HAL open science

Limit law of the standard right factor of a random Lyndon word

Regine Marchand, Elahe Zohoorian Azad

▶ To cite this version:

Regine Marchand, Elahe Zohoorian Azad. Limit law of the standard right factor of a random Lyndon word. Combinatorics, Probability and Computing, 2007, 16, pp.417-434. hal-00002167

HAL Id: hal-00002167

https://hal.science/hal-00002167v1

Submitted on 1 Jul 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ccsd-00002167, version 1 - 1 Jul 2004

LIMIT LAW OF THE LENGTH OF THE STANDARD RIGHT FACTOR OF A LYNDON WORD

RÉGINE MARCHAND AND ELAHE ZOHOORIAN AZAD

Abstract. Consider the set of finite words on a totally ordered alphabet with q letters. We prove that the distribution of the length of the standard right factor of a random Lyndon word with length n, divided by n, converges to:

$$\mu(dx) = \frac{1}{q}\delta_1(dx) + \frac{q-1}{q}\mathbf{1}_{[0,1)}(x)dx,$$

when n goes to infinity. The convergence of all moments follows. This paper completes thus the results of [2], giving the asymptotics of the mean length of the standard right factor of a random Lyndon word with length n in the case of a two letters alphabet.

1. Introduction

Consider a finite totally ordered alphabet $\mathcal{A}$ and for each $n \in \mathbb{N} = \{1, 2, 3, \dots\}$, denote by $\mathcal{A}^n$ the set of words with length n on $\mathcal{A}$. A Lyndon word with length n is a word in $\mathcal{A}^n$ which is strictly smaller, for the lexicographic order, than each of its proper suffixes. We denote by $\mathcal{L}_n$ the set of Lyndon words with length n.

The standard right factor v of a Lyndon word w is its smallest proper suffix for the lexicographic order. Any Lyndon word w can be written uv, in which u is a Lyndon word and v is the standard right factor of w. We call uv the standard factorization of the Lyndon word w. Lyndon words were introduced by Lyndon [11], to build a base of the free Lie algebra over $\mathcal{A}$. The standard factorization plays a central part in the building algorithm of this base. For each Lyndon word w, we can build a binary tree in the following manner: say that w is the root, and has two children, that are the factors u and v of the standard factorization of w. Since u and v are still Lyndon words, they can also be divided into two standard factors which are their children and so on (see figure 1). Then the average height of these trees characterizes the complexity of the building algorithm (see Chen, Fox and Lyndon [6] and Lothaire [10]). Thus the informations on the length of the standard right factor of a random Lyndon word are essential for the analysis of the building algorithm.

For any Lyndon word $w \in \mathcal{L}_n$, let $R_n(w)$ denote the length of its standard right factor. Endowing $\mathcal{L}_n$ with the uniform probability measure makes R_n a random variable on $\mathcal{L}_n$. Bassino, Clément and Nicaud [2], with the help of generating functions, prove that the expectation $\mathbb{E}(R_n)$, in the case of a two letters alphabet, is asymptotically equal to $3n/4$. The aim of this paper is to determine the limit distribution of R_n/n as n goes to infinity:

2000 *Mathematics Subject Classification.* 68R15,60B10,68Q25.

Key words and phrases. random word, Lyndon word, standard right factor, longest run, convergence in distribution.

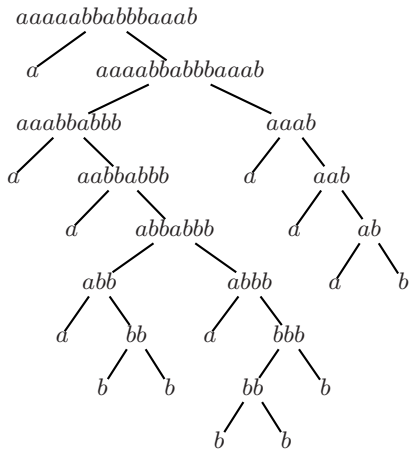

FIGURE 1. Binary tree associated to the standard decomposition of the Lyndon word *aaaaabbabbbaaab*

Theorem 1.1. *For a totally ordered alphabet with q letters, the normalized length R_n/n of a random Lyndon word of length n converges in distribution, when n goes to infinity, to*

$$\mu(dx) = \frac{1}{q}\delta_1(dx) + \frac{q-1}{q}\mathbf{1}_{[0,1)}(x)dx,$$

where δ_1 denotes the Dirac mass on the point 1, and dx Lebesgue's measure on $\mathbb{R}$. All the moments of R_n/n also converge to the corresponding moments of the limit distribution.

Remark. In the case $q = 2$, this result was conjectured by Bassino, Clément and Nicaud [2]. Simulations were also provided in this paper.

For the proof of this result, we focus first on the case of a two letters alphabet, and then we indicate the way to adapt the proof for the case of q letters.

Random Lyndon words are, in some sense, conditioned random words. In section 2, we obtain the number of Lyndon words with length n by dividing the number of primitive words with length n by n, (the shepherd's principle: counting the legs and dividing by four to obtain the number of sheeps). Thus the typical statistical behavior of a random word and of a random Lyndon word can be easily linked (see lemma 2.2).

Our analysis starts in section 3: we recall, among a number of well known properties of random words with length n, those useful for our purposes. In particular, we study the number of runs and the length of the longest run of "a", which is typically of order $\log_2 n$.

The key step is to prove that the two longest runs of "a" are approximately located along the word as two independent uniform random variables, and thus the distance D_n between the first longest one and the second longest one follows

approximately the uniform law on $[0..n]$. The distance D_n is of course closely related to the length of the standard right factor. We can distinguish two cases:

- If the word obtained from the Lyndon word by deleting the first "a" is still a Lyndon word, then the length of the standard right factor is equal to $n-1 \simeq n$, and this happens with a probability close to $1/2$ (this probability is replaced by $1/q$ in the case of an alphabet with q letters);
- Otherwise, the standard right factor begins by the second longest run of "a". In this case, the length of the standard right factor is equal to n minus the distance D_n, and is then approximately uniformly distributed on $[0..n]$.

To prove that D_n is approximately uniformly distributed on $[0..n]$, we cut a random word with length n into distinct "long blocks" with length of order $\log_2 n$ (section 4), in such a way that the long runs of "a" are at the beginnings of the long blocks. Then we prove that the uniform distribution on Lyndon words is invariant under uniform permutation of these blocks (section 5). Thus the positions of the two smallest (for the lexicographic order) long blocks are approximately uniformly distributed among all the possible positions of the long blocks. As n goes to infinity, the number of long blocks tends to infinity and their lengths are negligible when compared to n. This leads to our main result, Theorem 5.4, which says that the distance between the two smallest (for the lexicographic order) long blocks, divided by the length n of the word, follows asymptotically the uniform law on $[0,1]$. In section 6, we rephrase this result in terms of standard right factor and finally we generalize, in section 7, the obtained results to the case of an alphabet with q letters.

2. Random words and random Lyndon words

Let $\mathcal{A} = \{a, b\}$ be an ordered alphabet ($a < b$) and $\mathcal{A}^n$ be the set of all words with length n. If $w \in \mathcal{A}^n$, write $w = (w_1, \ldots, w_n)$ and define:

$$\tau w = (w_2, \ldots, w_n, w_1).$$

Then $< \tau >= \{Id, \tau, \ldots, \tau^{n-1}\}$ is the group of cyclic permutations of the letters of a word with length n. A word $w \in \mathcal{A}^n$ is called *primitive* if

$$(\exists k \in \{0, 1, \ldots, n-1\} \text{ such that } \tau^k w = w) \Rightarrow (k = 0).$$

Denote by $\mathcal{P}_n$ the set of primitive words in $\mathcal{A}^n$, by $\mathcal{N}_n$ its complement.

Remember that a Lyndon word with length n is a word in $\mathcal{A}^n$ which is strictly smaller, for the lexicographic order, than each of its proper suffix: it is equivalent to say that a word w with length n is a Lyndon word if and only if it is strictly smaller for the lexicographic order than every $\tau^k w$ with $k \in \{1, \ldots, n-1\}$. We denote by $\mathcal{L}_n$ the set of Lyndon words with length n.

The group $< \tau >$ of cyclic permutations acts on $\mathcal{A}^n$, and $\mathcal{P}_n$ and $\mathcal{N}_n$ are stable under this action. Each orbit associated to a primitive word w contains exactly n distinct words, and a unique Lyndon word, denoted by $\rho(w)$, which is the smallest word in the orbit for the lexicographic order: the application ρ is then the canonical projection of $\mathcal{P}_n$ on $\mathcal{L}_n$ associated to the action of $< \tau >$.

Example 2.1.

- *If $w = aabaaa$, then $\rho(w) = aaaaab$.*
- *If $A = \{aab, abb\}$ then $\rho^{-1}(A) = \{aab, aba, baa\} \cup \{abb, bab, bba\}$.*

As the set $\mathcal{N}_n$ of non-primitive words contains no Lyndon word, we have, by the shepherd's principle, that:

$$card(\mathcal{P}_n) = n \times card(\mathcal{L}_n).$$

Via the relation $card(\mathcal{P}_n) = \sum_{d|n} 2^{n/d}\mu(d)$, where μ is the Möbius function (see the book by Lothaire [10]), we are lead to:

$$card(\mathcal{L}_n) = \frac{2^n}{n}(1 + O(2^{-n/2})) \text{ and } card(\mathcal{N}_n) = O(2^{n/2}).$$

In the sequel, we will consider the two following probability spaces:

- the set $\mathcal{A}^n$ of words with length n, endowed with the uniform probability $\mathbb{P}_n$,
- the set $\mathcal{L}_n$ of Lyndon words with length n, endowed with the uniform probability $\tilde{\mathbb{P}}_n$.

The probability measure $\tilde{\mathbb{P}}_n$ can be seen as the conditional probability on $\mathcal{A}^n$, given $\mathcal{L}_n$. The next lemma is obvious, but it is very useful in our proofs because it allows to transfer results on random words to random Lyndon words by neglecting non-primitive words and using the shepherd's principle.

Lemma 2.2. *For $A \subset \mathcal{L}_n$, we have:*

$$| \tilde{\mathbb{P}_n}(A) - \mathbb{P}_n(\rho^{-1}(A)) | \leq O(2^{-n/2}).$$

Proof. It is sufficient to note that $\tilde{\mathbb{P}}_n(A) = \mathbb{P}_n(\rho^{-1}(A)|\mathcal{P}_n)$ and $\mathbb{P}_n(\mathcal{P}_n) = 1 - O\left(2^{-n/2}\right)$.

3. Number of runs and length of the longest run

This section deals with the number of runs and the length of the longest run in a random Lyndon word. The results exposed in this section are not new, but are presented in a convenient way for our proofs. The method is to get results for random words, and to transfer them to random Lyndon words via lemma 2.2.

Definition 3.1. *Let w be a word in $\mathcal{A}^n$. We denote by $N_n(w)$ the number of runs in w, by $X_1(w), X_2(w), \ldots, X_{N_n}(w)$ their lengths, by $M_n(w) = \max\{X_i(w),\ 1 \leq i \leq N_n(w)\}$ the length of the longest run in w and by $M_n^a(w)$ the length of the longest run of "a" in w.*

Example 3.2. $n = 9$

	N_6	(X_i)	M_6	M_6^a
$w = aabbbbaaa$	3	$2, 4, 3$	4	3
$\rho(w) = aaaaabbbb$	2	$5, 4$	5	5

Lemma 3.3 (Number of runs)**.** *For every $\gamma > 0$, there exists a constant $C_\gamma > 0$ such that for any $\varepsilon > 0$,*

$$\tilde{\mathbb{P}}_n\left(\left|N_n - \frac{n}{2}\right| \geq \gamma n^{-1/2+\varepsilon}\right) \quad \leq \quad O\left(\exp\left(-C_\gamma n^{2\varepsilon}\right)\right).$$

Proof. First, we prove the above inequality for the probability measure $\mathbb{P}_n$ on the set $\mathcal{A}^n$ of words with length n. The cardinal of the event $\{N_n = k\}$ corresponds

to the number of compositions of the integer n with k parts (see Andrews [1] and Pitman [12]):

$$\forall k \in \{1, 2, \ldots, n\},\ \mathbb{P}_n(N_n = k) = \frac{1}{2^{n-1}} \binom{n-1}{k-1}.$$

Thus $N_n - 1$ is a binomial random variable with parameters $(n-1, 1/2)$, whose large deviations are well known (see the book by Bollobás [5, Th. 7, p.13] for instance): there exists a positive constant C'_γ such that

$$\mathbb{P}_n\left(\left|N_n - \frac{n}{2}\right| \geq \gamma n^{-1/2+\varepsilon}\right) \quad \leq \quad O\left(\exp\left(-C'_\gamma n^{2\varepsilon}\right)\right).$$

To obtain the same inequality for the probability measure $\tilde{\mathbb{P}}_n$ on the set $\mathcal{L}_n$ of Lyndon words with length n, note that for a primitive word w, we have $N_n(w) - 1 \leq N_n(\rho(w)) \leq N_n(w)$. Thus, using Lemma 2.2, we obtain the announced result. ■

The next step is to study the length of the longest run of a word $w \in \mathcal{A}^n$. For this mean, we will use the following construction of the uniform probability measure on the set of all infinite words on $\mathcal{A}$:

Construction 3.4. *Let $(Z_i)_{i\in\mathbb{N}}$ be independent identically distributed geometrical random variables with parameter $1/2$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let ϵ be a Bernoulli random variable with parameter $1/2$ defined on Ω and independent of the Z_i's. To obtain a infinite random sequence of a and b, do the following:*

- *if $\epsilon = 1$, write Z_1 "a", followed by Z_2 "b", followed by Z_3 "a" and so on...*
- *if $\epsilon = 0$, write Z_1 "b", followed by Z_2 "a", followed by Z_3 "b" and so on...*

Truncating to keep the n first letters gives a random variable defined on Ω and uniformly distributed on $\mathcal{A}^n$. Thus, in this setting, the number of runs is:

$$N_n(\omega) = \inf\left\{k \in \mathbb{N},\ \sum_{i=1}^{k} Z_i(\omega) \geq n\right\},$$

and the lengths of runs are:

$$\forall i \in \{1, \ldots, N_n - 1\},\ X_i = Z_i \text{ and } X_{N_n} = n - \sum_{i=1}^{N_n - 1} Z_i \leq Z_{N_n}. \tag{1}$$

We denote by log the natural logarithm, by $\log_2(x) = \log x / \log 2$, by $\lceil x \rceil$ the smallest integer larger than x and by $\lfloor x \rfloor$ the largest integer smaller than x. The two next lemmas give estimates for the length of the longest run in a random Lyndon word. These estimates are related with the extreme values theory (see the books by Bingham, Goldie and Teugels [4], Resnick [13] or the work of Hitzenko and Louchard [9], or the initial works of Erdös and Rényi [7] and Erdös and Révész [8]).

Lemma 3.5 (Longest run, small values)**.** *For any $\varepsilon > 0$,*

$$\tilde{\mathbb{P}}_n\left(M_n \leq (1-\varepsilon)\log_2 n\right) \leq O\left(\exp\left(-\frac{n^\varepsilon}{4}\right)\right).$$

Proof. First, we prove the above inequality for the probability measure $\mathbb{P}_n$ on the set $\mathcal{A}^n$ of words with length n. The idea is that the number of the runs of "a" in a random word with length n is highly concentrated around $n/4$ and that it is easy to estimate the maximum of $n/4$ independent geometrical random variables.

Note that $M_n \geq M_n^a$. Using (1) and lemma 3.3 with $\gamma = 1$, we have:

$$\begin{aligned}
& \mathbb{P}_n(M_n \leq y) \\
\leq\ & \mathbb{P}_n(M_n^a \leq y) \\
\leq\ & \mathbb{P}_n\left(N_n \leq \frac{n}{2}\left(1 - n^{-\frac{1}{2}+\varepsilon}\right)\right) + \mathbb{P}_n\left(X_i \leq y,\ 1 \leq i \leq \frac{n}{4}\left(1 - n^{-\frac{1}{2}+\varepsilon}\right)\right) \\
\leq\ & \mathbb{P}_n\left(N_n \leq \frac{n}{2}\left(1 - n^{-1/2+\varepsilon}\right)\right) + \mathbb{P}_n\left(Z_i \leq y,\ 1 \leq i \leq \frac{n}{4}\left(1 - n^{-\frac{1}{2}+\varepsilon}\right)\right) \\
\leq\ & O\left(\exp\left(-C_1 n^{2\varepsilon}\right)\right) + (1 - 2^{\lfloor -y \rfloor})^{\lfloor \frac{n}{4}\left(1 - n^{-1/2+\varepsilon}\right) \rfloor}.
\end{aligned}$$

To lighten notations, we consider the $\frac{n}{4}\left(1 - n^{-\frac{1}{2}+\varepsilon}\right)$ first Z_i's rather than the Z_i's corresponding to runs of "a", which would have obliged us to distinguish whether the words begins with "a" or "b". Taking $y = (1-\varepsilon)\log_2 n$, we obtain easily:

$$\mathbb{P}_n\left(M_n^a \leq (1-\varepsilon)\log_2 n\right) \leq O\left(\exp\left(-\frac{n^\varepsilon}{4}\right)\right).$$

To obtain the same inequality for the probability measure $\tilde{\mathbb{P}}_n$ on the set $\mathcal{L}_n$ of Lyndon words with length n, note that, for a primitive word w, we have $M_n^a(w) \leq M_n^a(\rho(w))$. We can now use lemma 2.2 to conclude. ■

Lemma 3.6 (Longest run, large values)**.** *For any* $1 < A \leq 2$,

$$\tilde{\mathbb{P}}_n\left(M_n^a \geq A\log_2 n\right) \leq \tilde{\mathbb{P}}_n\left(M_n \geq A\log_2 n\right) \leq O\left(n^{1-A}\right).$$

Proof. As before, we begin by proving the above inequality for the probability measure $\mathbb{P}_n$ on the set $\mathcal{A}^n$. Note that we still have $M_n \geq M_n^a$. Thus for $y > 0$, we have:

$$\begin{aligned}
\mathbb{P}_n(M_n^a \leq y) &\geq \mathbb{P}_n(M_n \leq y) \\
&= \mathbb{P}_n(\forall i \in \{1,\dots,N_n\},\ X_i \leq y) \\
&\geq \mathbb{P}(\forall i \in \{1,\dots,N_n\},\ Z_i \leq y) \qquad (2) \\
&\geq \mathbb{P}(\forall i \in \{1,\dots,n\},\ Z_i \leq y) \qquad (3) \\
&\geq (1 - 2^{-\lfloor y \rfloor})^n.
\end{aligned}$$

Inequality (2) hold because of (1) and inequality (3) because $N_n \leq n$. Taking $y = A\log_2 n$, we obtain easily the announced upper bound.

To come back to Lyndon words, note that for a primitive word w, we have $M_n(\rho(w)) \leq \max\{M_n(w), X_1(w) + X_{N_n}(w)\}$. Thus we obtain:

$$\begin{aligned}
& \mathbb{P}_n\left(M_n(\rho(w)) \geq A\log_2 n\right) \\
\leq\ & \mathbb{P}_n\left(M_n(w) \geq A\log_2 n\right) + \mathbb{P}_n\left(X_1 + X_{N_n} \geq A\log_2 n\right) \\
\leq\ & \mathbb{P}_n\left(A\log_2 n \geq M_n(w)\right) + \mathbb{P}_n\left(X_1 \geq \frac{A}{2}\log_2 n\right) + \mathbb{P}_n\left(X_{N_n} \geq \frac{A}{2}\log_2 n\right).
\end{aligned}$$

Note that X_1 and X_{N_n} have the same law. Thus using (1), we get:

$$\mathbb{P}_n\left(X_1 \geq \frac{A}{2}\log_2 n\right) \leq \mathbb{P}_n\left(Z_1 \geq \frac{A}{2}\log_2 n\right) = O(n^{-\frac{A}{2}}) = o\left(n^{1-A}\right),$$

by the choice we made for A. Finally, by using Lemma 2.2, we get the desired result. ■

4. Building of long blocks and short blocks of a word

Let $0 < \varepsilon < 1$ and $B > 2$ be fixed in this section. Our aim here is to find, in a word $w \in \mathcal{A}^n$, some *long blocks* beginning by a long run of "a" in a word $w \in \mathcal{A}^n$, and we moreover want to choose them long enough to be distinct with high probability. We study then the positions of these long blocks along the word w. Here is our definition:

Definition 4.1 (Long blocks)**.** *Let w be a word with length n. The long blocks of w are the subwords of w that:*

- *begin with a run of "a" with length equal or greater than $(1-\varepsilon)\log_2 n$,*
- *end with a run of "b"*
- *have the smallest possible length larger than $3\log_2 n$.*

We denote by H_n the number of long blocks.

The next lemma estimates the number of long blocks for a random Lyndon word. Note that although the crude estimate we give could be sharpen, it is sufficient for our mean.

Lemma 4.2 (Number of long blocks)**.** *There exists a constant $D > 0$ such that*

$$\tilde{\mathbb{P}}_n\left(\frac{1}{4}n^\varepsilon \le H_n \le \frac{9}{4}n^\varepsilon\right) \quad \ge \quad 1 - O\left(\exp\left(-Dn^\varepsilon\right)\right).$$

Proof. We begin once again by proving the inequality for the probability measure $\mathbb{P}_n$ on the set $\mathcal{A}^n$. Set, for $i \ge 1$, $B_i = \mathbf{1}_{\{Z_i \ge (1-\varepsilon)\log_2 n\}}$. Then $(B_i)_{i\ge 1}$ are independent identically distributed Bernoulli random variables with parameter $p_{n,\varepsilon}$, which satisfies $n^{\varepsilon-1} \le p_{n,\varepsilon} \le 2n^{\varepsilon-1}$. Note that

$$\sum_{1\le 2i-1\le N_n-1} B_{2i-1}(w) \le H_n(w) \le \sum_{i=1}^n B_i(w) \qquad \text{if } w_1 = a,$$

$$\sum_{1\le 2i\le N_n-1} B_{2i}(w) \le H_n(w) \le \sum_{i=1}^n B_i(w) \qquad \text{if } w_1 = b.$$

Therefore, by large deviation results for sums of independent Bernoulli random variables (see for instance the book by Bollobas [5, Th. 7, p.13]), there exists $D_1 > 0$ such that:

$$\begin{aligned}\mathbb{P}_n\left(H_n \ge \frac{9}{4}n^\varepsilon\right) &\le \mathbb{P}\left(\sum_{i=1}^n B_i \ge \frac{9}{4}n^\varepsilon\right)\\ &\le \mathbb{P}\left(\sum_{i=1}^n B_i - n\mathbb{E}B_1 \ge \frac{1}{4}n^\varepsilon\right) \le O\left(\exp\left(-D_1 n^\varepsilon\right)\right).\end{aligned}$$

In the same manner, by looking only to the B_i's with odd indices (when the word begins with "a") or only to the B_i's with even indices (when the word begins with "b") and using lemma 3.3 in which $\gamma = 1$, we obtain the existence of $D_2 > 0$ such

that:

$$\begin{aligned}
&\mathbb{P}_n\left(H_n \leq \frac{1}{4}n^{\varepsilon}\right) \\
\leq\ & \mathbb{P}\left(N_n < \frac{n}{2}\left(1-n^{-1/2+\varepsilon}\right)\right) + \mathbb{P}\left(\sum_{i=1}^{n/2\left(1-n^{-1/2+\varepsilon}\right)} B_i \leq \frac{1}{4}n^{\varepsilon}\right) \\
\leq\ & O\left(\exp\left(-C_1 n^{2\varepsilon}\right)\right) + O\left(\exp\left(-D_2 n^{\varepsilon}\right)\right).
\end{aligned}$$

This proves the lemma for random words.

For random Lyndon words, note that if w is a primitive word, then $H_n(w) \leq H_n(\rho(w)) \leq H_n(w)+1$ and use lemma 2.2. ■

The length of the long blocks has been chosen large enough to ensure that two long blocks are distinct with high probability:

Lemma 4.3 (Inequality of long blocks)**.** *Denote by E_n the event that a word with length n has at least two equal disjoint subwords with length at least $3\log_2 n$. Then:*

$$\tilde{\mathbb{P}}_n(E_n) \leq O\left(n^{-1}\right).$$

Proof. We begin as usual with random words. By counting the number of possible subwords with length $3\log_2 n$ and their possible positions, we have:

$$\mathbb{P}_n\left(E_n\right) \leq O\left\{n^2(2^{3\log_2 n})(2^{-3\log_2 n})^2\right\} \leq O\left(n^{-1}\right).$$

Lemma 2.2 gives the same estimate for Lyndon words. ■

We also want that the long blocks do not overlap with high probability, or, in other words, that the beginnings of long blocks are far away enough with high probability. This is ensured by the next lemma:

Lemma 4.4 (Minimal distance between beginnings of long blocks)**.** *Let D_n be the event that there exist at least two long blocks which begin at a distance less than $8\log_2 n$. Then:*

$$\tilde{\mathbb{P}}_n(D_n) \leq O(n^{-(1-2\varepsilon)}\log_2 n).$$

Proof. As usual, we start with the case of random words:

$$\mathbb{P}_n(D_n) \quad \leq \quad \mathbb{P}_n\left(H_n \geq \frac{9}{4}n^{\varepsilon}\right) + \mathbb{P}_n\left(D_n \cap \left\{H_n < \frac{9}{4}n^{\varepsilon}\right\}\right)$$

Let us denote by F_n the last event. On F_n, at least one of the H_n subwords with length $8\log_2 n$ starting just after a run of "a" with length at least $(1-\varepsilon)\log_2 n$ must admit a subword of "a" with length $(1-\varepsilon)\log_2 n$ (which is the beginning of the next long block). By an estimate analogous to the one used in the previous lemma,

$$\mathbb{P}_n\left(F_n\right) \leq O\left\{\frac{9}{4}n^{\varepsilon}8\log_2 n 2^{-(1-\varepsilon)\log_2 n}\right\} = O(n^{-(1-2\varepsilon)}\log_2 n).$$

By lemma 4.2, the first term is negligible, and Lemma 2.2 concludes for Lyndon words.

Now we consider the set of "good" Lyndon words that satisfy all the previous properties:

Definition 4.5 (Good Lyndon words)**.** *Denote by $\mathcal{G}_n$ the set of Lyndon words w satisfying the following conditions:*

- *the maximal run of "a" satisfies $(1-\varepsilon)\log_2 n \leq M_n^a \leq 2\log_2 n$*
- *the maximal run satisfies $(1-\varepsilon)\log_2 n \leq M_n \leq 2\log_2 n$*
- *the number of long blocks satisfies $\frac{1}{4}n^\varepsilon \leq H_n \leq \frac{9}{4}n^\varepsilon$;*
- *the beginnings of long blocks are at a distance at least $8\log_2 n$, in the sense $\mathcal{G}_n \subset D_n^c$;*
- *the word w has no equal long blocks, in the sense $\mathcal{G}_n \subset E_n^c$.*

Note that on $\mathcal{G}_n$, the length of a long block is less than $3\log_2 n + 2 \times 2\log_2 n = 7\log_2 n$, and that there is no overlapping between two long blocks . The next lemma ensures that a large proportion of Lyndon words are good Lyndon words:

Lemma 4.6. *For every n large enough:*

$$\tilde{\mathbb{P}}_n(\mathcal{G}_n) \geq 1 - O\left(n^{-(1-2\varepsilon)}\log_2 n\right).$$

Proof. Everything has been proved in the previous lemmas 3.5, 3.6 (in which $A = 2$), 4.2, 4.3 and 4.4. ■

Now, we note that a good Lyndon word $w \in \mathcal{G}_n$ begins with a long block, ends with a run of "b", and all portions between long blocks begin with a run of "a" and end with a run of "b". We can thus give the following definition of short blocks:

Definition 4.7 (Short blocks)**.** *For a good Lyndon word $w \in \mathcal{G}_n$, we cut each section stretching between two long blocks into short blocks, made of two consecutive runs of "a" and "b"(in this order).*

Note that short blocks have length equal or smaller to $4\log_2 n$.

5. Permutations of blocks for good Lyndon words

In the previous section, we have cut any good Lyndon word w into blocks beginning with a run of "a" and ending with a run of "b": the long ones and and the short ones. The long ones correspond to long runs of "a", and the first long block (at the beginning of w) is, by definition of a Lyndon word, the smallest block for the lexicographic order. We are going to see that we can keep this first long block of w at the beginning of the word and permute the other blocks, without changing the distribution on the set of good Lyndon words.

In the following, "short" and "long" refer to the type of blocks, while "small" and "large" refer to the lexicographic order on words.

Definition 5.1 (Permutation of blocks for good Lyndon words)**.** *Consider $w \in \mathcal{G}_n$.*

1. We denote by $K_n(w)$ the total number of blocks, long and short, of w.

2. We denote by $(Y_i(w))_{0\leq i\leq K_n(w)-1}$ the blocks of w in their order of appearance along w. Certainly, the first block $Y_0(w)$ is the smallest block among all blocks of w.

3. Let $j_0(w)$ be the index of the second smallest block of w.

4. We denote by $\mathfrak{S}_{K_n(w)-1}$ the set of permutations of $\{1,\dots,K_n(w)-1\}$, and define

$$\sigma.w = Y_0(w)Y_{\sigma(1)}(w)\dots Y_{\sigma(K_n(w)-1)}(w),$$

for $\sigma \in \mathfrak{S}_{K_n(w)-1}$. Obviously, $\sigma.w \in \mathcal{G}_n$.

5. We define also $C(w) = \{\sigma.w, \sigma \in \mathfrak{S}_{K_n(w)-1}\}$, the set of all words which are obtained by the all the permutations of the blocks of w.

The two cases of right factor exposed in the introduction can be rephrased in the following manner: either the standard right factor is obtained by deleting the first "a", or it begins by the second smallest block of w, $Y_{j_0(w)}$.

In this section, we study the asymptotics of the position between the two smallest blocks, and we will rephrase this result in terms of standard right factor in the next section.

Our main tool is the immediate following property:

Lemma 5.2 (Invariance in law under the permutations of blocks)**.** *Let w_0 be a fixed good Lyndon word. Consider the set $\mathfrak{S}_{K_n(w_0)-1} \times C(w_0)$, endowed with the uniform probability. Then the random variable*

$$\begin{array}{rccc} W_{w_0}: & \mathfrak{S}_{K_n(w_0)-1} \times C(w_0) & \longrightarrow & C(w_0) \\ & (\sigma, w) & \longmapsto & \sigma.w \end{array}$$

follows the uniform law on $C(w_0)$.

Proof. It is sufficient to note that, by construction, each word in $C(w_0)$ has the same family of blocks. ■

Thus, roughly speaking, the second smallest block Y_{j_0} has the same probability to be at every possible place among all the blocks, and this is why its position along the word w, divided by n, should follows approximately the uniform law on $[0, 1]$. To formalize this intuition and to exploit this invariance property, we enlarge our probability space $\mathcal{G}_n$: consider a sequence $(U_i)_{i\in\mathbb{N}}$ of independent identically distributed random variables on a probability space $(X, \mathcal{X}, \mathbb{Q})$, following the uniform distribution on $[0, 1]$. We denote by $\bar{\mathbb{P}}_n$ the uniform probability on $\mathcal{G}_n$ and consider the product probability $\mathbb{Q} \otimes \bar{\mathbb{P}}_n$ on the product space $X \times \mathcal{G}_n$; this means that $(U_i)_{i\in\mathbb{N}}$ are independent of the choice of the random Lyndon word in $\mathcal{G}_n$.

Definition 5.3 (Random permutation)**.** *For $x \in X$, we define a uniform random permutation $\pi_x \in \mathfrak{S}_{K_n(w)-1}$ by the order statistics of $U_1(x), U_2(x), ..., U_{K_n(w)-1}(x)$:*

$$U_{\pi_x(1)}(x) < U_{\pi_x(2)}(x) < \cdots < U_{\pi_x(K_n(w)-1)}(x).$$

Therefore from the previous lemma, the random variable

$$W: \left\{ \begin{array}{rcl} X \times \mathcal{G}_n & \longrightarrow & \mathcal{G}_n \\ (x, w) & \longmapsto & \pi_x.w \end{array} \right.$$

follows the uniform law on $\mathcal{G}_n$. We can now study, under the uniform probability $\bar{\mathbb{P}}_n$ on $\mathcal{G}_n$, the position of the second smallest block defined by

$$d_n(w) = \frac{1}{n} \sum_{i=0}^{j_0(w)-1} |Y_i(w)|.$$

Here $|w|$ denotes the length of the word w. Thus, the random variable

$$\begin{array}{rccc} d_n o W: & X \times \mathcal{G}_n & \longrightarrow & [0, 1] \\ & (x, w) & \longmapsto & d_n(\pi_x.w) \end{array}$$

has the same law as the random variable d_n under $\bar{\mathbb{P}}_n$.

We will thus focus on this new random variable to use the property of invariance under the permutation of blocks. Remember that the convergence in L^2 implies the convergence in probability; thus, in the following, the notation $\|.\|_{\mathbb{P}}$ will denotes the L^2-norm associated to a probability measure $\mathbb{P}$.

Theorem 5.4 (Position of the second smallest block)**.** *We have:*

$$\|d_n oW - U_{j_0}\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \leq O\left(\left[\frac{\log_2 n}{n}\right]^{1/2}\right).$$

This implies in particular that the law of d_n under $\bar{\mathbb{P}}_n$ converges weakly to the uniform law on $[0,1]$ and that every moment of d_n converges to the corresponding moment of the uniform distribution.

Remark. Coming back to random words, this result implies that the normalized distance between the two smallest blocks (which roughly corresponds to the two largest runs of "a") asymptotically follows the uniform law on $[0,1]$.

Proof. We have:

$$\begin{aligned} d_n oW(x,w) = d_n(\pi_x.w) &= \frac{1}{n}\left(|Y_0(w)| + \sum_{j<\pi_x^{-1}(j_0(w))} |Y_{\pi_x(j)}(w)|\right) \\ &= \left(\frac{1}{n}|Y_0(w)| + \frac{1}{n}\sum_{j=1}^{K_n(w)-1} |Y_j(w)|\mathbf{1}_{\{U_j(x)<U_{j_0(w)}(x)\}}\right). \end{aligned}$$

By conditioning on w and $U_{j_0(w)}$, and using the fact that $\sum_i |Y_i| = n$, we obtain:

$$\begin{aligned} &\mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0(w)}\right) \\ = &\left(\frac{1}{n}|Y_0(w)| + \frac{1}{n}\sum_{\substack{1\leq j\leq K_n(w)-1 \\ j\neq j_0(w)}} |Y_j(w)|U_{j_0(w)}\right) \\ = &\left(U_{j_0(w)} + (1-U_{j_0(w)})\frac{|Y_0(w)|}{n} - U_{j_0(w)}\frac{|Y_{j_0(w)}(w)|}{n}\right). \end{aligned}$$

On $\mathcal{G}_n$, $|Y_0|$ and $|Y_{j_0}|$ are bounded by $7\log_2 n$, so

$$\begin{aligned} &\left\|\mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0(w)}\right) - U_{j_0(w)}\right\|^2_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\ = &\ \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left[\mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0(w)}\right) - U_{j_0(w)}\right]^2 \\ = &\ \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left[\left(1-U_{j_0(w)}\right)\frac{|Y_0(w)|}{n} - U_{j_0(w)}\frac{|Y_{j_0(w)}(w)|}{n}\right]^2 \\ \leq &\ \left[\frac{14\log_2 n}{n}\right]^2, \end{aligned}$$

which tends to 0 when n goes to infinity. Now,

$$\begin{aligned}
&\left\|d_n(\pi_x.w) - \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0(w)}\right)\right\|^2_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\
= \;& \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left[d_n(\pi_x.w) - \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0(w)}\right)\right]^2 \\
= \;& \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left[\sum_{j=1}^{K_n(w)-1}\frac{|Y_j(w)|}{n}\mathbf{1}_{\{U_j<U_{j_0(w)}\}} - \sum_{\substack{1\le j\le K_n(w)-1\\ j\neq j_0(w)}}\frac{|Y_j(w)|}{n}U_{j_0(w)}\right]^2 \\
= \;& \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left[\sum_{\substack{1\le j\le K_n(w)-1\\ j\neq j_0(w)}}\frac{|Y_j(w)|}{n}\left(\mathbf{1}_{\{U_j<U_{j_0(w)}\}} - U_{j_0(w)}\right)\right]^2 \\
(4)\le \;& \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left[\frac{1}{4}\sum_{\substack{1\le j\le K_n(w)-1\\ j\neq j_0(w)}}\left(\frac{|Y_j(w)|}{n}\right)^2\right] \\
(5)\le \;& \frac{7\log_2 n}{4n},
\end{aligned}$$

which tends to 0 when n goes to infinity. To obtain inequality (4), we conditioned first on w and $U_{j_0(w)}$; for inequality (5), we used the facts that, on $\mathcal{G}_n$, all blocks have length smaller than $7\log_2 n$ and that $\sum_i |Y_i| = n$.

Consequently,

$$\begin{aligned}
\|d_n oW - U_{j_0}\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \;\le\; & \left\|d_n(\pi_x.w) - \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0}(w)\right)\right\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\
& + \left\|\mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n}\left(d_n(\pi_x.w)|\, w, U_{j_0}(w)\right) - U_{j_0(w)}\right\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\
\le \;& O\left(\left[\frac{\log_2 n}{n}\right]^{1/2}\right).
\end{aligned}$$

The convergence of the other moments is a consequence of the convergence in law, as d_n is bounded by 1. ■

6. Limit distribution of the standard right factor

In this section, we establish the convergence of the distribution of the normalized length of the standard right factor of a random Lyndon word and give the limit distribution, which follows quite easily from the result of the previous section. Remember that $\tilde{\mathbb{P}}_n$ is the uniform probability on the set $\mathcal{L}_n$ of Lyndon words with length n and that $\bar{\mathbb{P}}_n$ is the uniform probability on the set $\mathcal{G}_n$ of good Lyndon words with length n. The length of the standard right factor of $w \in \mathcal{L}_n$ is denoted by $R_n(w)$, and we introduce the normalized length of the standard right factor $r_n(w) = R_n(w)/n$.

Theorem 6.1. *As n goes to infinity, r_n converges in distribution to*

$$\mu(dx) = \frac{1}{2}\delta_1(dx) + \frac{1}{2}\mathbf{1}_{[0,1)}(x)dx,$$

where δ_1 denotes the Dirac mass at point 1, and dx Lebesgue's measure on $\mathbb{R}$. All the moments of r_n also converge to the corresponding moments of the limit distribution.

Proof. First, we split the set $\mathcal{L}_n$ in two parts, corresponding to the two cases of the introduction:

- $\mathcal{L}_n^1 = a\mathcal{L}_{n-1}$ contains exactly the Lyndon words w whose standard right factor is obtained by deleting the first "a" of the word and has thus normalized length $r_n(w) = (n-1)/n$. Note that
$$\tilde{\mathbb{P}}_n(\mathcal{L}_n^1) = \frac{card(\mathcal{L}_{n-1})}{card(\mathcal{L}_n)} \sim \frac{1}{2}.$$
- $\mathcal{L}_n^2 = \mathcal{L}_n \backslash \mathcal{L}_n^1$ contains exactly the Lyndon words w whose standard right factor has normalized length $r_n(w)$ strictly smaller than $(n-1)/n$.

Now, forgetting the "bad" Lyndon words, using the inequality $r_n \leq 1$ and lemma 4.6, we obtain the following inequality:

$$\|r_n - r_n \mathbf{1}_{\mathcal{G}_n}\|_{\tilde{\mathbb{P}}_n} \leq (1 - \tilde{\mathbb{P}}_n(\mathcal{G}_n))^{1/2} \leq O\left(n^{-\frac{1-2\varepsilon}{2}} \sqrt{\log_2 n}\right). \tag{6}$$

But for $w \in \mathcal{L}_n^2 \cap \mathcal{G}_n$, the standard right factor begins with the second smallest block Y_{j_0} of w. Thus, in this case, with the notations of the previous section: $r_n(w) = 1 - d_n(w)$. Moreover, $\mathcal{L}_n^1 \cap \mathcal{G}_n$ and $\mathcal{L}_n^2 \cap \mathcal{G}_n$ are stable under the permutations of blocks. Thus with the same setting as in Theorem 5.4,

$$\begin{aligned} r_n(x,w)\mathbf{1}_{\mathcal{G}_n}(w) &= \left((1 - d_n(\pi_x.w))\mathbf{1}_{\mathcal{L}_n^2}(w) + \frac{n-1}{n}\mathbf{1}_{\mathcal{L}_n^1}(w)\right)\mathbf{1}_{\mathcal{G}_n}(w) \\ &\overset{\text{law}}{=} r_n(w)\mathbf{1}_{\mathcal{G}_n}(w), \end{aligned} \tag{7}$$

where the right hand side is a random variable from $X \times \mathcal{L}_n$, endowed with $\mathbb{Q} \otimes \tilde{\mathbb{P}}_n$ and the left hand side is from $\mathcal{L}_n$, endowed with the uniform probability $\tilde{\mathbb{P}}_n$. Keeping in mind the result of the previous theorem, we introduce for $(x,w) \in X \times \mathcal{L}_n$,

$$s_n(x,w) = \left((1 - U_{j_0(w)}(x))\mathbf{1}_{\mathcal{L}_n^2}(w) + \mathbf{1}_{\mathcal{L}_n^1}(w)\right)\mathbf{1}_{\mathcal{G}_n}(w).$$

Now, by Theorem 5.4,

$$\begin{aligned} & \|r_n(x,w)\mathbf{1}_{\mathcal{G}_n}(w) - s_n(x,w)\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\ \leq\ & \|r_n(x,w)\mathbf{1}_{\mathcal{G}_n}(w) - s_n(x,w)\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\ \leq\ & \left\|\left(r_n(x,w) - (1 - U_{j_0(w)}(x))\right)\mathbf{1}_{\mathcal{L}_n^2}(w) - \frac{1}{n}\mathbf{1}_{\mathcal{L}_n^1}(w)\right\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \\ \leq\ & \left\|(d_n(x,w) - U_{j_0(w)}(x))\right\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} + \frac{1}{n}\bar{\mathbb{P}}_n(\mathcal{L}_n^1)^{1/2} \\ \leq\ & O\left(\left[\frac{\log_2 n}{n}\right]^{1/2}\right). \end{aligned} \tag{8}$$

Note that the position of Y_{j_0}, the second smallest block of w in $\pi_x.w$, is governed by U_{j_0}, which is clearly a uniform random variable on $[0,1]$, independent of w. Note also that, thanks to lemma 4.6, $\bar{\mathbb{P}}_n(\mathcal{L}_n^1 \cap \mathcal{G}_n) \sim 1/2$, and then $\bar{\mathbb{P}}_n(\mathcal{L}_n^2 \cap \mathcal{G}_n) \sim 1/2$: consequently, the law of s_n under $\mathbb{Q} \otimes \tilde{\mathbb{P}}_n$ converges weakly to $\mu(dx) = \frac{1}{2}[\delta_1(dx) + \mathbf{1}_{[0,1]}(x)dx]$.

Now, as $\|r_n(x,w)\mathbf{1}_{\mathcal{G}_n}(w) - s_n(x,w)\|_{\mathbb{Q}\otimes\tilde{\mathbb{P}}_n}$ goes to 0 by (8), a classical result (see for instance the book by Billingsley [3, Th. 4.2, p.25] in the first edition) ensures that the distribution of $r_n\mathbf{1}_{\mathcal{G}_n}$, as a random variable on $X \times \mathcal{L}_n$, also converges to $\mu(dx) = \frac{1}{2}[\delta_1(dx) + \mathbf{1}_{[0,1]}(x)dx]$. Using (7), the distribution of $r_n\mathbf{1}_{\mathcal{G}_n}$, as a random variable on $\mathcal{L}_n$, also converges to the same limit. Finally, (6) ensures the convergence of the distribution of r_n to the same limit. ■

7. Generalization to the case of q letters

In this section, we generalize the previously obtained results to the case of a totally ordered alphabet with q letters: $\mathcal{A} = \{a_1, a_2, \ldots, a_q\}$, $q \in \{2, 3, 4, ...\}$ and $a_1 < a_2 < \cdots < a_q$. All the technics developed for the simple case of two letters can be readily adapted in this context and we just give the results and some indications for the adaptations needed.

1. Denote by $\mathcal{A}^n$ the set of words with length n and by $\mathcal{L}_n = \mathcal{L}_n(\{a_1, a_2, \ldots, a_q\})$ the subset of Lyndon words. The probability measures $\mathbb{P}_n$ and $\tilde{\mathbb{P}}_n$ are defined as before. As previously, we have:

$$card(\mathcal{L}_n) = \frac{q^n}{n}(1 + O(q^{-n/2})).$$

The link between random Lyndon words and random words still holds: if $A \subset \mathcal{L}_n$, we have:

$$\mid \tilde{\mathbb{P}}_n(A) - \mathbb{P}_n(\rho^{-1}(A)) \mid \leq O(q^{-n/2}).$$

2. Let w be a word in $\mathcal{A}_n$. As previously, we define its runs, its number of runs $N_n(w)$ and the length of these runs $X_1(w), \ldots, X_{N_n}(w)$. To build these random variables, we introduce a family $(Z_i)_{i\in\mathbb{N}}$ of independent identically distributed geometrical random variables with parameter $(q-1)/q$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and $(\epsilon_i)_{i\in\mathbb{N}}$, a family of independent and identically distributed random variables with uniform distribution on $\{a_1, a_2, ..., a_q\}$, and independent of the Z_i's. To obtain a random sequence of letters, do the following:

- Select the letter ϵ_1, and write a run of Z_1 such letters.
- Select the letter ϵ_2 conditioned to be distinct of ϵ_1, and write a run of Z_2 such letters.
- Proceed by recurrence: select the letter ϵ_{n+1} conditioned to be distinct of ϵ_n, and write a run of Z_{n+1} such letters.

Truncating the n first letters gives a random variable defined on Ω and uniformly distributed on $\mathcal{A}_n$. As previously,

$$N_n(\omega) = \inf\left\{k \in \mathbb{N},\ \sum_{i=1}^{k} Z_i(\omega) \geq n\right\}.$$

3. Estimate the number of runs by using the fact that $N_n - 1$ follows a binomial law with parameters $\left(n-1, \frac{q-1}{q}\right)$ as in lemma 3.3.

4. Estimate the length M_n of the longest run and the length $M_n^{a_1}$ of the largest run of a_1 as in lemma 3.5 and lemma 3.6 by using the same estimates on geometrical laws. The typical order of M_n and $M_n^{a_1}$ is $\log_q n$.

5. Define the long blocks:

Definition 7.1. *Let w be a word with length n. The long blocks of w are the subwords of w that:*

- *begin with a run of "a_1" with length equal or greater than* $(1-\varepsilon)\log_2 n$,
- *end just before an other run of "a_1" (and consequently end with a run of a letter distinct from "a_1"*
- *have the smallest possible length larger than* $3\log_q n$.

Their number H_n is, as in lemma 4.2, of order n^ε. To prove this, introduce, for $i \geq 1$, the variable $B_i = \mathbf{1}_{\{Z_i \geq (1-\varepsilon)\log_q n\}}$. Then the $(B_i)_{i\geq 1}$ are independent identically distributed Bernoulli random variables with parameter $p_{n,\varepsilon}$ satisfying $n^{\varepsilon-1} \leq p_{n,\varepsilon} \leq qn^{\varepsilon-1}$, thus we can have large deviation results.

6. We verify then that the long blocks do not overlap too often and are distinct with high probability, as in lemmas 4.3 and 4.4. Good Lyndon words are defined in the same manner as previously. Define the short blocks:

Definition 7.2. *For a good Lyndon word $w \in \mathcal{G}_n$, we cut each section stretching between two long blocks into short blocks, that begin with a run of 'a_1" and end just before the next run of "a_1".*

7. All is thus in place to permute the blocks as previously. With the same setting as before, we obtain:

Theorem 7.3. *We have:*

$$\|d_n oW - U_{j_0}\|_{\mathbb{Q}\otimes\bar{\mathbb{P}}_n} \leq O\left(\left[\frac{\log_q n}{n}\right]^{1/2}\right).$$

This implies in particular that the law of d_n under $\bar{\mathbb{P}}_n$ converges weakly to the uniform law on $[0,1]$ and that every moment of d_n converges to the corresponding moment of the limit law

8. To conclude for the length of the right factor, we split the set $\mathcal{L}_n$ in two parts:
   - $\mathcal{L}_n^1 = a_1\mathcal{L}_{n-1}(\{a_1, a_2, \ldots, a_q\}) \cup a_2\mathcal{L}_{n-1}(\{a_2, \ldots, a_q\}) \cup \cdots \cup a_{q-1}\mathcal{L}_{n-1}(\{a_{q-1}, a_q\})$ contains exactly the Lyndon words w whose standard right factor is obtained by deleting the first letter of the word and has thus normalized length $r_n(w) = (n-1)/n$,
   - $\mathcal{L}_n^{q,2} = \mathcal{L}_n^q \backslash \mathcal{L}_n^{q,1}$ contains exactly the Lyndon words w whose standard right factor has normalized length $r_n(w)$ strictly smaller than $(n-1)/n$.

The only difference is that

$$\tilde{\mathbb{P}}_n(\mathcal{L}_n^1) = \frac{card(\mathcal{L}_{n-1}(\{a_1, a_2, \ldots, a_q\})) + \cdots + card(\mathcal{L}_{n-1}(\{a_{q-1}, a_q\}))}{card(\mathcal{L}_n)} \sim \frac{1}{q},$$

which gives Theorem 1.1.

Acknowledgements

We wish to thank Philippe Chassaing for pointing this problem to us, and for many fruitful conversations.

References

[1] G. E. Andrews. *The theory of partitions.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1998. Reprint of the 1976 original.

[2] F. Bassino, J. Clément, and C. Nicaud. The standard factorization of lyndon words: an average point of view. *submitted to Elsevier Sciences, available at* `http://www-igm.univ-mlv.fr/~ bassino/biblio.html`, 2003.

[3] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. Second edition.
[4] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1989.
[5] B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.
[6] K.-T. Chen, R. H. Fox, and R. C. Lyndon. Free differential calculus. IV. The quotient groups of the lower central series. *Ann. of Math. (2)*, 68:81–95, 1958.
[7] P. Erdős and A. Rényi. On a new law of large numbers. *J. Analyse Math.*, 23:103–111, 1970.
[8] P. Erdős and P. Révész. On the length of the longest head-run. In *Topics in information theory (Second Colloq., Keszthely, 1975)*, pages 219–228. Colloq. Math. Soc. János Bolyai, Vol. 16. North-Holland, Amsterdam, 1977.
[9] P. Hitzenko and G. Louchard. Distinctness of compositions of an integer: A probabilistic analysis. *Random struct. Alg.*, 19:407–437, 2001.
[10] M. Lothaire. *Combinatorics on words*. Cambridge Mathematical Library.
[11] R. C. Lyndon. On Burnside's problem. *Trans. Amer. Math. Soc.*, 77:202–215, 1954.
[12] J. Pitman. Combinatorial stochastic processes. *Technical report*, 621, 2002.
[13] S. I. Resnick. *Extreme values, regular variation, and point processes*, volume 4 of *Applied Probability. A Series of the Applied Probability Trust*. Springer-Verlag, New York, 1987.

Institut Elie Cartan Nancy (mathématiques), Université Henri Poincaré Nancy 1, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex France
E-mail address: Regine.Marchand@iecn.u-nancy.fr, Elahe.Zohoorian@iecn.u-nancy.fr