The height of the Lyndon tree

Lucas Mercier and Philippe Chassaing

Institut Élie Cartan, campus scientifique, BP 239, F-54506 Vandœuvre lès Nancy, France

Abstract. We consider the set $\mathcal{L}_n$ of n-letters long Lyndon words on the alphabet $\mathcal{A} = \{0, 1\}$. For a random uniform element L_n of the set $\mathcal{L}_n$, the binary tree $\mathfrak{L}(L_n)$ obtained by successive standard factorization of L_n and of the factors produced by these factorization is the *Lyndon tree* of L_n. We prove that the height H_n of $\mathfrak{L}(L_n)$ satisfies

$$\lim_n \frac{H_n}{\ln n} = \Delta,$$

in which the constant Δ is solution of an equation involving large deviation rate functions related to the asymptotics of Eulerian numbers ($\Delta \simeq 5.092\ldots$). The convergence is the convergence in probability of random variables.

Résumé. Pour un mot L_n choisi au hasard uniformément dans l'ensemble des mots de Lyndon de longueur n sur l'alphabet $\{0, 1\}$, on montre que la hauteur H_n de l'arbre de Lyndon associé possède le comportement asymptotique suivant

$$\lim_n \frac{H_n}{\ln n} = \Delta.$$

La constante Δ est définie à l'aide de fonctions de taux liées au comportement asymptotique des nombres eulériens (incidemment, $\Delta \simeq 5.092\ldots$). La convergence est la convergence en probabilité des variables aléatoires.

Keywords: Lyndon word, Lyndon tree, branching random walk, Galton-Watson tree, Yule process, binary search tree

1 Introduction

1.1 Lyndon words and Lyndon trees

We recall some notations of Lothaire (1997) for readability. For an alphabet $\mathcal{A}$, $\mathcal{A}^n$ is the set of n-letters words, and the language, i.e. the set of finite words,

$$\{\emptyset\} \cup \mathcal{A} \cup \mathcal{A}^2 \cup \mathcal{A}^3 \cup \ldots,$$

is denoted by $\mathcal{A}^\star$. The length of a word $w \in \mathcal{A}^\star$ is denoted by $|w|$. A total order, $\prec$, on the alphabet $\mathcal{A}$, induces a corresponding lexicographic order, again denoted by $\prec$, on the language $\mathcal{A}^\star$: the word w_1 is smaller than the word w_2 (for the lexicographic order, $w_1 \prec w_2$) at one of the following conditions: either w_1 is a prefix of w_2, or there exist words p, v_1, v_2 in $\mathcal{A}^\star$ and letters $a_1 \prec a_2$ in $\mathcal{A}$, such that $w_1 = pa_1v_1$ and $w_2 = pa_2v_2$.

The notion of *Lyndon word* has many equivalent definitions, to be found, for instance, in Lothaire (1997). For any factorization $w = uv$ of w, vu is called a rotation of w, and the set $\langle w \rangle$ of rotations of w is called the *necklace* of w. A word w is primitive if $|w| = \#\langle w \rangle$.

Definition 1 *A word w is Lyndon if it is primitive and if it is the smallest element of its necklace.*

Example 1 *The word $w = \texttt{aabaab}$ is the smallest in its necklace*

$$\langle w \rangle = \{\texttt{aabaab}, \texttt{abaaba}, \texttt{baabaa}\}$$

but is not Lyndon; baac *is not Lyndon, nor* acba *or* cbaa*, but* aacb *is Lyndon.*

Here is a recursive characterization of Lyndon words:

Proposition 1 *One-letter words are Lyndon. A word w with length $n \geq 2$ is a Lyndon word if and only if there exists two Lyndon words u and v such that $w = uv$ and $u \prec v$.*

Among such decompositions of w, the decomposition with the longest second factor (or suffix) v is called the *standard* decomposition.

Example 2 $\texttt{0011} = (\texttt{001})(\texttt{1}) = (\texttt{0})(\texttt{011})$ *is a Lyndon word with two such decompositions. The latter is the standard decomposition.*

The set of Lyndon words is denoted by $\mathcal{L}$, and we set $\mathcal{L}_n = \mathcal{L} \cap \mathcal{A}^n$. The Lyndon tree (or standard bracketing tree, cf. Barcelo (1990)) of the Lyndon word w is a binary tree obtained by iteration of the standard decomposition:

Definition 2 (Lyndon tree) *For $w \in \mathcal{L}$, the Lyndon tree $\mathfrak{L}(w)$ of w is a labelled finite binary tree defined as follows:*

- *if $|w| = 1$, $\mathfrak{L}(w)$ has a unique node labelled w, and no edges;*
- *if (u, v) is the standard decomposition of w, then $\mathfrak{L}(w)$ is the binary tree with label w at its root, $\mathfrak{L}(u)$ as its left subtree and $\mathfrak{L}(v)$ as its right subtree.*

Remark 1 *Note that the labels of the leaves of a Lyndon tree are letters. Also, the label of an internal node is the concatenation of the labels of its two children, and, if $|w| = n$, $\mathfrak{L}(w)$ is a binary tree with n leaves, and $n - 1$ internal nodes.*

1.2 Context

The asymptotic behavior of the size of the right and left subtrees of $\mathfrak{L}(L_n)$, for n large, have been studied in Bassino et al. (2005); Chassaing and Zohoorian Azad (2010), for L_n a random element of $\mathcal{L}_n$. The height $h(\mathfrak{L}(L_n))$ of $\mathfrak{L}(L_n)$ is of interest for the analysis of algorithms, cf. Sawada and Ruskey (2003), but it seems to have resisted analysis up to now.

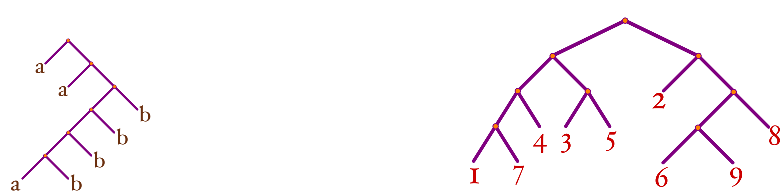

Fig. 1: a. $\mathcal{A} = \{\texttt{a}, \texttt{b}\}$ and $\mathfrak{L}(\texttt{a}^3\texttt{b}^4)$, b. $\mathcal{A} = \{\texttt{1}, \texttt{2}, \ldots, \texttt{9}\}$ and $\mathfrak{L}(\texttt{174352698})$.

1.3 Result

For a 2-letter alphabet, say $\mathcal{A} = \{0, 1\}$, and for $n \geq 1$, let L_n denote a uniform random word in $\mathcal{L}_n$. Let $(A(n,k))_{n,k}$ denote the Eulerian numbers, i.e. $A(n,k)$ is the number of permutations σ of n symbols having exactly k descents (k places where $\sigma(i) \geq \sigma(i+1)$) . Set

$$\Xi(\theta) = \lim_n \frac{1}{n} \ln(A(n, \lfloor \theta n \rfloor)/n!), \tag{1}$$

$$\Psi(\lambda,\mu,\nu) = \ln\left(\frac{(1+\mu)^{1+\mu}}{\mu^\mu} \frac{(e\lambda \ln 2)^\nu \ln 2}{\nu^\nu 2^\lambda}\right) + \Xi(\lambda - \mu), \tag{2}$$

$$\Delta = \sup_{\lambda,\mu,\nu>0} \frac{1+\nu+\mu}{\lambda} + \frac{\Psi(\lambda,\mu,\nu)}{\lambda \ln 2} = 5.092\ldots \tag{3}$$

See Lemma 1 for an expression of Ξ (Giladi and Keller, 1994, p. 299). We shall prove that:

Theorem 1

$$\frac{h(\mathfrak{L}(L_n))}{\ln n} \xrightarrow{\mathbb{P}} \Delta. \tag{4}$$

The two factors of the standard decomposition of a uniform Lyndon word are not uniform Lyndon words, and they are not even independent, which seems to preclude a recursive approach to the proof of this Theorem. We shall rather use a coupling method: in Section 2 we sketch the main steps of the construction, on the same probability space, of a random Lyndon tree, and of two well-known trees, the binary search tree of a random uniform permutation, and a Yule tree, in such a way that the height of the Lyndon tree is closely related to some statistics of the two other trees. Then Theorem 1 follows from a large deviation result presented in Section 3.

2 Coupling results

2.1 Reduction to a Bernoulli source

The Lyndon tree of a non-Lyndon word u is defined as the Lyndon tree of the unique Lyndon word in the necklace of u, if u is primitive. If u is periodic, we define the Lyndon word of u as the word $\mathtt{0}^{|u|-1}\mathtt{1}$, and the Lyndon tree of u is defined accordingly. Then the following algorithm:

- let W_∞ be a infinite word of uniformly random characters, obtained through the binary expansion of a number U uniformly distributed on $[0, 1]$;
- let W_n be the word W_∞ truncated after n letters, and let L_n be the Lyndon word of W_n.

produces a n-letter long random Lyndon word L_n, that fails to be uniform due to the small probability that W_n is periodic. However, the total variation distance between the probability distribution of L_n and the uniform distribution on $\mathcal{L}_n$ is $\mathcal{O}\left(2^{-n/2}\right)$ (cf. e.g. (Chassaing and Zohoorian Azad, 2010, Lemma 2.1)), thus any property that holds true asymptotically almost surely with respect to either distribution, holds true a.a.s. for both. We shall, from now on, consider that L_n is produced by the previous algorithm.

2.2 Exponential transform

In the first steps of the recursive construction of $\mathfrak{L}(L_n)$, the sizes of the factors of the successive standard decompositions are predicted by the positions of the longest runs of 0's, and the structure of the top levels of $\mathfrak{L}(L_n)$ is given by the lexicographic comparisons between the suffixes of L_n beginning at these longest runs. But when n is large, the number of runs of 0's is typically $n/4$, and several among these runs are tied for the title of the longest run. Actually the lengths of the runs behave pretty much as a sample of $n/4$ i.i.d. geometric random variables with parameter $1/2$, and, according to Brands et al. (1994), for any strictly increasing sequence n_k such that $\lim_k \log_2 n_k - \lfloor \log_2 n_k \rfloor = \alpha \in [0,1)$, the probability p_{ℓ,n_k} that $m \geq 1$ among the n_k elements of such a sample are tied for the maximum is given, approximately, by

$$p_{m,n_k} \simeq \sum_{j\in\mathbb{Z}} e^{-2^{\alpha+j}} \frac{(2^{\alpha+j-1})^m}{m!}. \tag{5}$$

Thus the number of ties does not converge in distribution, but has a set of limit distributions indexed by $\alpha \in \mathbb{R}/\mathbb{Z}$. Such a complex behavior does not bode well, so we shall rather analyze a transform of this problem, in the form of the Lyndon tree of a word with random length. Consider the finite word W^ℓ formed by a letter 1 followed by the truncation of W_∞ at the position τ_ℓ of the ℓth 0 in the first run of ℓ consecutive 0's of W_∞. Then W^ℓ is primitive, and L^ℓ denotes the Lyndon word of W^ℓ, *i.e.*:

$$W^\ell = 1\,\underbrace{010110\ldots1\overbrace{000000}^{\ell 0s}}_{\text{prefix of } W_\infty} \quad \text{and} \quad L^\ell = \overbrace{000000}^{\ell 0s}1\,\underbrace{010110\ldots1}_{\text{prefix of } W_\infty}.$$

If σ_ℓ is the position of the last 1 before τ_ℓ, L^ℓ is the concatenation of $0^\ell 1$ and of the truncation of the word W_∞ at position σ_ℓ.

Now, there exists a unique longest run of 0's in W^ℓ as well as in L^ℓ, and this run is ℓ letters long, to be compared with the behavior revealed by (5). Moreover, if Z_k denotes the number of runs longer than $\ell-k-1$, then $Z_0 = 1$ and $(Z_j)_{0\leq j\leq \ell-1}$ is a Galton-Watson process with offspring distribution $2^{-k}\mathbb{1}_{k\geq 1}$, so that Z_j has a geometric distribution with parameter 2^{-j}, see for instance Devroye (1992). The family tree of this Galton-Watson process gives a lot of information on $\mathfrak{L}(L^\ell)$, leading to the proof of Theorem 2 below. Note that the typical length of L^ℓ grows exponentially fast with ℓ:

$$\mathbb{E}[|L^\ell|] = 2^{\ell+1} - 1. \tag{6}$$

Thus, expectedly, the typical height of the Lyndon tree of L^ℓ grows linearly with ℓ:

Theorem 2 $\frac{h(\mathfrak{L}(L^\ell))}{\ell} \xrightarrow[\ell\to\infty]{\mathbb{P}} \Delta \ln 2.$

Due to space requirements, let us mention briefly how we obtain Theorem 1 using Theorem 2: we choose $\ell(n) = \log_2 n - \varepsilon_n$ in such a way that $\mathbb{P}\left(\tau_{\ell(n)} \geq n\right)$ is small, so that, with a large probability, $W^{\ell(n)}$ is a factor of W_n. Then we compare carefully $\mathfrak{L}(W^{\ell(n)})$ and $\mathfrak{L}(W_n)$. In the rest of the paper, we give more details about the proof of Theorem 2.

2.3 Reduction to a skeleton

In the top levels of the tree $\mathfrak{L}(L^\ell)$, the successive standard decompositions of the Lyndon word L^ℓ, at the smallest suffixes of L^ℓ, split the word L^ℓ at the longest runs of 0's. For ℓ large enough, the longest runs are sparse enough to preserve some degree of independence between the factors. This is not true anymore at the lowest levels of the tree $\mathfrak{L}(L^\ell)$. For this reason, it is easier to split the study of the Lyndon tree in two parts: the first one focuses on the top of the tree, where the runs of 0's are still above a threshold a_ℓ, and the second part studies a forest of shrubs at the bottom of the tree, each of them labeled with a factor of L^ℓ that contains only runs of 0's shorter than the treshold. The top part is a tree itself (a subtree of $\mathfrak{L}(L^\ell)$), and each shrub of the forest at the bottom of $\mathfrak{L}(L^\ell)$ is rooted at (or grafted on) a leaf of the top tree. We follow here the same path as Broutin and Devroye (2008), our *shrubs* playing the same rôle as their *spaghetti-like subtrees*. Let us define by induction the tree above the threshold k, with $k \geq 1$:

Definition 3 (Top tree) *If w denotes a Lyndon word, $\mathfrak{L}^k(w)$ is a finite labelled binary tree defined by:*

- *if w has one factor $\mathtt{0}^k$ or less (thus $\mathtt{0}^{k+1}$ is not a factor of w), $\mathfrak{L}^k(w)$ is a single node, labelled w;*
- *otherwise, let (u, v) be the standard decomposition of w. Then the root of $\mathfrak{L}^k(w)$ has label w, the left subtree of $\mathfrak{L}^k(w)$ is $\mathfrak{L}^k(u)$ and the right subtree is $\mathfrak{L}^k(v)$.*

$\mathfrak{L}^k(w)$ is called the top tree associated to w, with threshold k.

Since we focus on $\mathcal{L}(L^\ell)$, the threshold a_ℓ depends on ℓ. It has to be large enough for the top tree to retain the independence properties between the factors, but small enough that we can handle the shrubs, though they lack these nice independence properties. We set

$$a_\ell = \lfloor \log_2 \ell \rfloor = \Theta\left(\log \log \mathbb{E}\left[|L^\ell|\right]\right) \qquad \text{and} \qquad \mathfrak{T}_\ell = \mathfrak{L}^{a_\ell}(L^\ell). \tag{7}$$

2.4 A binary search tree

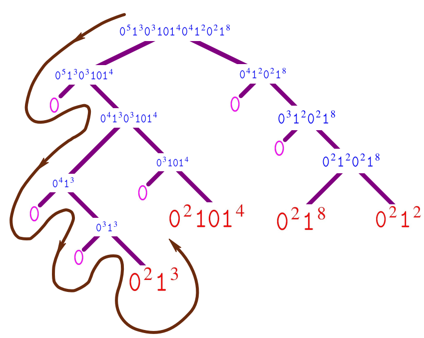

Fig. 2: For $L^5 = \mathtt{0}^5\mathtt{1}^3\mathtt{0}^3\mathtt{101}^4\mathtt{0}^4\mathtt{1}^8\mathtt{0}^2\mathtt{1}^2$, the tree $\mathfrak{T}_5 = \mathfrak{L}^2(L^5)$ has 6 needles and 4 blades. In brown, its contour traversal.

Let us take a closer look at $\mathfrak{T}_\ell$: observe that if $w \in \mathcal{L}$, then $\mathtt{0}w \in \mathcal{L}$, and the two factors of the standard factorization of $\mathtt{0}w$ are 0 and w. Thus either a leaf v of $\mathfrak{T}_\ell$ has label 0, and v is called a *needle*, or the label of v is a factor of w beginning with $\mathtt{0}^{a_\ell}\mathtt{1}$, and v is called a *blade*. The number N_ℓ of blades of $\mathfrak{T}_\ell$ has a geometric distribution with parameter $2^{-\ell+a_\ell}$, and the set of blades has a natural order related to the contour traversal (see Figure 2), that allows to identify it to $[\![1, N_\ell]\!]$. The number of needles has a simple expression in terms of a Galton-Watson process with geometric offspring distribution. In the analysis of the shape of $\mathfrak{T}_\ell$, the configuration of the needles is a special concern.

We shall need some notations: in the contour traversal of $\mathfrak{T}_\ell$, there is a sequence of $n_v - a_\ell \geq 0$ needles between a blade v and the previous blade (or between v and the root, if there exists no previous blade). The concatenation, starting at this sequence of needles, included, of the labels of the leaves in the order of the contour traversal of $\mathfrak{T}_\ell$, is a suffix of L^ℓ that can be written $\mathtt{0}^{n_v}\mathtt{1}\,t_v$, the run $\mathtt{0}^{n_v}$ being maximal in the sense that $\mathtt{0}^{n_v+1}\mathtt{1}\,t_v$ is not a suffix of L^ℓ.

The words of the sequence $(t_v)_{1\leq v\leq N_\ell}$ have different lengths, being proper suffixes of each other, so they are all different, and we can give a reformulation of the algorithm that produces $\mathfrak{T}_\ell$, or more generally $\mathfrak{L}^k(w)$, in terms of the family $T_\ell = ((n_v, t_v))_{1\leq v\leq N_\ell}$ of the blades (with labels $\mathtt{0}^k\mathtt{1}\,t_v$), in which only the n_v's and the relative order of the t_v's matter. With this reformulation of the algorithm, a slight perturbation of the t_v's produces a new tree, $\mathfrak{S}_\ell$, that is easier to handle than $\mathfrak{T}_\ell$ due to its property of independence of labels, but that has essentially the same profile as $\mathfrak{T}_\ell$ (i.e. it has the same repartition of blades with respect to the height). Let $\epsilon^{(j)}$ denote the sequence of integers defined, for $j \in I$, by

$$\epsilon_i^{(j)} = \delta_{i,j}.$$

For $\mathcal{R}$ a totally ordered set, $\mathbb{N}_0 \times \mathcal{R}$ inherits a lexicographic order, $\prec$, from $\mathcal{R}$: $(n, t) \prec (m, u)$ if $n > m$ or if $n = m$ and $t < u$. Let $B = (l_i, r_i)_{1\leq i\leq N}$ (resp. $L = (l_i)_{1\leq i\leq N}$, $R = (r_i)_{1\leq i\leq N}$) be a finite sequence of elements of $\mathbb{N}_0 \times \mathcal{R}$ (resp. $\mathbb{N}_0$, $\mathcal{R}$), with no repetitions in the sequence R. Assume that $l_j \geq k$ for each j, and that (l_1, r_1) is the smallest element of B, for $\prec$.

Definition 4 *The Lyndon tree $\mathfrak{L}^k(B)$ is defined by induction by:*

1. *If $N = 1$ and $l_1 = k$, $\mathfrak{L}^k(B)$ has no edge and its unique vertex, with label (k, r_1), is a blade.*
2. *Otherwise, consider the new sequence B' formed from $L - \epsilon^{(1)}$ and R and let i_0 denote the index of the smallest element in B', for $\prec$.*
   - *(a) If $i_0 = 1$, then $\mathfrak{L}^k(B)$ is the binary tree with a needle (labelled $\mathtt{0}$) as its left child and $\mathfrak{L}^k(B')$ as its right child.*
   - *(b) If $i_0 \geq 2$, then the left subtree of the binary tree $\mathfrak{L}^k(B)$ is $\mathfrak{L}^k((l_i, r_i)_{1\leq i\leq i_0-1})$ while the right subtree of $\mathfrak{L}^k(B)$ is $\mathfrak{L}^k((l_{i+i_0}, r_{i+i_0})_{0\leq i\leq N-i_0})$.*

Remark 2 *Since $\sum(l_i+1)$ is strictly decreasing at each step, and the l_i's are not allowed to drop under k, $\mathfrak{L}^k(B)$ is well-defined as long as the r_i's are distinct. The N blades of $\mathfrak{L}^k(B)$ are labelled $(k, r_i)_{1\leq i\leq N}$, and, during the contour traversal, they appear in this order.*

Remark 3 *For $\mathcal{R} = \{t_v \mid 1 \leq v \leq N_\ell\}$ and $T_\ell = ((n_v, t_v))_{1\leq v\leq N_\ell}$ defined in this section,*

$$\mathfrak{L}^{a_\ell}(T_\ell) = \mathfrak{T}_\ell,$$

or more precisely, the shapes are the same, but the labels are different. When the label of some node of $\mathfrak{L}^{a_\ell}(T_\ell)$ is (k, t_v), the corresponding label of $\mathfrak{T}_\ell$ is the prefix of $\mathtt{0}^k\mathtt{1}t_v$ that stops before the beginning of the next occurrence of $\mathtt{0}^k$.

For the analysis of $\mathfrak{T}_\ell$, the fact that the t_v's are suffixes of t_1, precluding any form of independence, is bothering. In order to fix the problem, in T_ℓ, we replace the sequence $(t_v)_{1\leq v\leq N_\ell}$ with a new sequence $(s_v)_{1\leq v\leq N_\ell}$ of infinite binary words, close to the t_v's but independent, defined as follows: let $(\zeta_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of uniform infinite words, independent of T_ℓ and let s_v be the concatenation of p_v, the prefix formed by the first a_ℓ letters of t_v, with ζ_v. When t_{N_ℓ} is shorter than a_ℓ letters, it is completed with the appropriate number of $\mathtt{0}$'s, before the concatenation with ζ_{N_ℓ}. This way, we obtain a new sequence $S_\ell = ((n_v, s_v))_{1\leq v\leq N_\ell}$, and we set

$$\mathfrak{S}_\ell \ = \ \mathfrak{L}^{a_\ell}\left(S_\ell\right).$$

Differences between $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ occurs scarcely, only when at least a_ℓ letters are used to distinguish two suffixes, so that $\mathfrak{T}_\ell$ and $\mathfrak{S}_\ell$ are close in some sense, see Proposition 3. The probability distribution of S_ℓ is given by:

Proposition 2 *The sequence of words $\tilde{S}_\ell = (\mathtt{0}^{n_v-a_\ell}\mathtt{1}s_v)_{2\leq v\leq N_\ell}$, followed by the word $\mathtt{0}^{n_1-a_\ell}\mathtt{1}s_1$, is distributed as a sequence of uniform infinite random words observed until the first occurrence of the prefix $\mathtt{0}^{\ell-a_\ell}$, this first occurrence $\mathtt{0}^{n_1-a_\ell}\mathtt{1}s_1$ being eventually truncated of any $\mathtt{0}$ in excess of $\mathtt{0}^{\ell-a_\ell}\mathtt{1}\ldots$, so that $n_1 = \ell$.*

Remark 4 *By properties of repeated coin flipping, equivalently, $n_1 = \ell$ and for $v \geq 2$, $n_v - a_\ell$ is a geometric random variable with parameter $\frac{1}{2}$, conditioned to be smaller than $\ell - a_\ell$, s_v is an infinite uniform random word, for all v n_v and s_v are independent, N_ℓ is geometric with parameter $2^{a_\ell-\ell}$, and the sequence $(n_v, s_v)_{v\in\mathbb{N}}$ is independent of N_ℓ.*

Remark 5 *When ℓ grows, with a large probability the prefix p_v is short compared to t_v and the corresponding, much longer suffix of t_v determines, with p_v, the shape of the corresponding shrub at the bottom of $\mathfrak{L}_\ell$. These suffixes are independent of S_ℓ, and for that matter, independent of $\mathfrak{S}_\ell$, which happens to be crucial in the study of the bottom of the tree.*

For a blade $v \in [\![1, N_\ell]\!]$, let h_v (resp. $\tilde{h}_v$) be its height in $\mathfrak{S}_\ell$ (resp. in $\mathfrak{T}_\ell$). Set

$$d_v = \left|h_v - \tilde{h}_v\right|, \qquad D_\ell = \max_{1\leq v\leq N_\ell} d_v.$$

Then

Proposition 3

$$\lim_\ell \mathbb{P}\left(D_\ell \ \geq \ \frac{\ell}{\sqrt{\ln \ell}}\right) \ = \ 0.$$

Due to space requirements, we omit the proof, that uses branching random walks arguments, as in Biggins (1977). We shall now be interested in the height of leaves of $\mathfrak{S}_\ell$. Due to Proposition 2, conditionally given N_ℓ, the ranks of the terms of the sequence $\tilde{S}_\ell = (\mathtt{0}^{n_v-a_\ell}\mathtt{1}s_v)_{2\leq v\leq N_\ell}$ with respect to the lexicographic order form a uniform permutation on $N_\ell - 1$ symbols. As a consequence, the subtree $\mathfrak{A}_\ell$ of $\mathfrak{S}_\ell$ induced by the root and the blades, once the needles and the first blade erased, forms a *binary search tree*. With the help of a coupling of $\mathfrak{A}_\ell$ with a Yule process, Chauvin et al. (2005) provides a very precise analysis of the depths of the leaves of a binary search tree. Though, the depths of blades in $\mathfrak{S}_\ell$, though they depend on their depths in $\mathfrak{A}_\ell$, are also affected by the positions of the needles, and we need to tweak the arguments of Chauvin et al. (2005) in order to include the needles in their analysis.

2.5 A Yule process

A *Yule process* $\mathcal{Y}$ (Athreya and Ney, 1972, page 109) models a population in which each individual lives forever, and gives birth to new individuals at the times of a Poisson process with rate λ. We assume that the population starts at time 0 with a single individual, called the *ancestor*. One can keep track of the genealogy of the population through the *Yule tree* $\mathfrak{Y}$ (cf. Chauvin et al. (2005)), a family tree of the population: for each individual, a vertical life line is drawn downward, on the right of the life line of its father, and starting at an ordinate given by *minus* the date of birth of this individual ; this vertical life line is connected to the life line of its father by a dotted horizontal line. Let $\mathfrak{Y}_t$ denote the family tree $\mathfrak{Y}$ truncated at time t, i.e. at ordinate $-t$.

The Yule tree has yet another construction, starting from a sample $Y = (Y_n)_{n\geq 1}$ of i.i.d. exponential random variables with rate λ: consider the sequence $Z^{(t)}$ defined by

$$T_t = \inf\{n \geq 1 \,|\, Y_n \geq t\}, \quad Z^{(t)} = (Z_k^{(t)})_{0\leq k\leq T_t-1} = (t, Y_1, Y_2, \ldots, Y_{T_t-1})\,; \tag{8}$$

now, picture each term $Z_k^{(t)}$ of the sequence $Z^{(t)}$ by a vertical line of the corresponding length $Z_k^{(t)}$, drawn at abscissa $T_t - k$, and connect with an horizontal line the top of the kth line, but the first, to the next line on its left whose height exceeds $Z_k^{(t)}$, to obtain a family tree $\mathfrak{E}_t$. Then $\mathfrak{E}_t$ and $\mathfrak{Y}_t$ have the same probability distribution, by the lack of memory of the exponential distribution, and, from now on, we shall retain the notation $\mathfrak{Y}_t$ for both of them.

Let us denote by U_v the real number whose dyadic development is $0^{n_v-a_\ell}1s_v$. Then, as a consequence of Proposition 2, $(U_2, U_3, \ldots, U_{N_\ell}, U_1)$ is distributed as a sequence of uniform random variables observed until the first term smaller than $2^{-\ell+a_\ell}$ occurs, this last term U_1 being eventually multiplied by a power of 2, so as to belong to $[2^{-\ell+a_\ell-1}, 2^{-\ell+a_\ell}[$. Equivalently, the sequence $(X_2, X_3, \ldots, X_{N_\ell}, X_1)$ defined by

$$X_i \;=\; -\log_2 U_i$$

is distributed as a sequence of exponential random variables observed until the first term larger than $\ell - a_\ell$ occurs, this last term X_1 being eventually shifted by an integer, so as to belong to $[\ell - a_\ell, \ell - a_\ell + 1[$. Then the construction of Figure 3, based on the sequence $(\ell - a_\ell, X_2, X_3, \ldots, X_{N_\ell})$, gives a Yule family tree $\mathfrak{Y}_{\ell-a_\ell}$ with lifetime $\ell - a_\ell$, and with intensity $\ln 2$. Now, as observed for instance in Chauvin et al. (2005), $\mathfrak{Y}_{\ell-a_\ell}$ seen as a planar tree, with no edge length, or with edge length 1, is the binary search tree $\mathfrak{A}_\ell$: their planar tree structure depends only on the relative order of the words $0^{n_v-a_\ell}1s_v$ for $\mathfrak{A}_\ell$, and of the real numbers X_v, through the same algorithm, and the mapping $0^{n_v-a_\ell}1s_v \longrightarrow X_v$ is monotone. Thus the depth of the blade v in the binary search tree $\mathfrak{A}_\ell$ induced by $\mathfrak{S}_\ell$ is the depth of the leaf v corresponding to

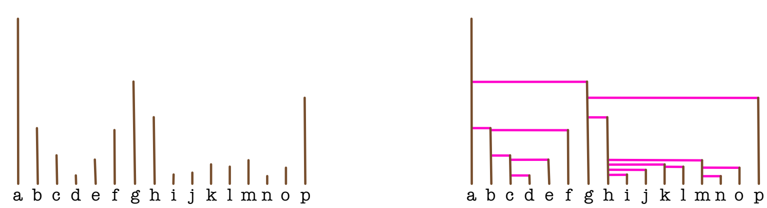

Fig. 3: A sample of exponential random variables, until hitting t, and the related tree $\mathfrak{E}_t$, distributed as $\mathfrak{Y}_t$ (in which the pink horizontal lines are to be seen as vertices rather than as edges).

X_v in $\mathfrak{Y}_{\ell-a_\ell}$. The difference between the depth of the blade v in $\mathfrak{A}_\ell$ and its depth in $\mathfrak{S}_\ell$ is the number of needles on the path to the root, whose expression is given by Proposition 4 below.

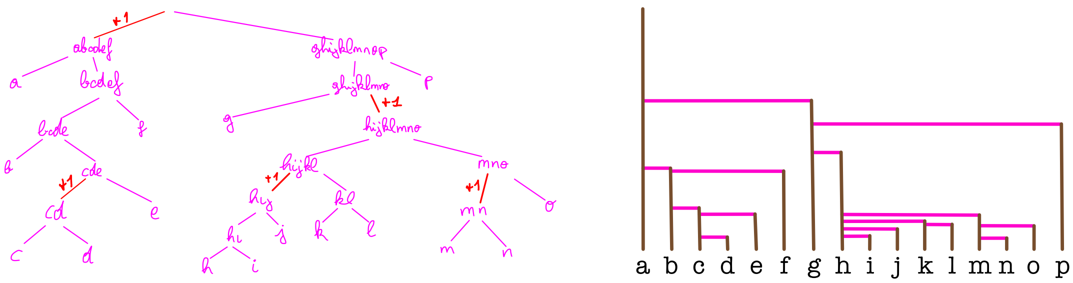

Fig. 4: Trees $\mathfrak{S}_\ell$ and $\mathfrak{Y}_\ell$. The sign $+1$ marks the presence of a needle.

For a given blade v of $\mathfrak{S}_\ell$, consider the marked point process Π_v formed by the vertices of $\mathfrak{Y}_{\ell-a_\ell}$ on the path from v to the root, the leaf v and the root excluded. This path is naturally identified with $[0, \ell - a_\ell]$, the leaf being at 0 and the root at $\ell - a_\ell$, for instance, and the mark being 0 or 1, respectively, according to the side of the branching, say left or right, leading to a decomposition

$$\Pi_v \;=\; \Pi_v^{(0)} \cup \Pi_v^{(1)}.$$

By convention, the mark for both points 0 and $\ell - a_\ell$ is one, and unless mentioned otherwise, they are not included in the point processes. For a point process $\pi = \{\xi_1 < \xi_2 < \ldots < \xi_{k-1} < \xi_k\}$ in some interval $[0, m]$, $G(\pi)$ denotes

$$G(\pi) \;=\; \sum_{s=1}^{k+1} \lfloor \xi_r - \xi_{r-1} \rfloor, \tag{9}$$

in which $\xi_{k+1} = m$ and $\xi_0 = 0$. We have

Proposition 4 *For* $1 \leq v \leq N_\ell$*,*

$$0 \leq h_v - (\#\Pi_v + 1 + G(\Pi_v^{(1)})) \leq 1.$$

Here $\#\Pi_v + 1$ accounts for the depth of v in $\mathfrak{A}_\ell$, $\lfloor \xi_1 \rfloor$ is $n_v - a_\ell$, and one can check that $\lfloor \xi_r - \xi_{r-1} \rfloor = k$ if the corresponding labels satisfy $(n_{w_r} - k, s_{w_r}) \prec (n_{w_{r-1}} - k, s_{w_{r-1}})$ but $(n_{w_r} - k - 1, s_{w_r}) \succ (n_{w_{r-1}} - k, s_{w_{r-1}})$, in which case the corresponding edge in $\mathfrak{A}_\ell$ is obtained by erasing k needles of $\mathfrak{S}_\ell$. The special rôle of $\Pi_v^{(1)}$ in Proposition 4 reflects the asymmetry of the Lyndon tree. Since $G(\Pi_v^{(1)})$ tends to be large when $\#\Pi_v^{(1)}$ is small, and small when $\#\Pi_v^{(1)}$ is large, the difference between the height and the saturation level should be smaller for the Lyndon tree than for the binary search tree.

3 Large deviations and final result

3.1 The many-to-one formula

A leaf v of $\mathfrak{Y}_\ell$ is said to be *of type* (m, n, A) if its left (resp. right) depth in $\mathfrak{Y}_\ell$ is m (resp. n) and if $\Pi_v^{(1)}$ belongs to A. Let $\pi_{\ell,m,n,A}$ denote the average number of leaves of type (m, n, A) in $\mathfrak{Y}_\ell$ and let $\mathbb{U}_{n,\ell}$ denote the uniform probability distribution on the simplex $\{0 < \xi_1 < \xi_2 < \ldots < \xi_n < \ell\}$. Then

Proposition 5

$$\pi_{\ell,m,n,A} \;=\; (\ell \ln 2)^{m+n} 2^{-\ell}\, \mathbb{U}_{n,\ell}(A)/m!n!. \tag{10}$$

Up to a factor 2^ℓ, the right hand of (10) is the probability that two independent Poisson processes with intensity $\ln 2$ on $[0,\ell]$, $\Pi^{(0)}$ (resp. $\Pi^{(1)}$), have m (resp. n) points, and that $\Pi^{(1)}$ belongs to A. This is an instance of the many-to-one formula for branching random walks.

When A_k (resp. A_I) is the set of point processes Π on $[0,\ell]$ such that $G(\Pi) = k$ (resp. such that $G(\Pi) \in I$) , we set

$$\pi_{\ell,m,n,A_k} \;=\; \pi_{\ell,m,n,k}, \qquad \text{resp. } \pi_{\ell,m,n,I}.$$

3.2 Final argument and expression of Δ

For the final argument, it will be convenient to think of $\Psi(\lambda,\mu,\nu)$ as the following limit:

$$\Psi(\lambda,\mu,\nu) \;=\; \lim_n n^{-1}\, \ln\left(\pi_{\lambda_n n,\nu_n n,n,\mu_n n}\right), \tag{11}$$

when the sequence $(\lambda_n,\nu_n,\mu_n) = (\ell/n, m/n, k/n)$ converges to (λ,μ,ν), though we do not really need the existence of this limit for the proof of Theorem 2. We follow Broutin and Devroye (2008): roughly, the average number of blades of type $(\nu,1,\mu)\ell/\lambda$ in $\mathfrak{Y}_\ell$ or in $\mathfrak{S}_\ell$ is approximately $e^{\ell\Psi(\lambda,\mu,\nu)/\lambda}$, and their depth is approximately $\frac{1+\nu+\mu}{\lambda}\,\ell$. On each of these blades we graft shrubs that are Lyndon trees for Lyndon words that contain approximately 2^{a_ℓ} runs of `0`'s, all shorter than a_ℓ, and the same number of runs of `1`'s, whose lengths are independent geometric random variables with parameter $1/2$. We prove that the maximum height of a set of k such shrubs behaves like the maximum of a sample of k independent geometric random variables with parameter $1/2$, i.e. the maximum is essentially $\log_2 k$. As a consequence, the total height of the highest leaf in any shrub that is grafted on a blade of type $(\nu,1,\mu)\ell/\lambda$ in $\mathfrak{S}_\ell$ is approximately

$$\frac{1+\nu+\mu}{\lambda}\,\ell \;+\; \log_2\left(e^{\ell\Psi(\lambda,\mu,\nu)/\lambda}\right) \;=\; \frac{1+\nu+\mu}{\lambda}\,\ell \;+\; \ell\,\frac{\Psi(\lambda,\mu,\nu)}{\lambda\ln 2}.$$

Thus the highest leaf in such a tree is $\Delta\ell$ high, in which

$$\Delta \;=\; \sup_{\lambda,\mu,\nu>0} \left((1+\nu+\mu)\ln 2 \;+\; \Psi(\lambda,\mu,\nu)\right)/\lambda\ln 2. \tag{12}$$

Due to Proposition 3, the same is true when the shrubs are grafted on the blades of $\mathfrak{T}_\ell$, producing $\mathfrak{L}(L^\ell)$.

3.3 Large deviations for G under $\mathbb{U}_{n,\ell}$

Let us compute Ψ. For a sequence b of positive numbers, set

$$|b| \;=\; \sum_{j\in I} b_j, \quad \langle b\rangle \;=\; \sum_{j\in I} j b_j, \quad \text{and} \quad \mathcal{H}(b) \;=\; -\sum_{j\in I} b_j \ln(b_j),$$

and if the entries are integers, set:

$$b! \;=\; \prod b_i! \quad \text{and} \quad \binom{|b|}{b} \;=\; \frac{|b|\,!}{\prod b_i!},$$

whenever they are defined. If $|b| = 1$, $\mathcal{H}(b)$ is the Shannon entropy of b. Under $\mathbb{U}_{n,\ell}$, rather than the ξ_j's, we shall consider the vector $\gamma = (\gamma_j)_{1\leq j\leq n+1}$ of gaps between the order statistics, defined by

$$\gamma_j = \xi_j - \xi_{j-1},\ 1 \leq j \leq n+1,$$

with the usual convention, $\xi_0 = 0$ and $\xi_{n+1} = \ell$. The random vector γ is uniformly distributed on

$$\mathcal{D}_{n,\ell} \ =\ \{\sum_{j=1}^{n+1} x_j = \ell \quad \text{and} \quad \forall j,\ x_j \geq 0\},$$

and its distribution is denoted $\mathbb{U}_{n,\ell}$ again. Also, set $\rho = (\rho_j)_{1\leq j\leq n+1}$, in which $\rho_j = \lfloor \gamma_j \rfloor$. Then $G = |\rho|$. The probability $\mathbb{U}_{n,\ell}(\rho = r)$ depends only on $|r|$ and is given by:

$$\mathbb{U}_{n,\ell}\,(\rho = r) = n!\,\ell^{-n}\mathbb{P}(\ell - |r| - 1 < \sum_{j=1}^{n} U_j < \ell - |r|), \tag{13}$$

in which the U_i's are a sequence of i.i.d. uniform random variables on $[0,1]$, $\mathbb{P}\,(\ldots)$ is the volume of the domain $\{\rho = r\}$, and $\ell^n/n!$ is the volume of $\mathcal{D}_{n,\ell}$. Consider the sequence $\beta = (\beta_j)_{j\geq 0}$ defined by

$$\beta_j = \#\left\{1 \leq i \leq n+1 \,|\, \rho_i = j\right\},\ \text{satisfying } |\beta| = n+1,\ \text{and} \quad \langle\beta\rangle = |\rho| = G(\xi),$$

in order to take advantage of the symmetric rôle of the ρ_j's. Then

$$\mathbb{U}_{n,\ell}\,(\beta = b) \ =\ \frac{n!}{\ell^n}\binom{|b|}{b}\,\mathbb{P}(\ell - \langle b\rangle - 1 < \sum_{j=1}^{n} U_j < \ell - \langle b\rangle).$$

For distributions b such that $b_j = c_j n + o(n)$, $|c| = 1$, and $\langle c\rangle = \mu$, we find that

$$\frac{1}{n}\ \ln\left(\frac{n!}{\ell^n}\binom{|b|}{b}\right) \ =\ -1 - \ln\lambda + \mathcal{H}\,(c) + o(1), \tag{14}$$

and we know [(i)] that

$$\lim_n n^{-1}\ln\mathbb{P}(\theta n - 1 < \sum_{j=1}^{n} U_j < \theta n) \ =\ \Xi(\theta). \tag{15}$$

According to formulas (6.12) to (6.16) (Giladi and Keller, 1994, p. 299), Ξ is given by

Lemma 1 *For* $0 < \theta < 1$,

$$\Xi(\theta) \ =\ \ln\sinh\alpha - \alpha\coth\alpha + 1 - \ln\alpha,$$

in which α *is given implicitly by* $-\alpha^{-1} + 1 + \coth\alpha = 2\theta$.

[(i)] Due to an interpretation of $A(n,k)$ in terms of uniform random variables going back to Tanny (1973).

Now, if $|c| = 1$ and $\langle c \rangle = \mu$,

$$\mathcal{H}(c) \; \leq \; (1+\mu)\ln(1+\mu) - \mu \ln \mu, \tag{16}$$

with equality only if

$$c_k = \mu^k (1+\mu)^{-k-1}\, \mathbb{1}_{k \geq 0} = d_k^{(\mu)},$$

cf. Lemma 8.3.1 in (Ash, 1965, p. 238). As usual in large deviation theory, only the leading term, provided by (16), contributes, so that, using (13), (14) and (15), we obtain

$$\begin{aligned} n^{-1} \ln\left(\mathbb{U}_{n,\ell}\left(G = \mu n\right)\right) &= n^{-1} \ln\left(\mathbb{U}_{n,\ell}\left(\langle \beta \rangle = \mu n\right)\right) \\ &\simeq n^{-1} \ln\left(\mathbb{U}_{n,\ell}\left(\beta = d^{(\mu)} n\right)\right) \\ &= -1 - \ln \lambda + \mathcal{H}\left(d^{(\mu)}\right) \; + \; \Xi(\lambda - \mu). \end{aligned}$$

Together with (10), this leads to the expression of Ψ given in (2).

References

R. Ash. *Information theory*. Interscience Publishers John Wiley & Sons, 1965.

K. B. Athreya and P. E. Ney. *Branching processes*. Springer-Verlag, 1972.

H. Barcelo. On the action of the symmetric group on the free Lie algebra and the partition lattice. *J. Combin. Theory Ser. A*, 55(1):93–129, 1990.

F. Bassino, J. Clément, and C. Nicaud. The standard factorization of Lyndon words: an average point of view. *Discrete Math.*, 290(1):1–25, 2005.

J. D. Biggins. Chernoff's theorem in the branching random walk. *J. Appl. Probability*, 14(3):630–636, 1977.

J. J. A. M. Brands, F. W. Steutel, and R. J. G. Wilms. On the number of maxima in a discrete sample. *Statist. Probab. Lett.*, 20(3):209–217, 1994.

N. Broutin and L. Devroye. An analysis of the height of tries with random weights on the edges. *Combin. Probab. Comput.*, 17(2):161–202, 2008.

P. Chassaing and E. Zohoorian Azad. Asymptotic behavior of some factorizations of random words, 2010.

B. Chauvin, T. Klein, J.-F. Marckert, and A. Rouault. Martingales and profile of binary search trees. *Electron. J. Probab.*, 10:no. 12, 420–435 (electronic), 2005.

L. Devroye. A limit theory for random skip lists. *Ann. Appl. Probab.*, 2(3):597–609, 1992.

E. Giladi and J. B. Keller. Eulerian number asymptotics. *Proc. Roy. Soc. London Ser. A*, 445(1924): 291–303, 1994.

M. Lothaire. *Combinatorics on words*. Cambridge University Press, 1997.

J. Sawada and F. Ruskey. Generating Lyndon brackets. *J. Algorithms*, 46(1):21–26, 2003.

S. Tanny. A probabilistic interpretation of Eulerian numbers. *Duke Math. J.*, 40:717–722, 1973.