Arbres aléatoires et analyse de singularités

Christian Costermans

Directeurs de mémoire :

Anne Philippe

Hoang Ngoc Minh

Université Lille 1 - Laboratoire Painlevé

Université Lille 2 - Centre de Bio-informatique (C.I.B.)

Contexte

- Travaux de l'équipe de bio-informatique, Lille 2
- Etude des puces à ADN

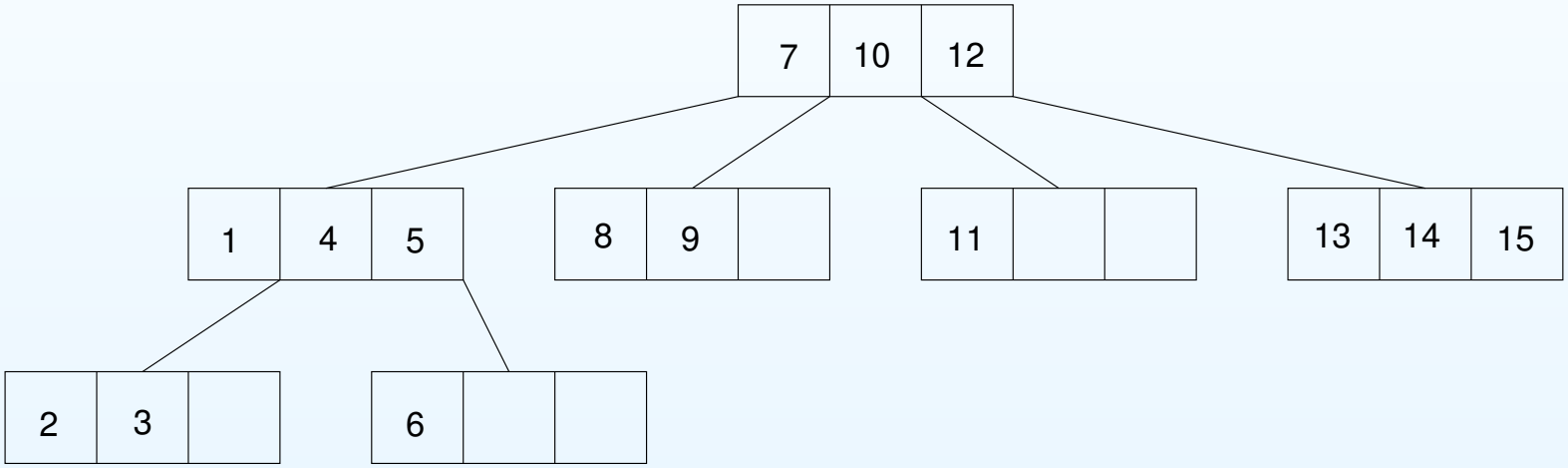
  Une puce = 1920 gènes sur-, sous- ou non-exprimés, représentés par $+, -$ ou 0.

  $\longrightarrow$ Stockage sous forme d'arbre pondérés, ternaires.

  $\longrightarrow$ Détermination d'une distribution, puis simulation.
- Problème: coût algorithmique pour la détermination d'un plus long préfixe commun ?

Arbre de recherche m-aires

- Structure de données importante en informatique
- Généralise la notion d'arbre binaire de recherche



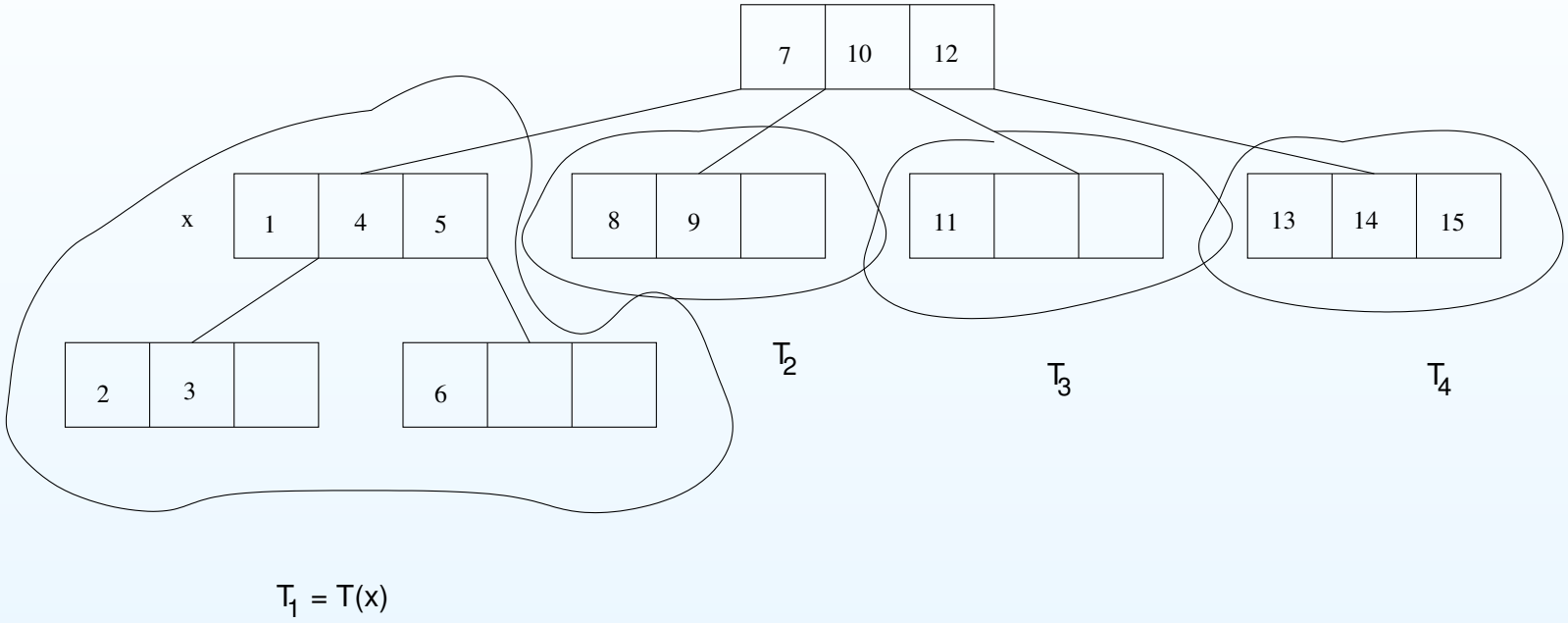
Ici, l'arbre est construit à partir de la suite
$(7, 10, 12, 1, 4, 5, 8, 9, 11, 13, 2, 6, 3, 15, 14)$

Modèles de probabilité

- Arbres de Catalan : tous les arbres m-aires de recherche à n clés sont équiprobables

- Modèle de permutation aléatoire : toutes les permutations de $\{1, 2, \ldots, n\}$ sont équiprobables

Notations



On note $|T| = 15$, et $T_1, T_2 \ldots$ les sous-arbres de T.

Fonctionnelles additives sur des arbres aléatoires

$$f(T) = \sum_{i=1}^{m} f(T_i) + c_{|T|}$$

$(c_n)_{n \geq m-1}$: “suite-test” (dépendant de l’algorithme étudié)

- Exemples:
  1. $c_n = 1$: espace requis (Chern & Hwang, 2001)
  2. $c_n = n - (m-1)$: analyse du Quicksort

Fonctionnelle additive $\Longleftrightarrow$ complexité d’un algorithme récursif

Objectifs

Pour des algorithmes récursifs sur des arbres aléatoires

- Etudier les relations entre le coût de l'appel récursif et la complexité de l'algorithme
- Trouver des équivalents asymptotiques pour cette complexité
- Obtenir des distributions limites de cette complexité

Outils:

- Analyse de singularités (analyse complexe)
- "Théorèmes de transfert" reliant le développement asymptotique de la fonction-test avec celui de la fonctionnelle

Complexité moyenne : relation de récurrence

- Rappel :
$$f(T) = \sum_{i=1}^{m} f(T_i) + c_{|T|}$$
- On pose : $f_n = \mathbb{E}(f(T))$ pour $|T| = n$
- Donc, avec T_i désignant le i-ème sous-arbre, on a pour tous i et j
$$\mathbb{P}(|T_i| = j) = \binom{n-j-1}{m-2} \Big/ \binom{n}{m-1}$$
- Coût moyen de complexité associé
$$\mathbb{E}(f(T)) = f_n = m \sum_{j=0}^{n-(m-1)} \mathbb{P}(|T_i| = j) f_j + c_n$$
$$= \frac{m}{\binom{n}{m-1}} \sum_{j=0}^{n-(m-1)} \binom{n-j-1}{m-2} f_j + c_n$$

Complexité moyenne : solution de la récurrence

- Pour simplifier, on considère le cas $c_1 = \cdots = c_{m-2} = 0$
- On obtient alors, en posant
  $A(z) = \sum_{n\geq 0} f_n z^n$ et $B(z) = \sum_{n\geq 0} c_n z^n$:

$$A(z) = \sum_{j=1}^{m-1} \frac{(1-z)^{-\lambda_j}}{\psi'(\lambda_j)} \int_0^z B^{(m-1)}(\zeta)(1-\zeta)^{\lambda_j+m-2} d\zeta$$

- Le **développement asymptotique** de la fonction génératrice $B(z)$ de la suite (c_n) **se transfère** en un **développement asymptotique** de $A(z)$, fonction génératrice de la suite des espérances.
- On utilise **l'analyse de singularités** [Flajolet & Odlyzko, 1990] pour obtenir un équivalent asymptotique de la moyenne.

Complexité asymptotique - Analyse de singularités

- Permet la détermination de l'ordre de grandeur asymptotique des coefficients de Taylor de fonctions analytiques

$$\begin{aligned} f(z) = O(g(z)) &\Rightarrow f_n = O(g_n) \\ f(z) = o(g(z)) &\Rightarrow f_n = o(g_n) \\ f(z) \sim g(z) &\Rightarrow f_n \sim g_n. \end{aligned}$$

- Schéma du raisonnement général

$$\begin{aligned} (c_n) &\Longrightarrow (f_n) \\ B(z) = \sum_{n\geq 0} c_n z^n &\Longrightarrow A(z) = \sum_{n\geq 0} f_n z^n \\ Asympt(B(z)) &\Longrightarrow Asympt(A(z)) \\ Asympt(A(z)) &\Longrightarrow Asympt(f_n) \end{aligned}$$

Complexité asymptotique - Analyse de singularités

Rappel :

$$A(z) = \sum_{j=1}^{m-1} \frac{(1-z)^{-\lambda_j}}{\psi'(\lambda_j)} \int_0^z B^{(m-1)}(\zeta)(1-\zeta)^{\lambda_j+m-2} d\zeta$$

où $A(z) = \sum_{n\geq 0} f_n z^n$ et $B(z) = \sum_{n\geq 0} c_n z^n$.

Nous nous plaçons sous l'hypothèse que les résultats d'analyse de singularités sont

- (1) stables par intégration
- (2) stables par dérivation

  Par exemple, $f(z) = O((1-z)^\beta)$ dans un domaine Δ implique $f'(z) = O((1-z)^{\beta-1})$ dans un domaine Δ

Complexité asymptotique - Arbre binaire de recherche

Soit $X_n = f(T)$ pour $|T| = n$; on a la relation

$$X_n = X_{K_n} + X_{n-1-K_n} + c_n, \qquad \text{avec } X_n \text{ et } K_n \text{ indépendantes}$$

Or $P(K_n = k) = \frac{1}{n}$,
donc

$$f_n = \mathbb{E}(X_n) = c_n + \frac{2}{n}\sum_{k=0}^{n-1} f_k$$

d'où

$$A(z) = B(z) + 2\int_0^z A(w)\frac{dw}{1-w}$$

et

$$A(z) = (1-z)^{-2}\int_0^z B'(w)(1-w)^2 dw$$

Complexité asymptotique - Arbre binaire de recherche

Donc, pour passer de $B(z)$ à $A(z)$:

$$B(z) \xrightarrow{\frac{d}{dz}} \ldots \xrightarrow{\times(1-z)^2} \ldots \xrightarrow{\int} \ldots \xrightarrow{\times(1-z)^{-2}} A(z)$$

En particulier, si $c_n = n^\alpha (\alpha > 1)$, alors $B(z) \sim \Gamma(\alpha+1)(1-z)^{-\alpha-1}$
Ainsi,

$$\Gamma(\alpha+1)(1-z)^{-\alpha-1} \xrightarrow{\frac{d}{dz}} \Gamma(\alpha+1)(\alpha+1)(1-z)^{-\alpha-2} \xrightarrow{\times(1-z)^2} \Gamma(\alpha+1)(\alpha+1)(1-z)^{-\alpha}$$

$$\xrightarrow{\int} \Gamma(\alpha+1)\frac{-\alpha-1}{-\alpha+1}(1-z)^{-\alpha+1} \xrightarrow{\times(1-z)^{-2}} A(z) \sim \Gamma(\alpha+1)\frac{\alpha+1}{\alpha-1}(1-z)^{-\alpha-1}$$

Par conséquent,

$$f_n \sim \Gamma(\alpha+1)\frac{\alpha+1}{\alpha-1}\frac{n^\alpha}{\Gamma(\alpha+1)} = \frac{\alpha+1}{\alpha-1}n^\alpha$$

Convergence en loi - Moments d'ordre supérieur

Pour k $\in \mathbb{N}$ fixé, le moment d'ordre k, $\mu_n(k) = \mathbb{E}(f(T)^k)$ vérifie une relation de récurrence de la forme :

$$\mu_n(k) = \frac{m}{\binom{n}{m-1}} \sum_{j=0}^{n-(m-1)} \binom{n-j-1}{m-2} \mu_j(k) + r_n(k)$$

avec $r_n(k)$ définie à l'aide de c_n et des moments $\mu_n(p)$ pour $p < k$ et on a encore

$$A_k(z) = \sum_{j=1}^{m-1} \frac{(1-z)^{-\lambda_j}}{\psi'(\lambda_j)} \int_0^z B_k^{(m-1)}(\zeta)(1-\zeta)^{\lambda_j+m-2} d\zeta$$

en posant $A_k(z) = \sum_{n\geq 0} \mu_n(k) z^n$ et $B_k(z) = \sum_{n\geq 0} r_n(k) z^n$

Convergence en loi - Principes

- Centrer la fonctionnelle en utilisant l'équivalent asymptotique de la moyenne
- Constater que tous les moments de la fonctionnelle centrée vérifient une relation de récurrence de la même forme
- Utiliser un théorème de transfert pour obtenir les développements asymptotiques de la fonctionnelle centrée
- En déduire la convergence en loi en utilisant la méthode des moments (si possible)

Convergence en loi - Théorème de transfert général

- Relie le comportement de la suite $(\mu_n(k))_{n\geq 0}$ à la suite $(r_n(k))_{n\geq 0}$

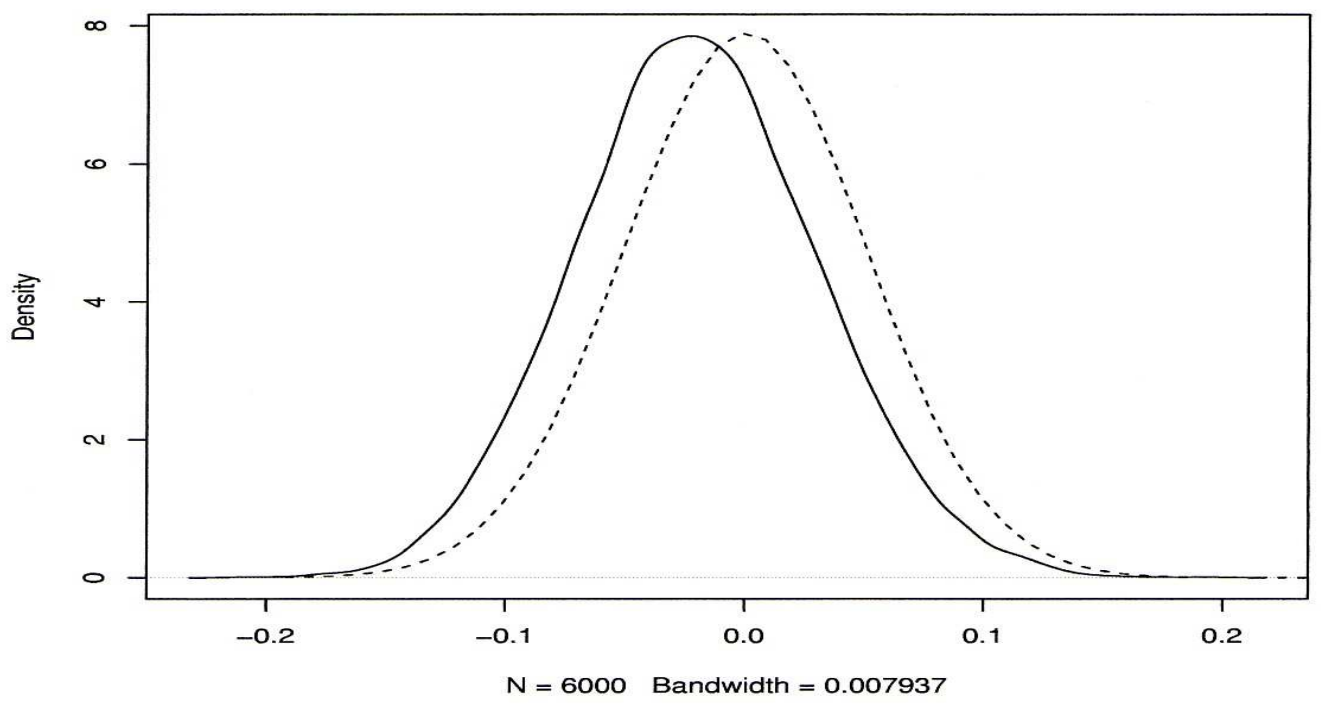
Entrée $(r_n(k))$	Sortie $(\mu_n(k))$
$r_n(k) = o(n)$	$\mu_n(k) \sim \frac{K_1}{H_m - 1} n$
$r_n(k) \sim Kn$	$\mu_n(k) \sim \frac{K}{H_m - 1} n \log n$
$r_n(k) \sim Kn^v, \quad v > 1$	$\mu_n(k) \sim \frac{K}{1 - \frac{m!\Gamma(v+1)}{\Gamma(v+m)}} n^v$

- Rappel : $r_n(1) = c_n$ et $\mu_n(1) = f_n$

Convergence en loi - Distributions asymptotiques

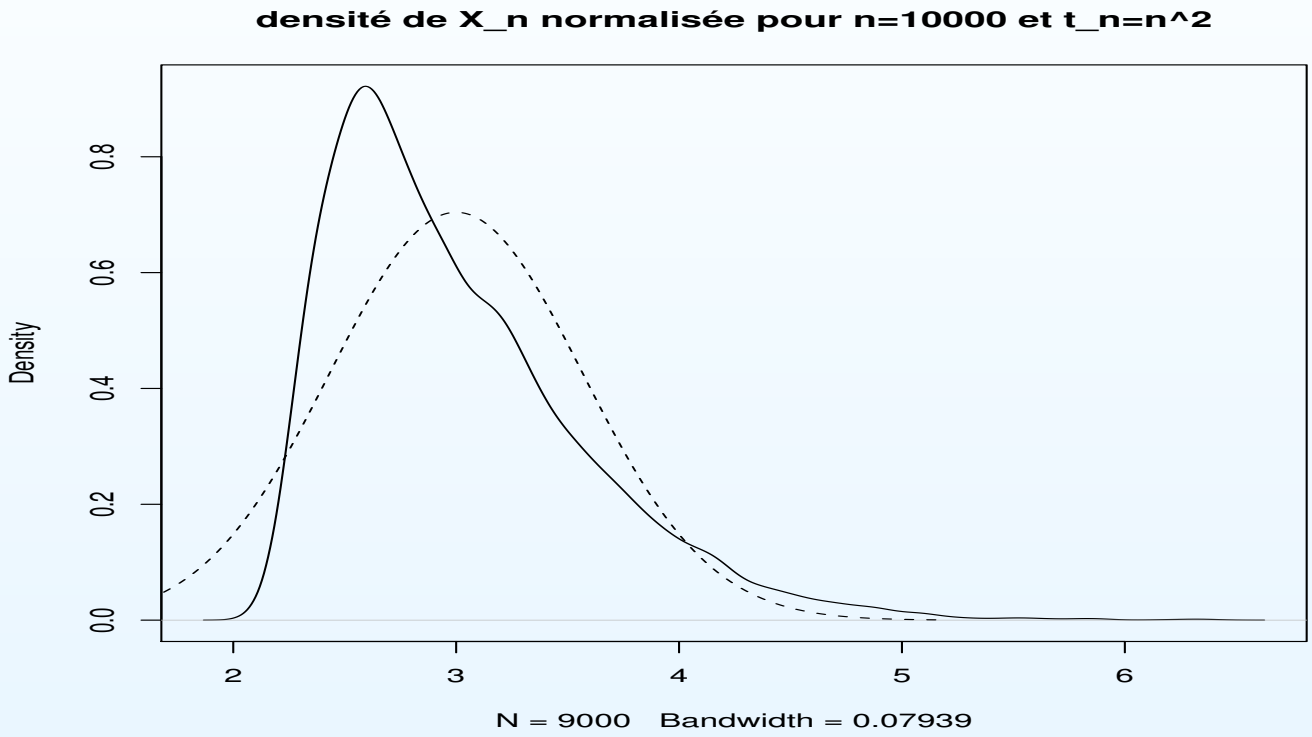
- suite-test “faible”: $c_n = O(n^{1/2} L(n))$:
  - Normalité asymptotique si $m \leq 26$
  - Pas de distribution limite (en général) si $m \geq 27$
- suite-test “modérée”: $c_n \sim n^\gamma L(n), \quad \frac{1}{2} < \gamma < 1$:
  - Convergence vers une distribution non gaussienne si $m \leq m_0$ (où $m_0 \geq 26$)
  - Pas de distribution limite (en général) si $m \geq m_0 + 1$
- suite-test “élevée”: $c_n \sim n^\gamma L(n)$ avec $\gamma > 1$:
  - Convergence, pour tout m, vers une distribution non gaussienne

Illustration : distribution de X_n pour m=2 et $t_n = n^{0.1}$



- Convergence vers une loi normale
- Ecart des moyennes - on retrouve le développement asymptotique : $\frac{1}{\sqrt{n}} \frac{0.1+1}{0.1-1} n^{0.1} \approx -0.02$

Distribution de X_n pour m=2 et $t_n = n^2$



Convergence vers une loi non-gaussienne