The height of the Lyndon tree

Philippe Chassaing
Résumé par Nicolas Broutin

Séminaire de Combinatoire Philippe Flajolet
Institut Henri Poincaré, Séance du 31 janvier 2013

Abstract

We consider the set $\mathcal{L}_n$ of n-letter Lyndon words on the alphabet $\{0,1\}$. For a random uniform element L_n of the set $\mathcal{L}_n$, the binary tree 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 the Lyndon tree of ln satisfies $\lim_n(H_n/\ln n) = \Delta$, in which the constant Δ is solution of an equation involving large deviation rate functions related to the asymptotics of Eulerian numbers ($\Delta \approx 5.091...$). The convergence is the convergence in probability of random variables. Joint work with Lucas Mercier.

1 Introduction and main result

For a word $w \in \{0,1\}^* := \cup_{n\geq 0}\{0,1\}^n$, define its *necklace* $\langle w\rangle$ to be the collection of its cyclic rotations. A word w is said to be *primitive* if the cardinal of its necklace $\#\langle w\rangle$ equals its length $|w|$. A *Lyndon word* is a word w that is primitive and lexicographically the smallest in its necklace (the lexicographic order is written $\prec$) [1, 4]. Lyndon words have a natural recursive structure since $w \in \mathcal{A}^*$ is Lyndon if either $|w| = 1$, or $|w| > 1$ and there exists two Lyndon words u and v such that $w = uv$ and $u \prec v$. The decomposition that maximizes the length of the second factor (v above) is called the *standard factorization.* This decomposition defines a binary tree $\mathfrak{L}(w)$ of a Lyndon word w, where the leaves a labelled with letters. Let $\mathcal{L}_n$ denote the set of Lyndon words on length n on $\{0,1\}$. The main question that is addressed concerns the asymptotics as $n \to \infty$ for the height of the *Lyndon tree* $\mathfrak{L}(L_n)$ when L_n is chosen uniformly at random in $\mathcal{L}_n$.

Let $(A(n,k))_{n,k}$ denote the Eulerian numbers and define

$$\Xi(\theta) = \lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{A(n,k)}{n!}\right)$$

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

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

Then, it is proved that

Theorem 1. *Let L_n be a uniformly random Lyndon word of length n, and let $h(\mathfrak{L}(L_n))$ be the height of the corresponding Lyndon tree. Then*

$$\frac{h(\mathfrak{L}(L_n))}{\ln n} \xrightarrow[n\to\infty]{} \Delta$$

in probability.

2 Approach and sketch of proof

One of the main difficulties is that the uniform distribution on the Lyndon words is not maintained by the recursive decomposition. The proof uses a number of approximations which reduce the problem to simpler versions of the tree.

A more convenient distribution on $\mathcal{L}_n$. First, it is more convenient to work with Lyndon words which are not uniform in $\mathcal{L}_n$. For a word $w \in \mathcal{A}^*$, one can define its Lyndon word $\ell(w)$ as follows: if w is not primitive, then $\ell(w) = 0^{|w|-1}1$, otherwise let $\ell(w)$ be the lexicographically smallest word in the necklace $\langle w \rangle$. For a uniformly random word W_n on $\mathcal{A}^*$, $\ell(W_n)$ is not uniform in $\mathcal{L}_n$, but the total variation distance between the distribution of $\ell(W_n)$ and the uniform distribution is $O(2^{-n/2})$. So in order to prove Theorem 1 it suffices to prove the corresponding result for $\ell(W_n)$. From now on, L_n denotes $\ell(W_n)$.

Lyndon words of random length. Since the standard factorization looks for Lyndon subwords, which must be the smallest in their necklace, the high level structure of the decomposition is given by the lengths and positions of the long runs of 0 (these are the only locations where a Lyndon word may start). In order to simplify the analysis which would involve ties between some of the longest runs, Lyndon words of random length are considered. Let W_∞ be the word consisting of a sequence of independent uniformly random letters. Let W^ℓ denote the word formed by the letter 1, followed by the truncation of W_∞ at the position of the ℓ-th zero in the first run of ℓ consecutive zeros. Then W^ℓ is primitive and one lets L^ℓ denote the corresponding Lyndon word. Note that $\mathbf{E}[|L^\ell|] \sim 2^{\ell+1}$.

In L^ℓ, the structure of the runs of zeros of decreasing lengths $\ell - k - 1$ is that of a Galton–Watson process with geometric(1/2) offspring distribution [3]. One then shows that proving Theorem 1 reduces to showing the following corresponding result for the height of $\mathfrak{L}(L^\ell)$:

Theorem 2.

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

Decomposition of the contributions to the height. As long as the Lyndon words of the decomposition are sufficiently long, the long runs of zero are sufficiently sparse, and the corresponding factors sufficiently independent. This stops being true when the words get short, and one wants to decompose the tree into two regions depending on the length of the runs of zeros: one first focus on the top of the tree denoted $\mathfrak{T}_\ell$, where the runs all have length at least $a_\ell := \lfloor \log_2 \ell \rfloor$, and then studies the remaining part which consists in a forest of pendant shrubs.

The top of the tree and a binary search tree. Roughly, the distribution of the locations of the longest run of zero is uniformly distributed in the word, so that *when it does a proper cut* the standard factorization cuts the word at a uniformly random location. This is precisely the kind of split that happens for the random binary search tree, and one can approximate the shape of the top of the tree by that of a random binary search tree. However, there is a small effect due to the following phenomenon: when a given run is much longer than the next one, for some time the standard factorization simply extracts some of the initial zeros one after another until a truely new run can take over, and produce a new uniform cut of the word. (Each such zero expelled is referred to as a *needle*.) Note that this effect is not negligible since it happens at every macroscopic cut with a positive probability. For this reason, the structure of the top of the tree is that of a random binary search tree in which some edges have been split in order to take this into account the effect of needles on the depth of leaves.

The game is then to estimate the number of leaves which lie at every depth k, so that one can in turn estimate the height h_k of the highest shrub grafted from a leaf at this depth (the highest of a bunch of independent random trees, so their number is crucial). The height of the tree is then $\sup_k(k + h_k)$. However, the phenomenon of needles discussed above makes the analysis much tricker, and especially the estimation of the number of leaves at a certain depth. By generalizing the analyses of the height of binary search trees using branching processes, the authors then obtain the first order asymptotics for the number of leaves of $\mathfrak{T}_\ell$ of certain types which allows to control the contributions of the binary search tree part and the needles part. Without going in the details of the definition of the types the parameters of interest are the left depth (number of edges going left on the path to a node) the right depth, and the number of needles. These are (essentially) tracked using the parameters $\ell = \lambda n$, νn for the right-depth, n for the left-depth, and μn for the needles. Then:

- the depth of a leaf of $\mathfrak{T}_\ell$ of type $(\nu n, n, \mu n) = (\nu, 1, \mu)\ell/\lambda$ is approximately $(1 + \mu + \nu)\ell/\lambda$;
- there are about $e^{\ell\Psi(\lambda,\mu,\nu)/\lambda}$ such leaves;
- the maximum height of a bunch of k independent shrubs is about $\log_2 k$ (for one this is geometric).

So the maximum height of a leaf $\mathfrak{L}(L^\ell)$ that goes through a leaf of the top $\mathfrak{T}_\ell$ of a given type is

$$\frac{1 + \nu + \mu}{\lambda} \cdot \ell + \log_2 e^{\ell\Psi(\lambda,\mu,\nu)},$$

hence the highest leaf is about $\Delta\ell$ high, with

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

The number of leaves of $\mathfrak{T}_\ell$ of a given type. The last unknown in (1), Ψ is the large deviation a rate function: the sheer number of leaves at a given level allow some events to occur that would not occur if one would consider any the path to any single leaf. Estimating the number of leaves of a given type reduces precisely to obtain the exponential rate of decay of the corresponding events for a single leaf as in [2].

References

[1] 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.

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

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

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