}(1)$, avec vitesse de convergence en $O\left(\frac{2^n}{n!}\right)$.

Généralisation Utilisation d'une urne $(k+2) \times (k+2)$

Théorème

- ▶ $U_{k,n}$ converge en loi, $U_{k,n} \xrightarrow[n \rightarrow \infty]{} U_k$,
- ▶ $U_k \sim \mathcal{Poisson}\left(\frac{1}{k}\right)$, avec vitesse de convergence en $O\left(\frac{(2k)^n}{n!}\right)$.

3. Urnes de croissance préférentielle

Motivation : Caractériser les urnes 2×2 **additives** (coefficients positifs).

Approche : Avoir une classe d'urnes avec des fonctions génératrices "exploitables".

3. Urnes de croissance préférentielle

Motivation : Caractériser les urnes 2×2 additives (coefficients positifs).

Approche : Avoir une classe d'urnes avec des fonctions génératrices "exploitables".

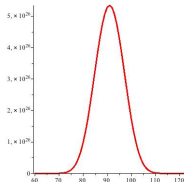
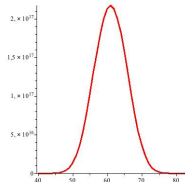
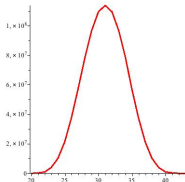
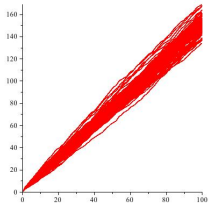
Théorème

La classe d'urnes équilibrées $\begin{pmatrix} 2\alpha & \beta \\ \alpha & \alpha + \beta \end{pmatrix}$, avec $\alpha > 0$, $\beta \geq 0$, a une fonction génératrice bivariée algébrique.

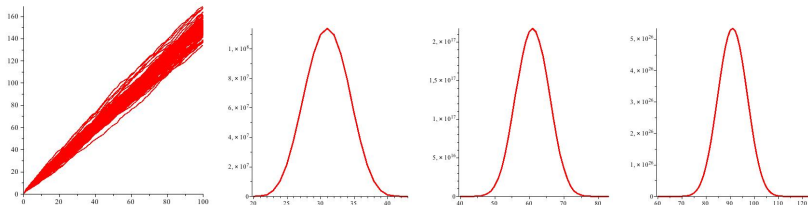
La fonction génératrice des histoires $H(x, z)$ est racine du polynôme en Y

$$(z - a - b(x)) Y^{2\alpha+\beta} + b(x) Y^\alpha + a$$

Simulations – Moyenne – Variance



Simulations – Moyenne – Variance



Proposition

Soit X_n la variable aléatoire comptant le nombre de boules de couleur x dans l'urne au temps n . Alors

$$\mathbb{E}(X_n) = \frac{\alpha(2\alpha + \beta)}{\alpha + \beta}n + \frac{\alpha}{\alpha + \beta} \frac{\Gamma\left(\frac{1}{2\alpha + \beta}\right)}{\Gamma\left(\frac{\alpha + 1}{2\alpha + \beta}\right)} n^{\frac{\alpha}{2\alpha + \beta}} + \frac{\alpha}{\alpha + \beta} + O\left(n^{\frac{\alpha}{2\alpha + \beta} - 1}\right),$$

$$\mathbb{V}(X_n) = \frac{\alpha^3(2\alpha + \beta)}{(\alpha + \beta)^2}n + O\left(n^{\frac{\alpha + \beta}{2\alpha + \beta}}\right) \quad \text{où } \Gamma(x) := \int_0^{\infty} t^{x-1} e^{-t} dt.$$

Exploitation analytique

1. L'équation de la fonction génératrice présente deux singularités, lorsque x varie autour de 1.
2. Il y a confluence lorsque x tend vers 1. Cela donne des développements asymptotiques différents selon le voisinage que l'on considère.

Exploitation analytique

1. L'équation de la fonction génératrice présente deux singularités, lorsque x varie autour de 1.
2. Il y a confluence lorsque x tend vers 1. Cela donne des développements asymptotiques différents selon le voisinage que l'on considère.

Démarche :

- ▶ utilisation d'un contour normalisé (grâce à la moyenne et la variance), autour des singularités. Addition des deux contributions.
- ▶ changement de variable à la Lagrange, et méthode de col.

Exploitation analytique

1. L'équation de la fonction génératrice présente deux singularités, lorsque x varie autour de 1.
2. Il y a confluence lorsque x tend vers 1. Cela donne des développements asymptotiques différents selon le voisinage que l'on considère.

Démarche :

- ▶ utilisation d'un contour normalisé (grâce à la moyenne et la variance), autour des singularités. Addition des deux contributions.
- ▶ changement de variable à la Lagrange, et méthode de col.

But : obtenir une convergence vers une loi normale.

4. Urnes des k -arbres

Motivation : modélisation de réseaux sociaux [Panholzer–Seitz 2010]

Définition

Un k -arbre T est

- ▶ soit une k -clique
- ▶ soit il existe un sommet f ayant une k -clique comme voisin et $T \setminus f$ est un k -arbre

Ordonné : fils distinguables.

Croissant : sommets numérotés dans l'ordre d'apparition.

4. Urnes des k -arbres

Motivation : modélisation de réseaux sociaux [Panholzer–Seitz 2010]

Définition

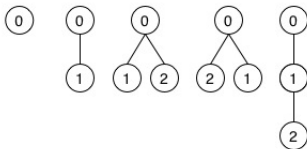
Un k -arbre T est

- ▶ soit une k -clique
- ▶ soit il existe un sommet f ayant une k -clique comme voisin et $T \setminus f$ est un k -arbre

Ordonné : fils distinguables.

Croissant : sommets numérotés dans l'ordre d'apparition.

1-arbres ordonnés croissants (dits PORT)



4. Urnes des k -arbres

Motivation : modélisation de réseaux sociaux [Panholzer–Seitz 2010]

Définition

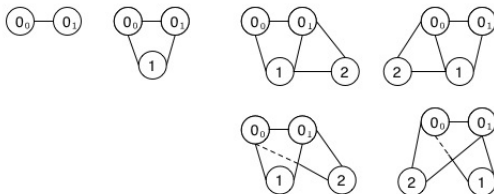
Un k -arbre T est

- ▶ soit une k -clique
- ▶ soit il existe un sommet f ayant une k -clique comme voisin et $T \setminus f$ est un k -arbre

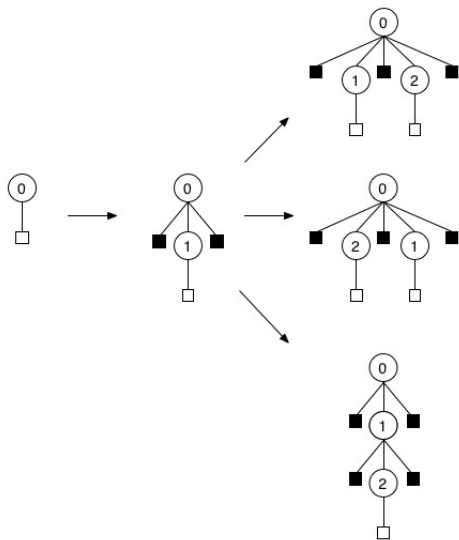
Ordonné : fils distinguables.

Croissant : sommets numérotés dans l'ordre d'apparition.

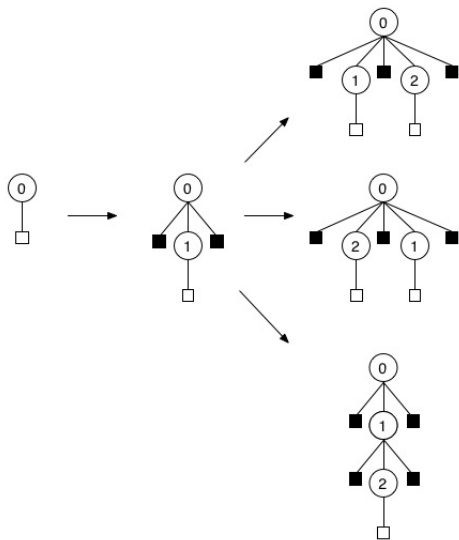
2-arbres ordonnés croissants



Modélisation



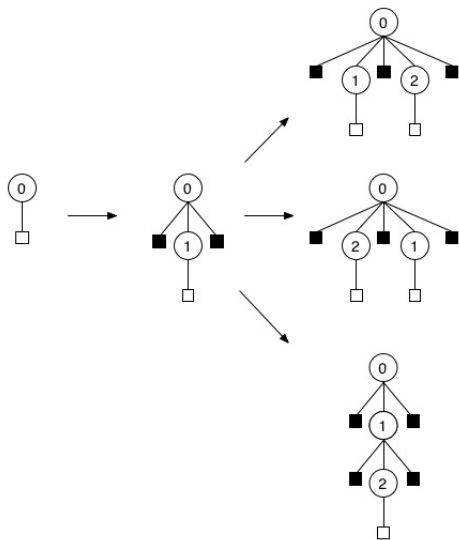
Modélisation



Pour les 1-arbres

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

Modélisation



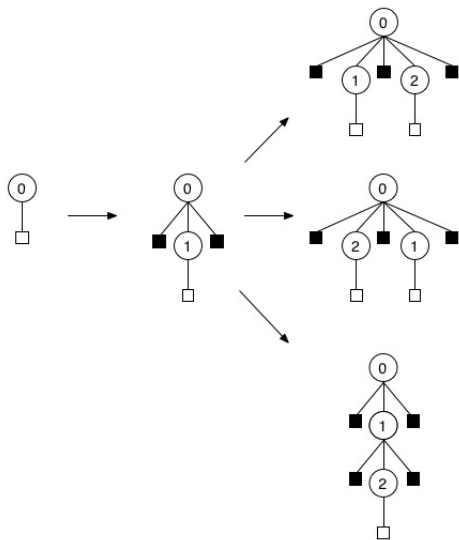
Pour les 1-arbres

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

Pour les k -arbres

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

Modélisation



Pour les 1-arbres

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

Pour les k -arbres

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

$$\frac{\frac{d}{dz} H(x, z)}{H(x, z)^k (H(x, z) - x + 1)^2} = 1$$

Loi normale limite

Soit X_n la variable aléatoire comptant le nombre de boules \square dans l'urne U_1 après n étapes.

Cas $k = 1$: nombre de feuilles dans un PORT de taille n .

Théorème

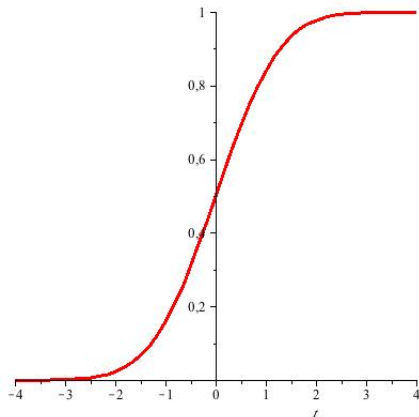
- ▶ X_n a une moyenne μ_n et une variance σ_n^2
- ▶ $\frac{X_n - (2/3)n}{\sqrt{n}/3}$ converge en loi vers $\mathcal{N}(0, 1)$,
avec vitesse de convergence $O\left(\frac{1}{\sqrt{n}}\right)$.

$$\mu_n = \mathbb{E}(X_n) = \frac{2}{3}(n+1/2) + O\left(\frac{1}{n}\right), \quad \sigma_n^2 = \mathbb{V}(X_n) = \frac{1}{9}(n+1/2) + O\left(\frac{1}{n}\right),$$

$$\mathbb{P}\left\{\frac{X_n - \frac{2}{3}n}{\sqrt{\frac{n}{9}}} \leq t\right\} = \Phi(t) + O\left(\frac{1}{\sqrt{n}}\right), \quad \text{où } \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{v^2}{2}} dv.$$

Loi normale limite – Convergence de la fonction de répartition

$$\mathbb{P} \left\{ \frac{X_n - \frac{2}{3}n}{\sqrt{\frac{n}{9}}} \leq t \right\} \xrightarrow{n \rightarrow \infty}$$



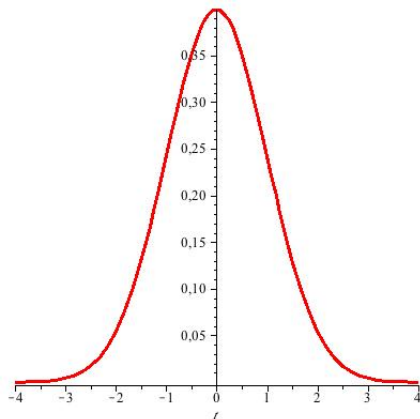
Théorème

Notons $p_{n,k} = \mathbb{P}\{X_n = k\}$. La distribution des X_n satisfait une **loi locale limite** de type **gaussienne** avec vitesse de convergence $O\left(\frac{1}{\sqrt{n}}\right)$, i.e.

$$\sup_{t \in \mathbb{R}} \left| \frac{\sqrt{n}}{3} p_{n, \lfloor 2n/3 + t\sqrt{n}/3 \rfloor} - \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \right| \leq \frac{1}{\sqrt{n}}.$$

Loi Locale Limite – Convergence vers la fonction de densité

$$\frac{\sqrt{n}}{3} \mathbb{P} \left\{ X_n = \left\lfloor \frac{2n}{3} + t \frac{\sqrt{n}}{3} \right\rfloor \right\} \xrightarrow{n \rightarrow \infty}$$

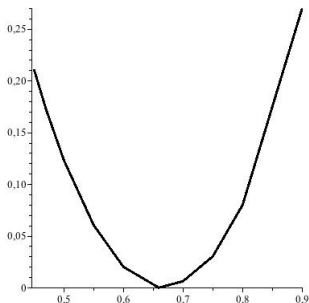


Grandes déviations

- ▶ borne exponentiellement petite sur la grande déviation par rapport à la valeur moyenne : quantification sur les **événements rares**

Théorème

- ▶ si $0.42 < t < 2/3$, $\mathbb{P}(X_n \leq tn) \approx e^{-nW(t)}$ (queue gauche)
- ▶ si $2/3 < t < 0.73$, $\mathbb{P}(X_n \geq tn) \approx e^{-nW(t)}$ (queue droite)



Conclusion et Perspectives

Divers problèmes modélisés par les urnes et traités par la combinatoire analytique :

- ▶ Urnes en grande dimension
 - ▶ premiers pas avec la combinatoire analytique
 - ▶ variété de problèmes modélisés
- ▶ Urnes bicolores additives
 - ▶ classe algébrique à deux paramètres : **générale**
 - ▶ urne des k -arbres : **premiers résultats** sur les urnes additives avec Loi centrale limite, Loi locale limite et principe de grandes déviations, avec vitesse de convergence.