Hauteur des arbres de Lyndon

A. Briquet, L. Mercier

Institut Elie Cartan

Aléa, 20 février 2014

Définitions

Définition

Un mot de Lyndon est un mot plus petit (pour l'ordre lexicographique) que tous ses suffixes propres.

Exemple : aab est plus petit que ab et b : c'est un mot de Lyndon.
aba est plus grand que a : ce n'est pas un mot de Lyndon.

Propriété

Tout mot de Lyndon peut se décomposer en deux mots de Lyndon. Cette décomposition n'est pas unique. La décomposition qui a le facteur de gauche le plus court est appelée la décomposition standard.

Exemple : abc peut se décomposer en $a \mid bc$ ou $ab \mid c$. La décomposition standard est $a \mid bc$.

Génération aléatoire des mots de Lyndon

On associe à chaque lettre i de l'alphabet (à m lettres) une probabilité p_i.
On génère un mot de n lettres où chaque lettre suit une loi $p = (p_i)$ i.i.d., et on utilise le mot pour générer un mot de Lyndon de la manière suivante :
Exemple :
$220012 \longrightarrow 200122 \longrightarrow 001222 \longrightarrow 012220 \longrightarrow 122200 \longrightarrow 222001$
On a généré le mot 220012. On en tire le mot de Lyndon 001222.

Propriété

Si les m lettres de l'alphabet sont équiprobables, alors le mot obtenu avec l'algorithme ci-dessus est uniforme sur l'ensemble des mots de Lyndon de longueur n à m lettres.

On note

$$p^* = \max_{i \neq 0}(p_i)$$

Définitions

Définition

Un arbre de Lyndon est obtenu à partir d'un mot de Lyndon, en le décomposant récursivement (décomposition standard).

Exemple

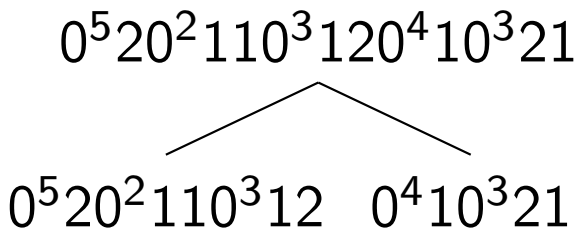
$$0^5 20^2 110^3 120^4 10^3 21$$

Définitions

Définition

Un arbre de Lyndon est obtenu à partir d'un mot de Lyndon, en le décomposant récursivement (décomposition standard).

Exemple

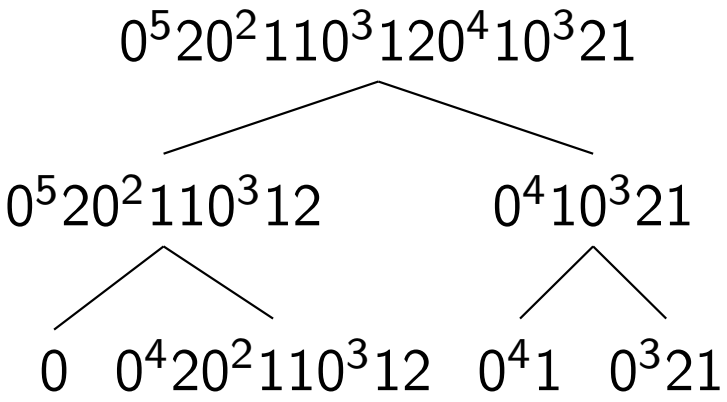


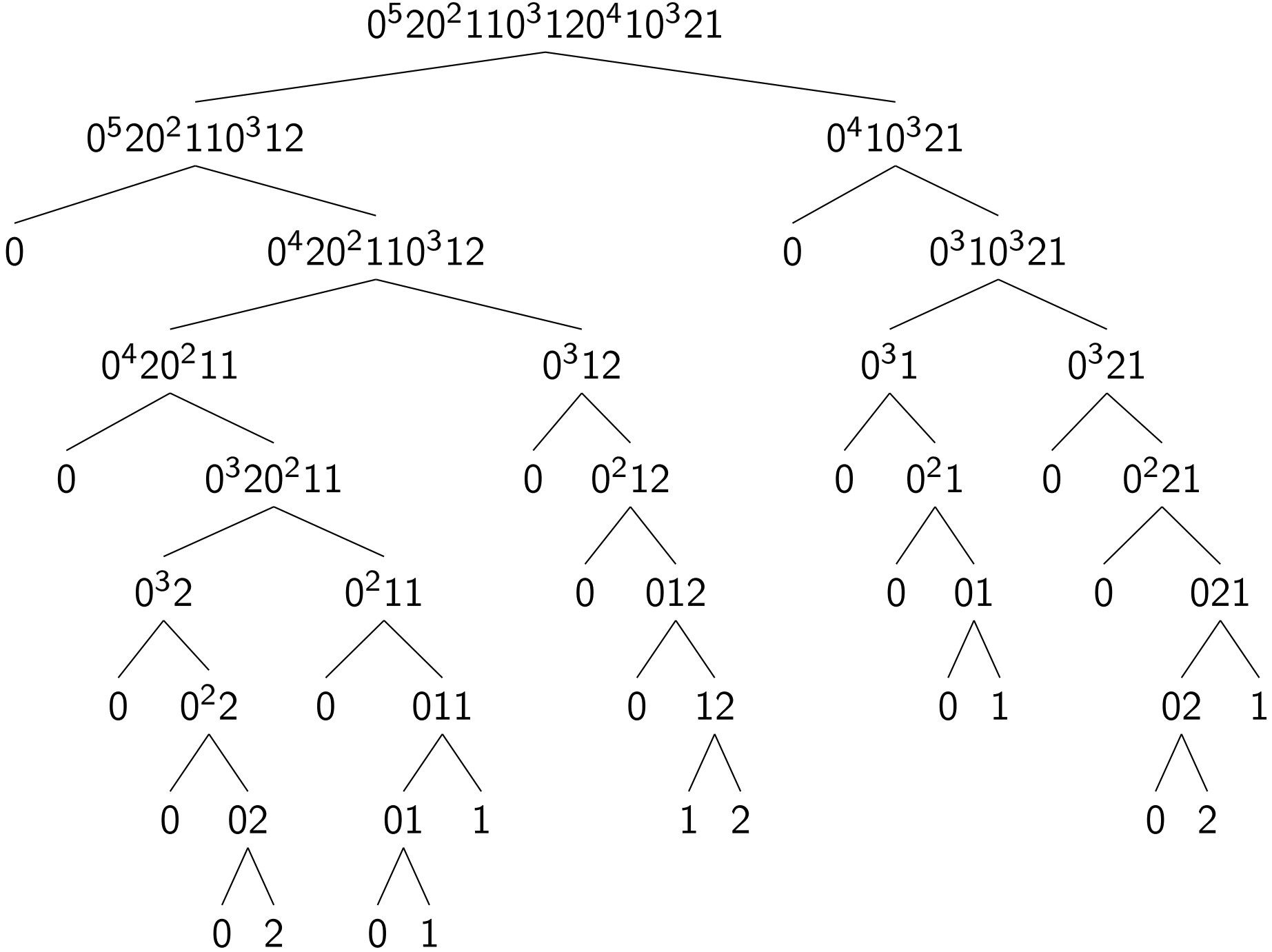
Définitions

Définition

Un arbre de Lyndon est obtenu à partir d'un mot de Lyndon, en le décomposant récursivement (décomposition standard).

Exemple




$0^5 2 0^2 1 1 0^3 1 2 0^4 1 0^3 2 1$
$0^5 2 0^2 1 1 0^3 1 2$
$0^4 1 0^3 2 1$
0
$0^4 2 0^2 1 1 0^3 1 2$
0
$0^3 1 0^3 2 1$
$0^4 2 0^2 1 1$
$0^3 1 2$
$0^3 1$
$0^3 2 1$
0
$0^3 2 0^2 1 1$
0
$0^2 1 2$
0
$0^2 1$
0
$0^2 2 1$
$0^3 2$
$0^2 1 1$
0
012
0
01
0
021
0
$0^2 2$
0
011
0
12
0
1
02
1
0
02
01
1
1
2
0
2
0
2
0
1

Résultats

On note
$\Psi(\lambda,\rho,\gamma) = \log(p_0(-\log(p_0))^{\rho+\lambda}(e/\lambda)^{\lambda}(1+\rho/\gamma)^{\gamma}(1+\gamma/\rho)^{\rho}) + \Xi(\rho,\gamma)$,
où $\Xi(\gamma,\rho) = \lim 1/n \log(\mathbb{P}(\frac{1-\gamma}{\rho}n \leq \sum_{i=1}^{n} U_i \leq \frac{1-\gamma}{\rho}n + 1))$ et

$$\Delta^*(p_0,p^*) = \sup_{\Psi(\lambda,\rho,\gamma)>0} \{\lambda+\rho+\gamma - \frac{\Psi(\lambda,\rho,\gamma)}{\log(p^*)}\}/\log(1/p_0).$$

Théorème

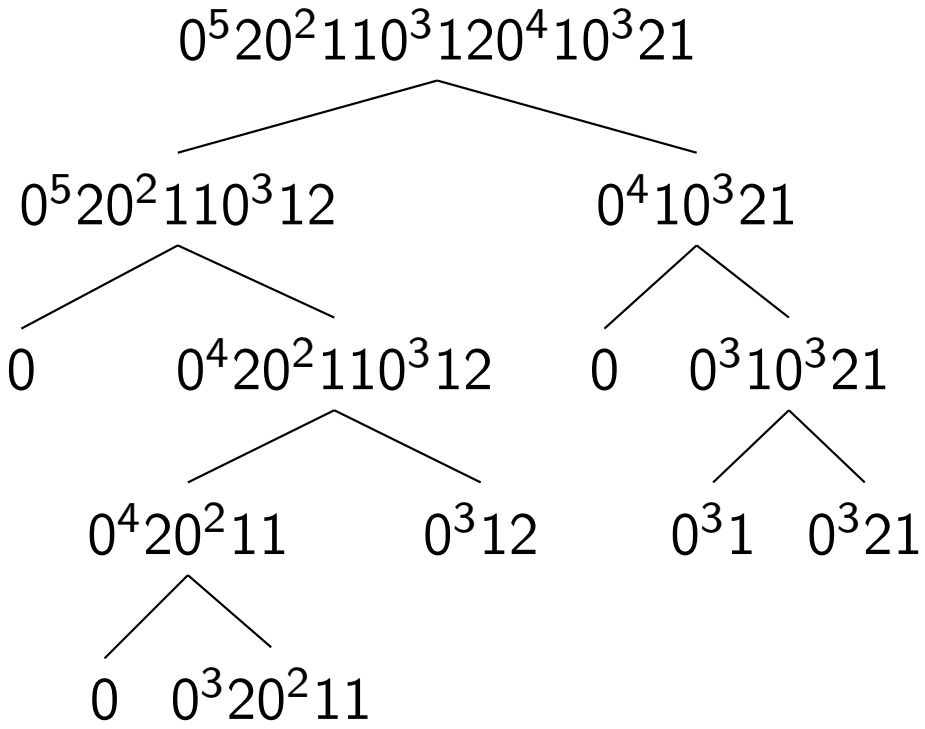
La hauteur H_n de l'arbre de Lyndon à n feuilles converge en probabilité vers $\Delta^*(p_0,p^*)\ln n$.

On note $\Delta_*(p_0) = \inf_{\Psi(\lambda,\rho,\gamma)>0}\{\lambda+\rho+\gamma\}/\log(1/p_0)$.

Conjecture

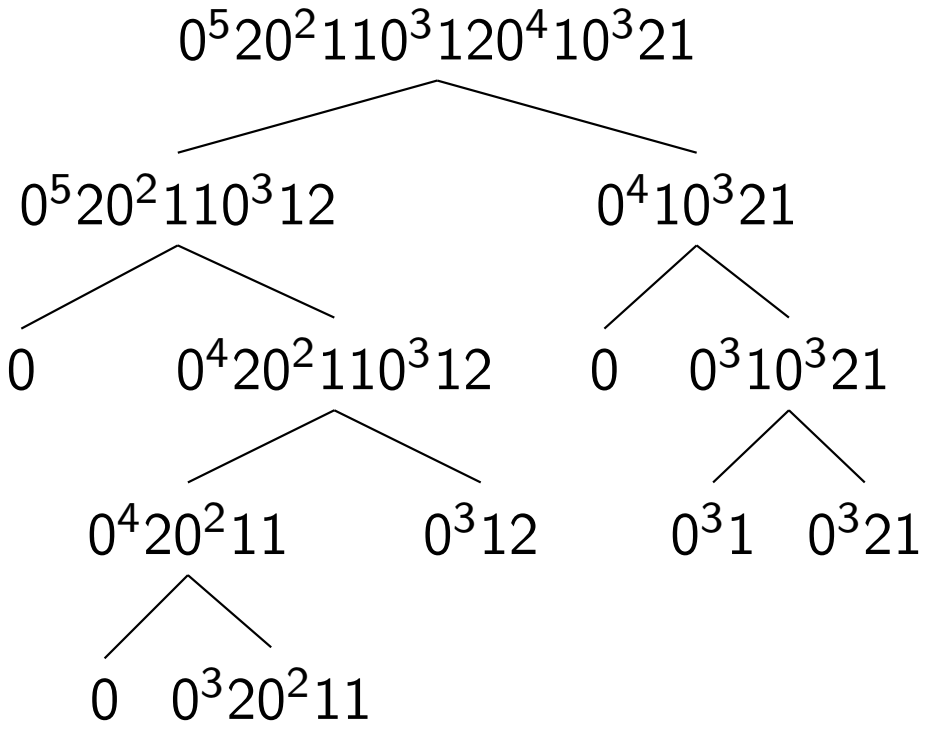
La hauteur de la feuille étiquetée par un 1 (ou toute autre lettre que 0) la plus basse dans l'arbre de Lyndon à n feuilles converge en probabilité vers $\Delta_*(p_0)\ln n$.

Arbre de Lyndon seuillé



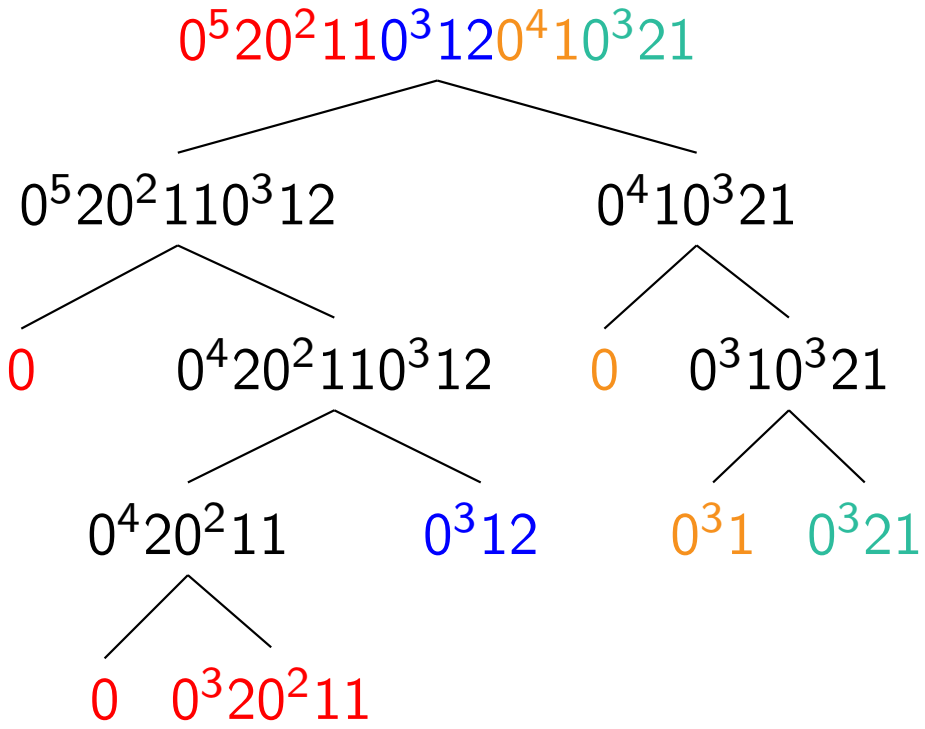
On arrête la décomposition quand chaque feuille contient au plus un facteur $0^i \rightarrow$ le seuil vaut i.
Dans l'exemple, on s'arrête à $0^3 \rightarrow$ seuil $= 3$.

Arbre de Lyndon seuillé



On appelle les feuilles qui contiennent juste un seul 0 des aiguilles, et les autres des palmes.
Les sous-arbres correspondant aux palmes sont appelés les buissons.

Arbre de Lyndon seuillé



On appelle les feuilles qui contiennent juste un seul 0 des aiguilles, et les autres des palmes.
Les sous-arbres correspondant aux palmes sont appelés les buissons.

Distribution sur les réels

On veut passer d'un mot w (pas forcément de Lyndon) à un réel $x \in [0,1]$.
Pour cela, on lit le mot en base m.
Exemple : avec un alphabet à 3 lettres,
mot 0102 $\longrightarrow$ réel 0,0102 en base 3 $\longrightarrow 3^{-2} + 2 * 3^{-4}$ en base 10.

On regarde un réel X créé à partir d'un mot infini avec des lettres $\sim (p_i)$ i.i.d.

Propriété

Si (p_i) est équiprobable, le réel obtenu est uniforme sur [0,1].

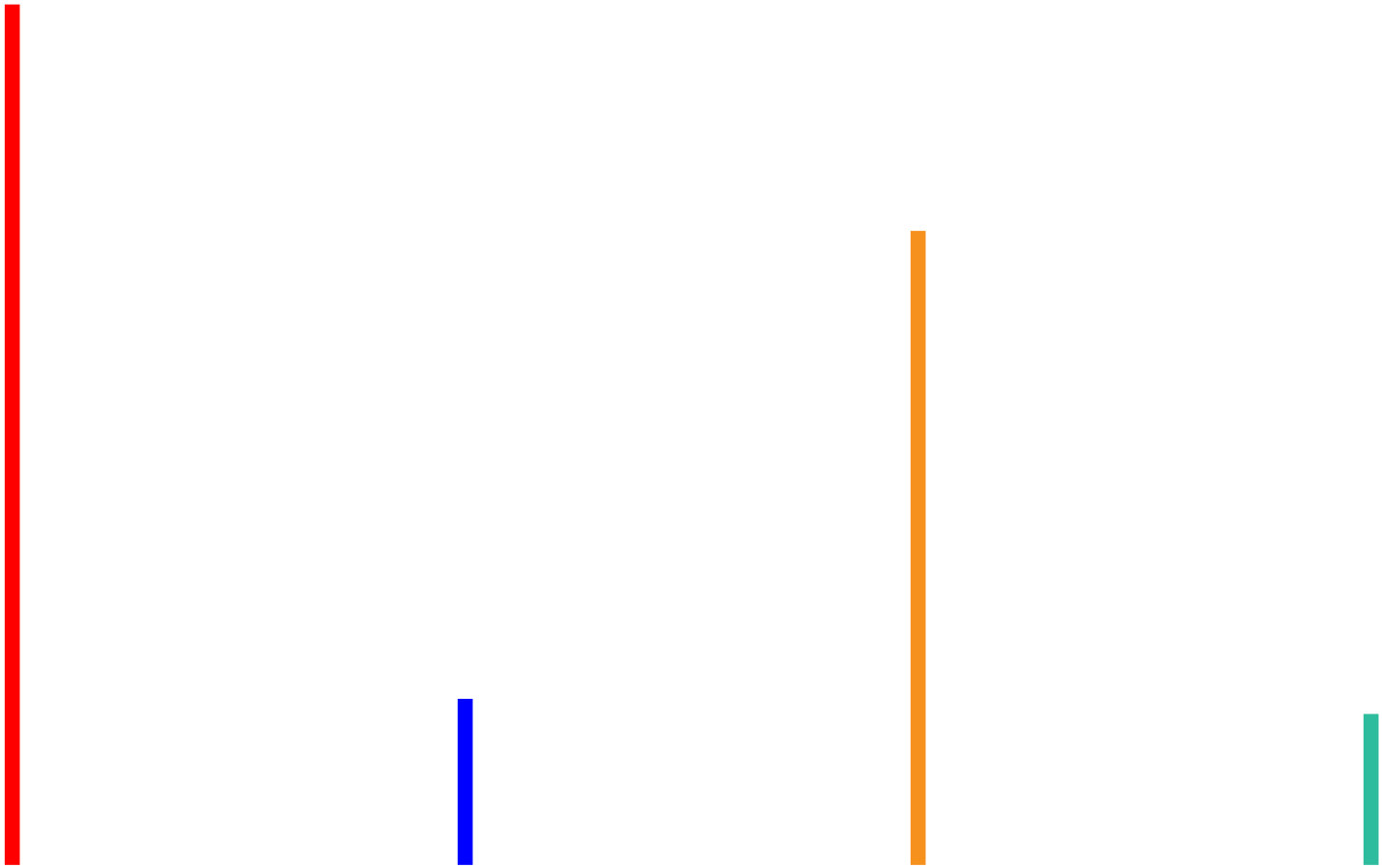
Sinon, on a une expression explicite de sa fonction de répartition F seulement en quelques points.
Exemple : $F(m^{-1}) = p_0$

On sait néanmoins que $F(X) \sim \mathcal{U}([0,1])$.

Construction de l'arbre de Yule

$0^5 20^2 110^3 120^4 10^3 21$

$0^2 20^2 11 \mid 12 \mid 0^1 1 \mid 21$



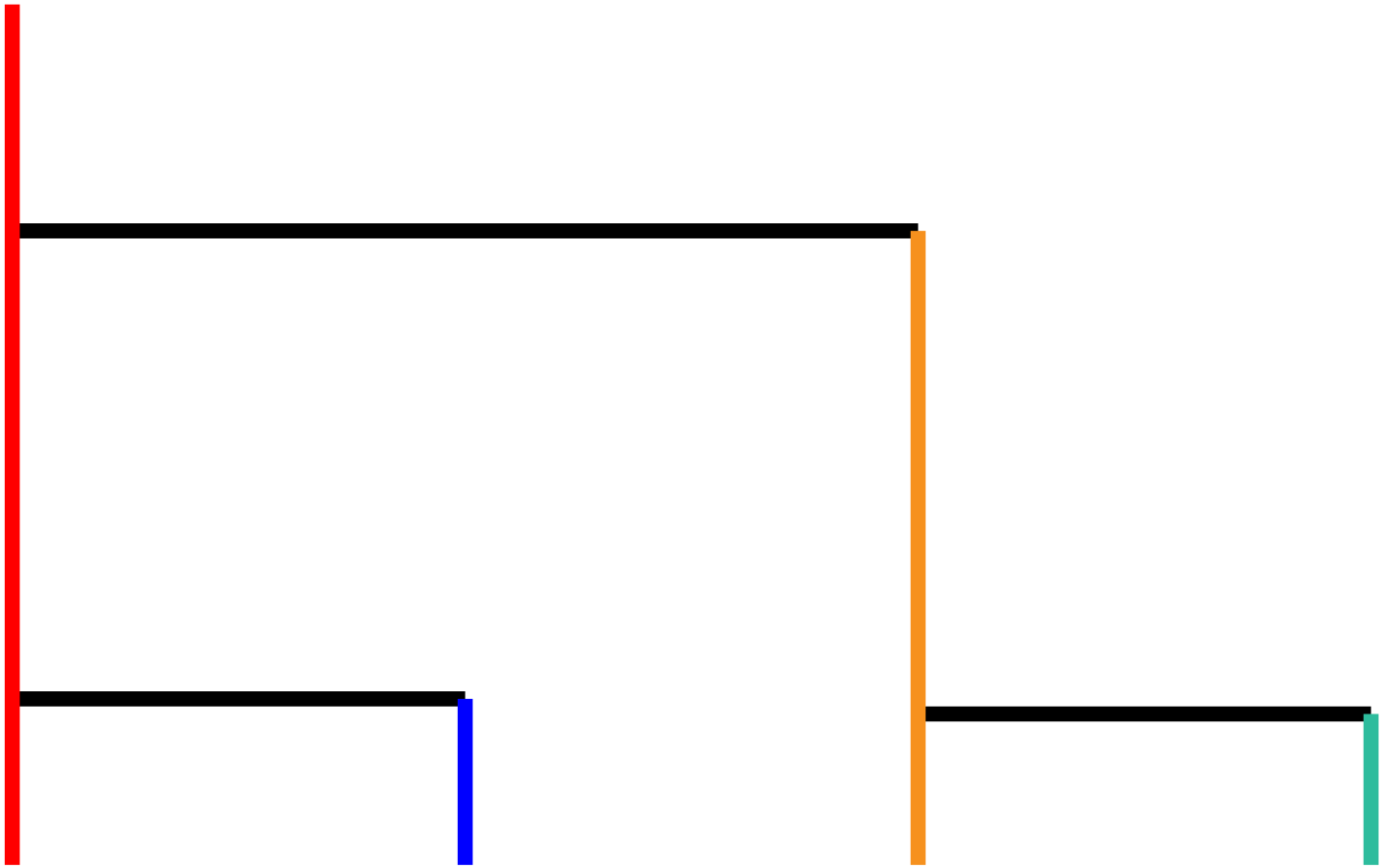
$\log_{\frac{1}{p_0}}(F(0,0^2 2\ldots))$ $\log_{\frac{1}{p_0}}(F(0,12\ldots))$ $\log_{\frac{1}{p_0}}(F(0,12\ldots))$ $\log_{\frac{1}{p_0}}(F(0,21\ldots))$

2.966227... 0.5335294... 1.1053397... 0.5350265...

Construction de l'arbre de Yule

$0^5 2 0^2 1 1 0^3 1 2 0^4 1 0^3 2 1$

$0^2 2 0^2 1 1 \mid 1 2 \mid 0^1 1 \mid 2 1$

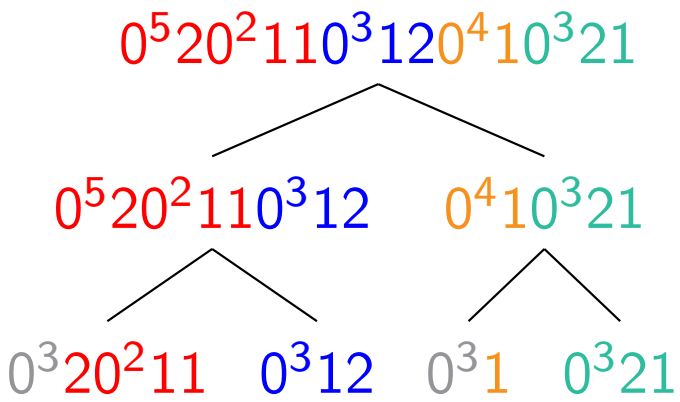


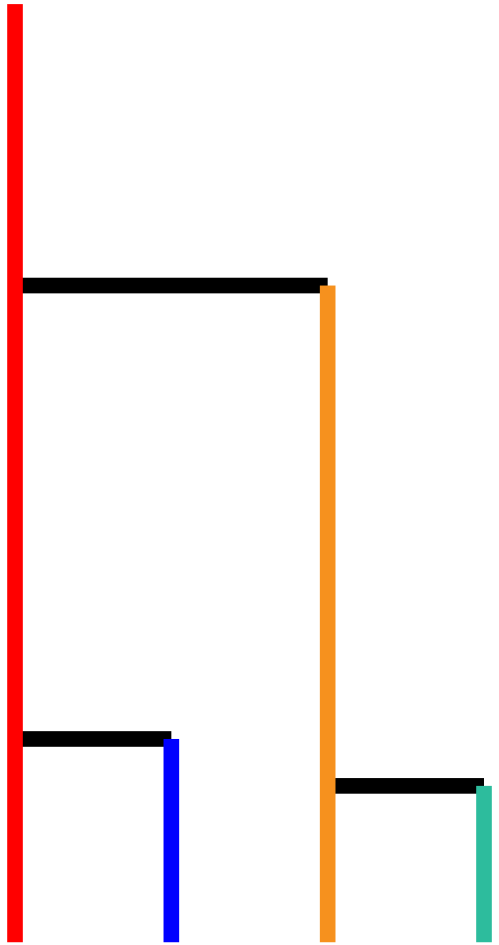
$\log_{\frac{1}{p_0}}(F(0,0^2 2\ldots))$ $\log_{\frac{1}{p_0}}(F(0,12\ldots))$ $\log_{\frac{1}{p_0}}(F(0,12\ldots))$ $\log_{\frac{1}{p_0}}(F(0,21\ldots))$

$2.966227\ldots$ $0.5335294\ldots$ $1.1053397\ldots$ $0.5350265\ldots$

Lien entre hauteur dans l'arbre de Yule et de Lyndon

Si on oublie les étapes de la décomposition où on éjecte juste un zéro, la structure des arbres est identique.





Lien entre hauteur dans l'arbre de Yule et de Lyndon

Pour une palme w_i, on considère le chemin dans l'arbre seuillé entre cette palme et la racine.

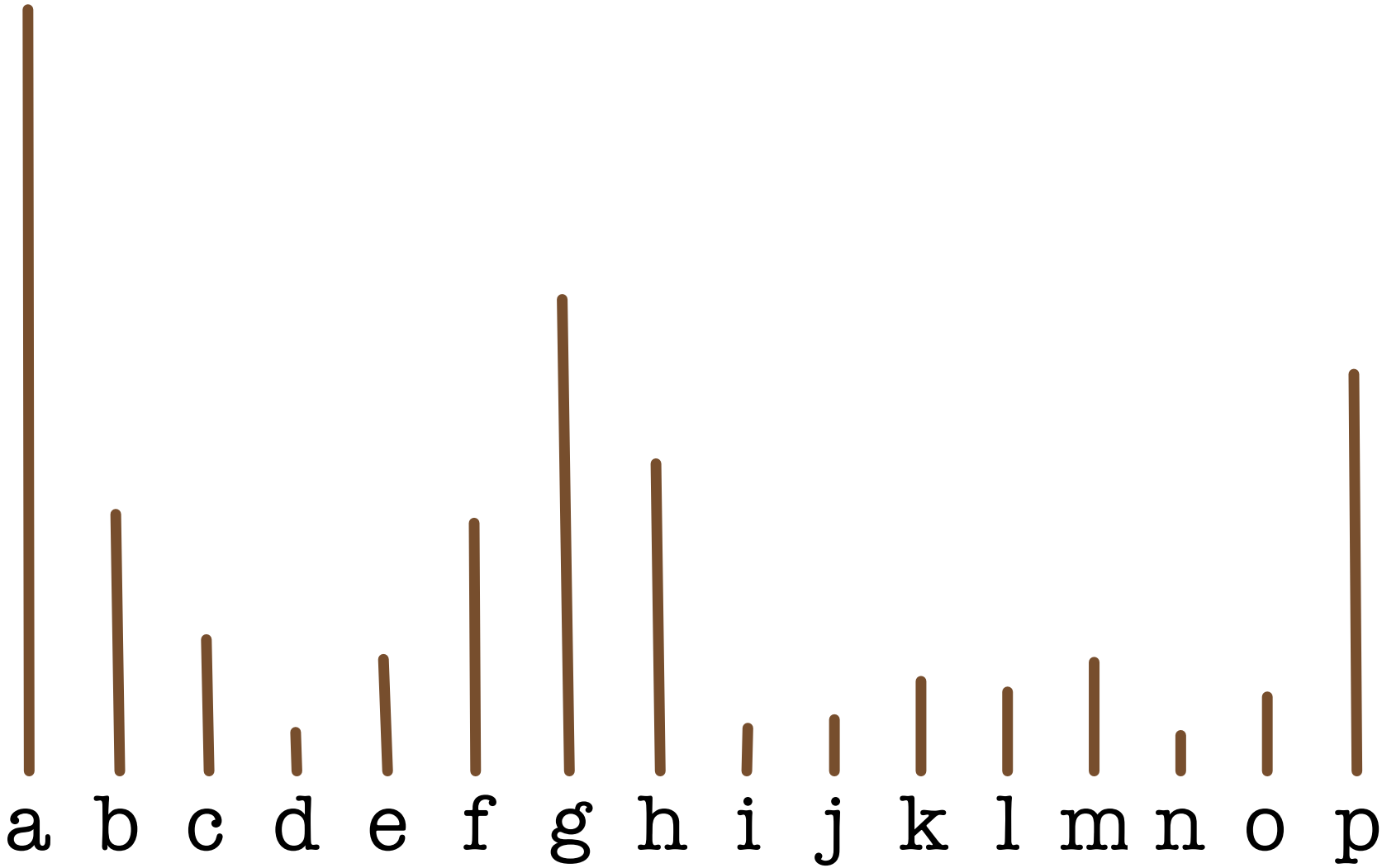
Lemme

Un embranchement entre w_i et w_j correspond à au moins k zéros éjectés si et seulement si

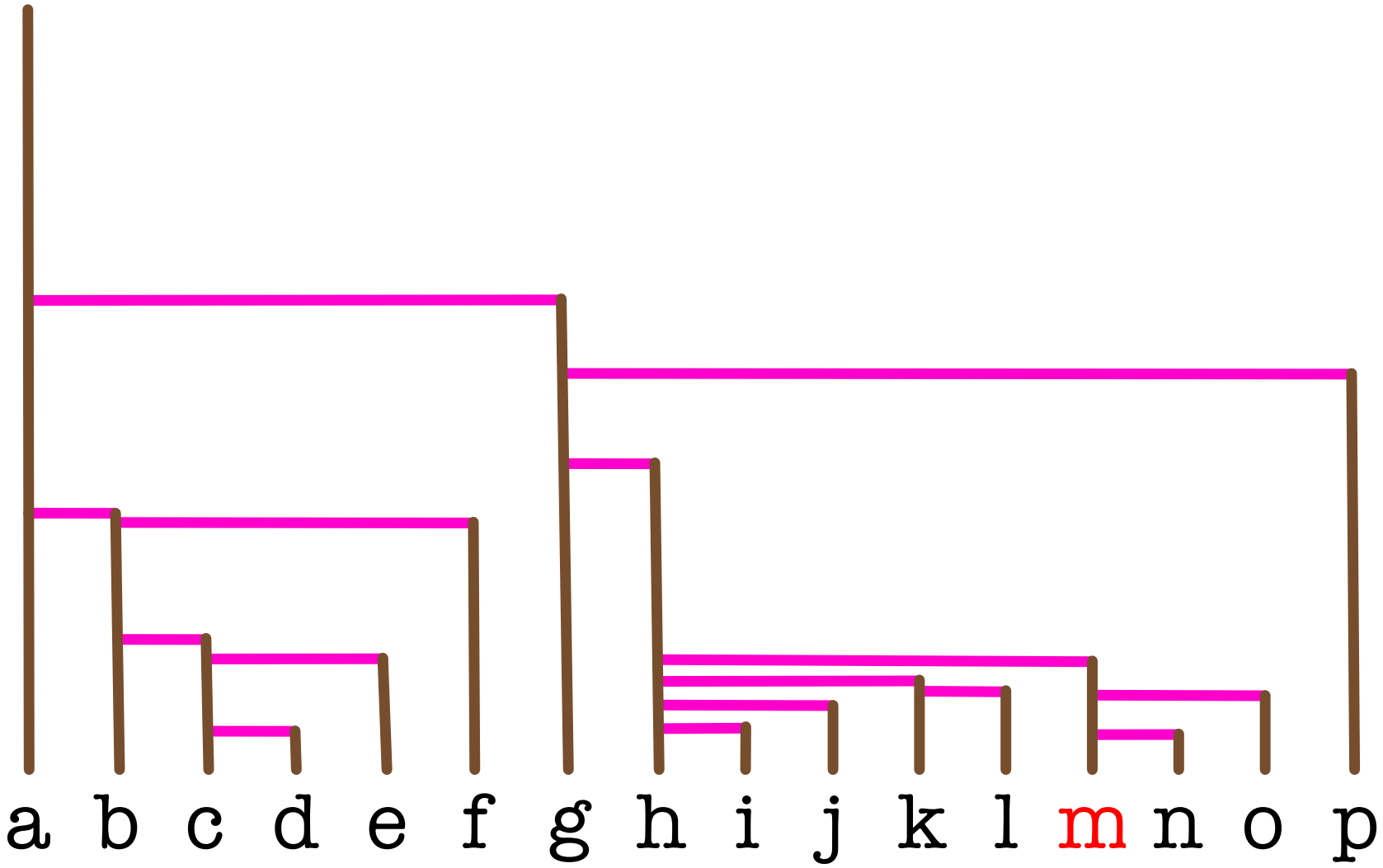
$$w_i < 0^k w_j$$

$$\iff -\log_{p_0}(F(x_i)) > k - \log_{p_0}(F(x_j))$$

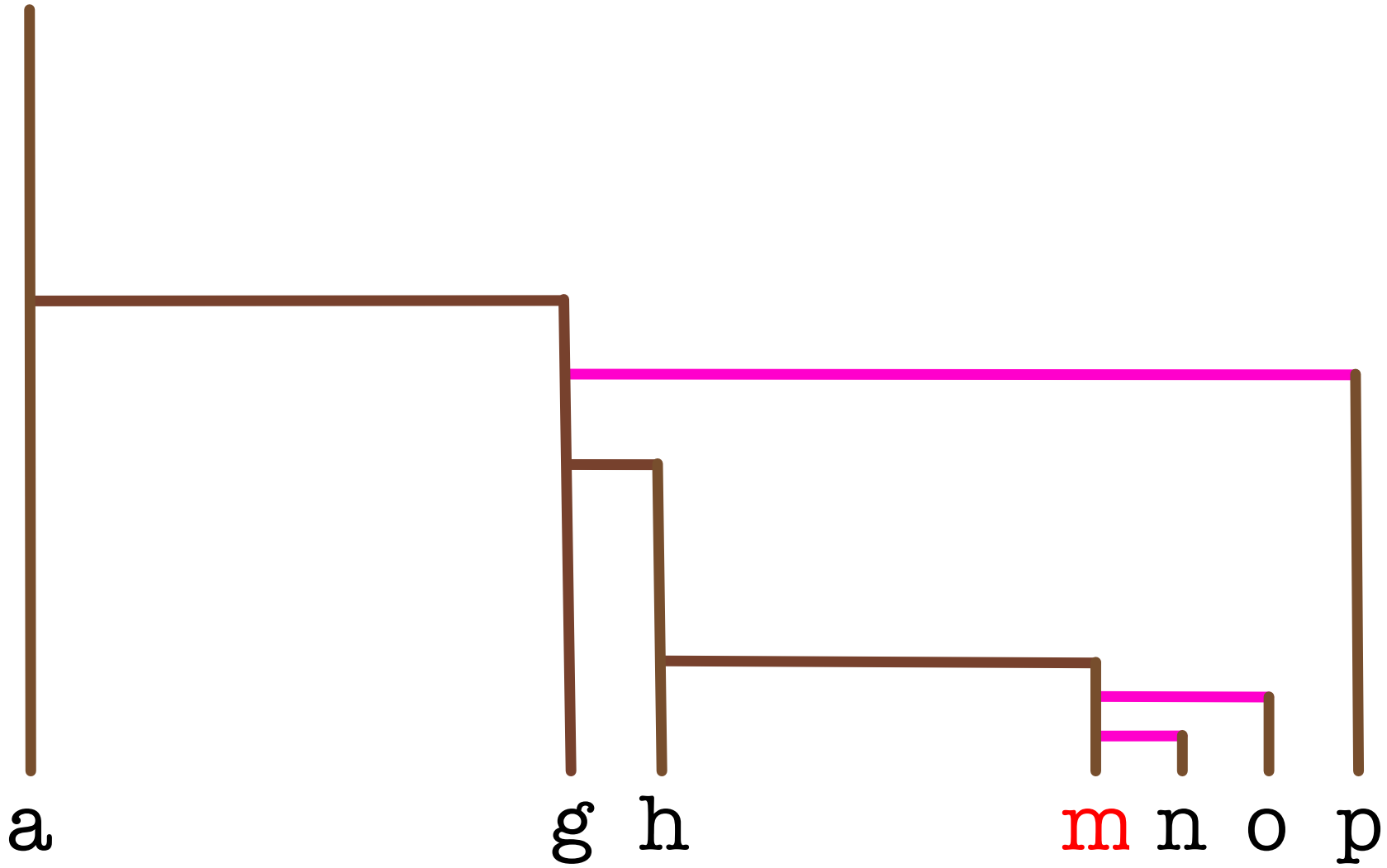
Lien entre hauteur dans l'arbre de Yule et de Lyndon



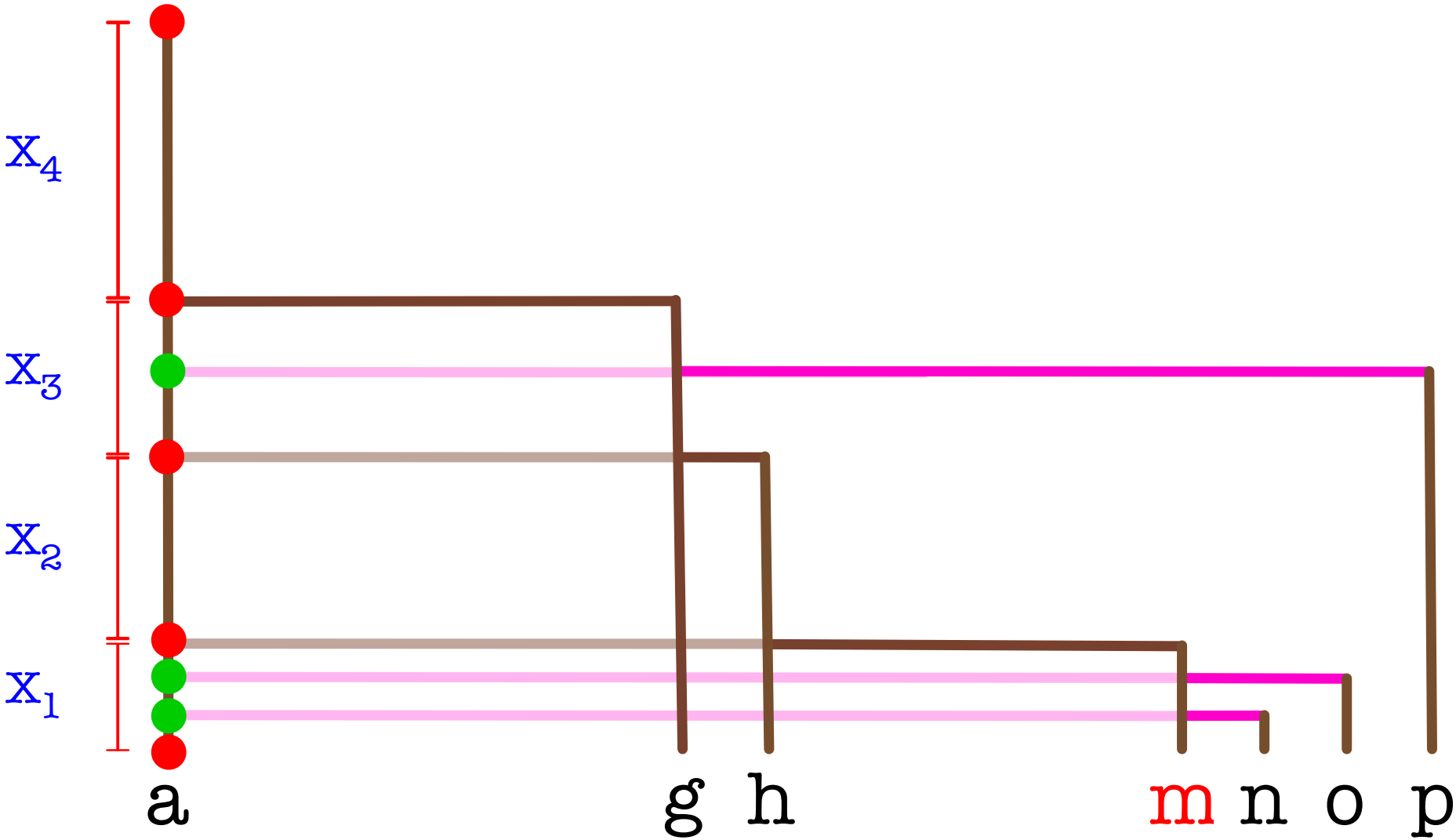
Lien entre hauteur dans l'arbre de Yule et de Lyndon



Lien entre hauteur dans l'arbre de Yule et de Lyndon



Lien entre hauteur dans l'arbre de Yule et de Lyndon



Lien entre hauteur dans l'arbre de Yule et de Lyndon

Les embranchements vers la droite et vers la gauche dans ce chemin forment deux processus de Poisson Π_i^0 et
$\Pi_i^1 = \{X_0^i = 0 < X_1^i < \ldots < X_k^i < X_{k+1} =$ nombre de zéros au début du mot - seuil $\}$.
On note $G(\Pi_i^1) = \sum_{j=0}^{k} \lfloor X_{j+1}^i - X_j^i \rfloor$.

Proposition

La hauteur h_i de la feuille w_i dans l'arbre seuillé est telle que :

$$0 \leq h_i - (\#\Pi_i^0 + \#\Pi_i^1 + 1 + G(\Pi_i^1)) \leq 1$$

Dénombrement des palmes

Notation

On note

$$e^{k\Psi(\lambda,\rho,\gamma)}$$

le nombre moyen de palmes de type $\#\Pi_i^0 = \lambda k, \#\Pi_i^1 = \rho k, G(\Pi_i^1) = \gamma k$ *dans l'arbre qui commence par un facteur* 0^k.

L'expression de Ψ ne dépend que de p_0.

On a donc environ $e^{k\Psi(\lambda,\rho,\gamma)}$ buissons qui viendront se brancher sur des palmes de ce type.
Hauteur de ces palmes :

$$k(\lambda + \rho + \gamma)$$

Greffe des buissons

Les buissons ont des hauteurs géométriques.

$$\mathbb{E}(\max(n\mathcal{G}(q))) \sim -\frac{\ln(n)}{\ln(1-q)}$$

donc le plus haut des buissons qui se greffera sur une palme de ce type aura pour hauteur :

$$-\frac{\ln(e^{k\Psi(\lambda,\rho,\gamma)})}{\ln(1-q)} = -\frac{k\Psi(\lambda,\rho,\gamma)}{\ln(1-q)}$$

On aura donc une hauteur totale de

$$k\left(\lambda+\rho+\gamma-\frac{\Psi(\lambda,\rho,\gamma)}{\ln(1-q)}\right)$$

Lien entre hauteur d'une feuille et records vers le bas

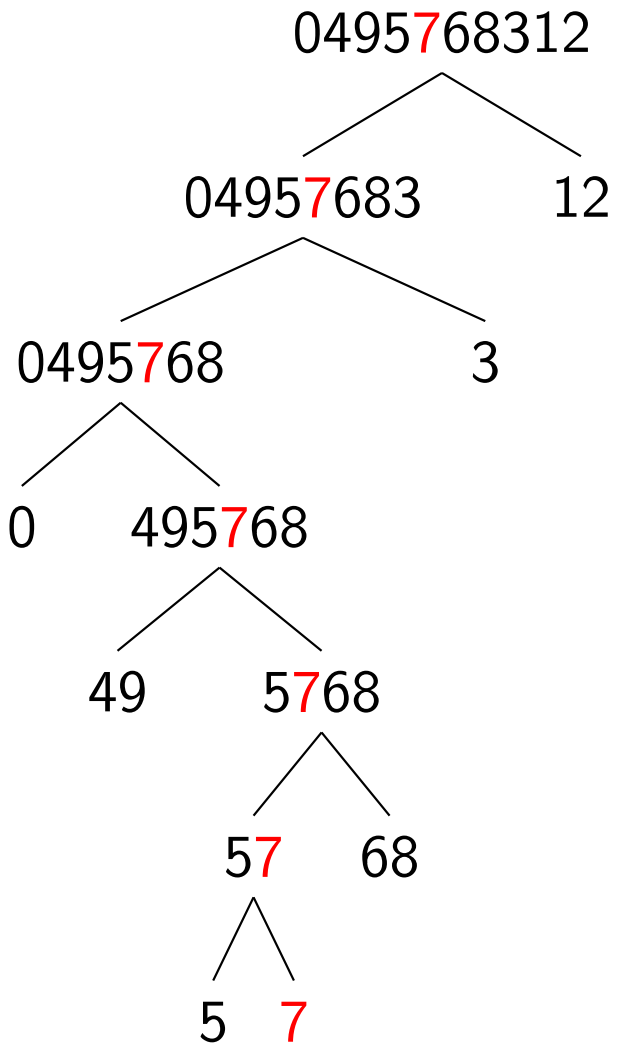
Pour minorer la hauteur d'un buisson, on utilise tout simplement la longueur d'un run de i^*, qui suit une loi $\mathcal{G}(1-p^*)$.

Pour majorer, on va s'intéresser à la hauteur de la feuille correspondant à une lettre.
On parcourt le mot de gauche à droite (resp. de droite à gauche) en partant de la lettre étudiée, et on note les endroits où l'on voit un mot plus petit que ce que l'on a vu jusque là. Cela forme le processus des records vers le bas de droite (resp. de gauche).

Propriété

La hauteur d'une feuille est égale aux cardinaux des processus des records vers le bas de droite et de gauche -1.

Cherchons la hauteur de la feuille 7



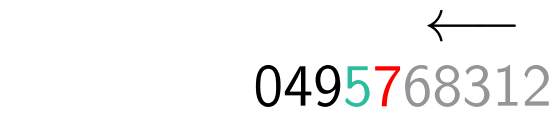
Cherchons la hauteur de la feuille 7

$\longleftarrow$

0495768312

Cherchons la hauteur de la feuille 7

Cherchons la hauteur de la feuille 7



Cherchons la hauteur de la feuille 7

Cherchons la hauteur de la feuille 7

Cherchons la hauteur de la feuille 7

$$\longrightarrow$$

$$0495768312$$

Cherchons la hauteur de la feuille 7

$$\longrightarrow$$

$$0495768312$$

Cherchons la hauteur de la feuille 7

$$\longrightarrow$$

$$0495768312$$

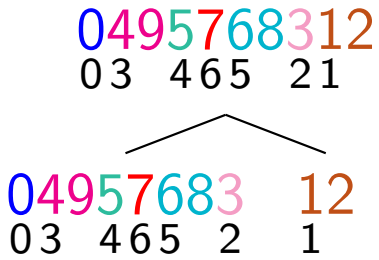
Cherchons la hauteur de la feuille 7

$$\longrightarrow$$

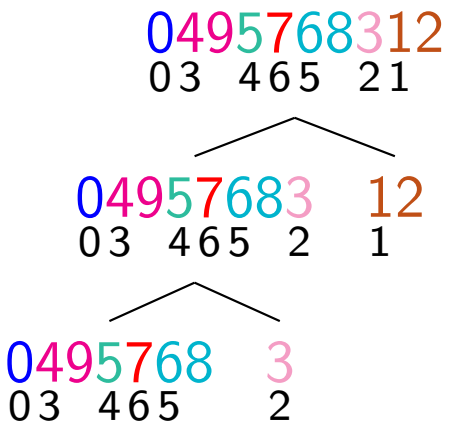
$$0495768312$$

Cherchons la hauteur de la feuille 7

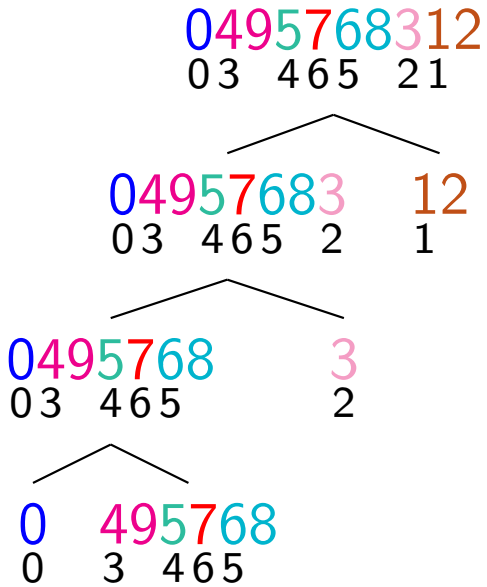
Cherchons la hauteur de la feuille 7



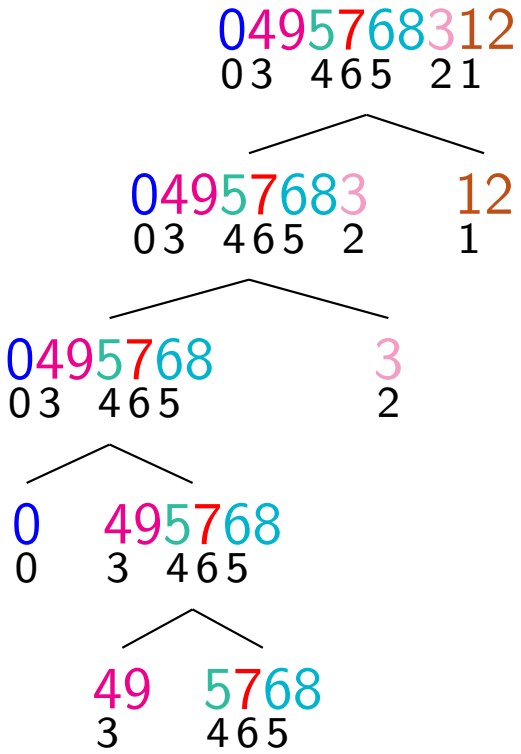
Cherchons la hauteur de la feuille 7



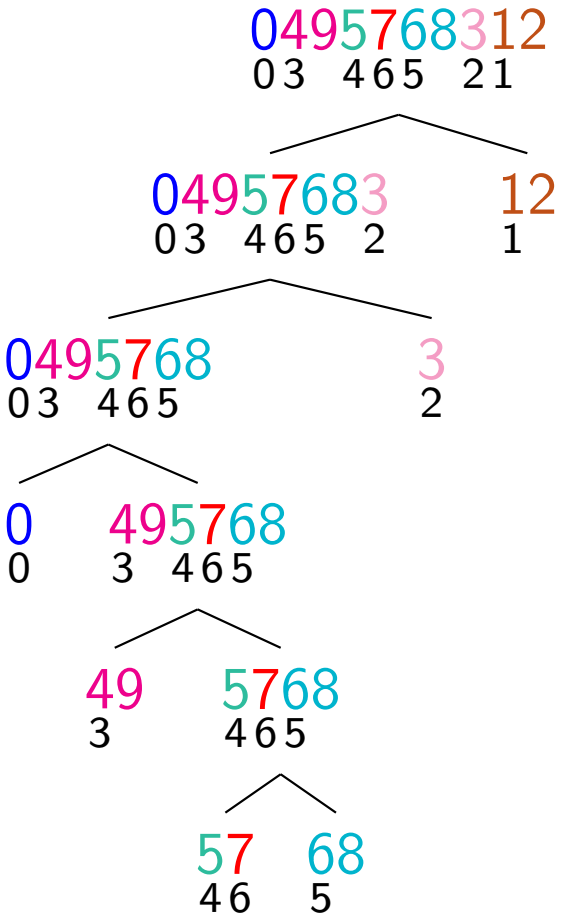
Cherchons la hauteur de la feuille 7



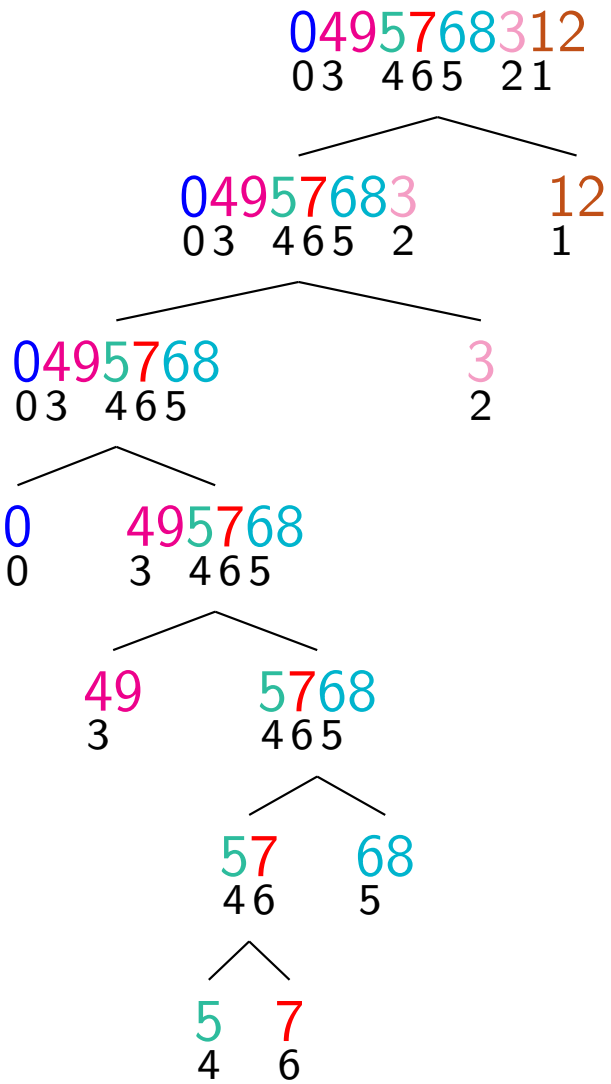
Cherchons la hauteur de la feuille 7



Cherchons la hauteur de la feuille 7



Cherchons la hauteur de la feuille 7



Hauteur des buissons

On utilise cette relation pour obtenir la majoration : dans un buisson commençant par 0^i, la hauteur de la feuille, puis la hauteur du buisson, est plus petite que la somme de Ci $\mathcal{G}(1-p^*)$ i.i.d., où C ne dépend que du nombre de lettres de l'alphabet.

Minoration $\rightarrow$ on utilise la longueur d'un run de i^*, qui suit une géométrique de paramètre $1-p^*$.

Problème

On minore par une géométrique et on majore par une somme de géométriques i.i.d., est-ce suffisant pour résoudre notre problème ?

Oui : $\mathbb{E}(\text{maximum de } n \text{ variables } \mathcal{G}(1-p^*) \text{ i.i.d.}) \sim -\log_{p^*}(n)$ et $\mathbb{E}(\text{maximum de } n \text{ sommes de } i \text{ variables } \mathcal{G}(1-p^*) \text{ i.i.d.}) \sim -\log_{p^*}(n)$ avec forte probabilité pour $i = O(\ln(n))$.

Hauteur

On note
$\Psi(\lambda,\rho,\gamma) = \log(p_0(-\log(p_0))^{\rho+\lambda}(e/\lambda)^{\lambda}(1+\rho/\gamma)^{\gamma}(1+\gamma/\rho)^{\rho}) + \Xi(\rho,\gamma)$,
où $\Xi(\gamma,\rho) = \lim 1/n \log(\mathbb{P}(\frac{1-\gamma}{\rho}n \leq \sum_{i=1}^{n} U_i \leq \frac{1-\gamma}{\rho}n + 1))$ et

$$\Delta^*(p_0,p^*) = \sup_{\Psi(\lambda,\rho,\gamma)>0} \{\lambda+\rho+\gamma - \frac{\Psi(\lambda,\rho,\gamma)}{\log(p^*)}\}\log(1/p_0).$$

Théorème

La hauteur H_n de l'arbre de Lyndon à n feuilles converge en probabilité vers $\Delta^*(p_0,p^*)\ln n$.

Les palmes telles que $\#\Pi_i^0 = \lambda k, \#\Pi_i^1 = \rho k, G(\Pi_i^1) = \gamma k$ contribuent à une hauteur de :

$$k\left(\lambda+\rho+\gamma - \frac{\Psi(\lambda,\rho,\gamma)}{\log(p^*)}\right),$$

ce qui donne le résultat avec $k \sim \log_{1/p_0} n$.

Hauteur

On note
$\Psi(\lambda,\rho,\gamma) = \log(p_0(-\log(p_0))^{\rho+\lambda}(e/\lambda)^{\lambda}(1+\rho/\gamma)^{\gamma}(1+\gamma/\rho)^{\rho}) + \Xi(\rho,\gamma)$,
où $\Xi(\gamma,\rho) = \lim 1/n \log(\mathbb{P}(\frac{1-\gamma}{\rho}n \leq \sum_{i=1}^{n} U_i \leq \frac{1-\gamma}{\rho}n + 1))$ et

$$\Delta^*(p_0,p^*) = \sup_{\Psi(\lambda,\rho,\gamma)>0} \{\lambda+\rho+\gamma - \frac{\Psi(\lambda,\rho,\gamma)}{\log(p^*)}\}\log(1/p_0).$$

Théorème

La hauteur H_n de l'arbre de Lyndon à n feuilles converge en probabilité vers $\Delta^*(p_0,p^*)\ln n$.

On note $\Delta_*(p_0) = \inf_{\Psi(\lambda,\rho,\gamma)>0}\{\lambda+\rho+\gamma\}\log(1/p_0)$.

Conjecture

La hauteur de la feuille étiquetée par un 1 (ou toute autre lettre que 0) la plus basse dans l'arbre de Lyndon à n feuilles converge en probabilité vers $\Delta_*(p_0)\ln n$.

Niveau de saturation

Définition

Le niveau de saturation d'un arbre est la hauteur de sa feuille la plus basse.

On déduit directement directement de l'arbre de Yule la limite du niveau de saturation de l'arbre à S_n :

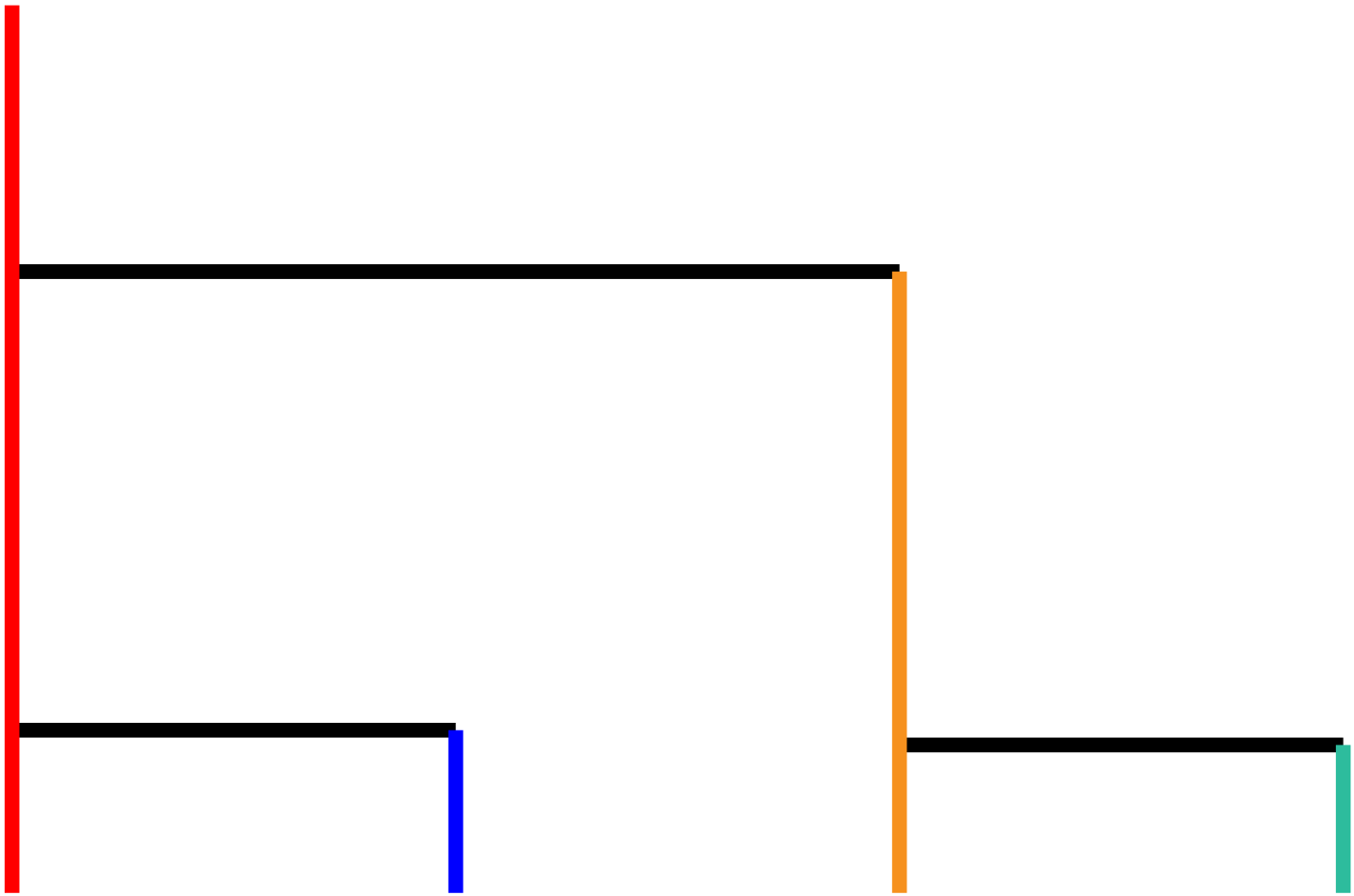
Théorème

On note $a_n = 1 - p_0 \sum_{i=0}^{n} - \log(p_0)^i / i!$ *la queue de distribution d'une variable de loi* $\mathcal{P}(-\log_{p_0})$.

$$\mathbb{P}(S_n = k) \underset{n\to\infty}{\longrightarrow} (1 - a_k \prod_{i=1}^{k-2} a_{k-2-i}^{2^{i-1}}) a_k \prod_{i=1}^{k-1} a_i^{k-i-1}$$

On remarque que cette loi ne dépend que de p_0.

Niveau de saturation



$\log_{p_0}(F(0,0^2 2\ldots))$ $\log_{p_0}(F(0,12\ldots))$ $\log_{p_0}(F(0,12\ldots))$ $\log_{p_0}(F(0,21\ldots))$

2.966227... 0.5335294... 1.1053397... 0.5350265...

Perspectives

- Meilleure approximation (équivalent du résultat de Mathew Roberts sur l'ABR)
- Généraliser le modèle ?