HAL
open science

# Asymptotic behavior of some factorizations of random words

Elahe Zohoorian Azad, Philippe Chassaing

## To cite this version:

Elahe Zohoorian Azad, Philippe Chassaing. Asymptotic behavior of some factorizations of random words. Random Structures and Algorithms, In press, 10.1002/rsa.21073 . hal-00475379v2

**HAL Id: hal-00475379**

**https://hal.science/hal-00475379v2**

Submitted on 4 Nov 2021

**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.

**Noname manuscript No.**
(will be inserted by the editor)

# Asymptotic behavior of some factorizations of random words

**Elahe Zohoorian Azad · Philippe Chassaing**

the date of receipt and acceptance should be inserted later

**Abstract** In this paper we consider the normalized lengths of the factors of some factorizations of random words. First, for the *Lyndon factorization* of finite random words with $n$ independent letters drawn from a finite or infinite totally ordered alphabet according to a general probability distribution, we prove that the limit law of the normalized lengths of the smallest Lyndon factors is a variant of the stickbreaking process. Convergence of the distribution of the lengths of the longest factors to a Poisson-Dirichlet distribution follows. Secondly we consider the *standard factorization* of random *Lyndon word* : we prove that the distribution of the normalized length of the standard right factor of a random $n$-letters long Lyndon word, derived from such an alphabet, converges, when $n$ is large, to:

$$\mu(dx) = p_1\delta_1(dx) + (1-p_1)\mathbf{1}_{[0,1)}(x)dx,$$

in which $p_1$ denotes the probability of the smallest letter of the alphabet.

**Keywords** Random word · Lyndon word · Standard right factor · Longest run · Poisson-Dirichlet distribution

## Contents

1 Introduction . . . . . 2

E. Zohoorian Azad
School of mathematics and computer sciences
Damghan University
P.O.Box 36715-364
Damghan, Iran

P. Chassaing
Institut Elie Cartan de Lorraine
Université de Lorraine
Campus Scientifique, BP 239
54506 Vandoeuvre-lès-Nancy Cedex France

1.1 Random words and random Lyndon words . . . 4
1.2 Main results . . . 4
1.3 Context . . . 6
2 Sketch of proofs . . . 7
2.1 Lyndon factorization and runs . . . 7
2.2 A refined factorization . . . 8
2.3 The factorization shuffle . . . 9
3 Statistical properties of runs, under $\mathbb{P}_n$ or under $\mathbb{L}_n$ . . . 12
4 Long blocks of words and good words . . . 13
5 Proof of Theorem 1.2 . . . 16
6 Proof of Theorem 1.4 . . . 21
7 Runs statistics: proofs . . . 27
7.1 Asymptotically almost sure properties in $\mathcal{A}^n$ vs $\mathcal{P}_n$: proof of Lemma 3.1 . . . 27
7.2 Alternative representations for $\mathbb{P}_n$ . . . 28
7.3 Number of runs: proof of Lemma 3.2 . . . 29
7.4 Number of long runs : proof of Lemma 3.3 . . . 30
7.5 Large values of the longest runs : proof of Lemma 3.4 . . . 31
7.6 $\Lambda_n$ is large : proof of Proposition 5.1 . . . 32

## 1 Introduction

In this paper we address the statistical properties of two well studied factorizations related to Lyndon words: the *Lyndon factorization* of a word, and the *standard factorization* of a Lyndon word. Applications of these two factorizations are discussed in [BP07, Section 4.2]. Let us recall some notations and definitions from [Lot97, Reu93] for readability. For an ordered alphabet $\mathcal{A} = \{\mathtt{a}_1 \prec \mathtt{a}_2 \prec \dots\}$, finite or infinite, $\mathcal{A}^n$ is the set of $n$-letters words, and the *language*, i.e. the set of finite words, is

$$\mathcal{A}^\star = \{\emptyset\} \cup \mathcal{A} \cup \mathcal{A}^2 \cup \mathcal{A}^3 \cup \dots,$$

while $\mathcal{A}^+$ denotes $\mathcal{A}^\star \setminus \{\emptyset\}$. The length of a word $\mathtt{w} \in \mathcal{A}^\star$ is denoted by $|\mathtt{w}|$. The language $\mathcal{A}^\star$ is endowed with an operation, the *concatenation* $\mathtt{uv} = \mathtt{w}$ of two words $\mathtt{u}$ and $\mathtt{v}$, that is also a *factorization* of $\mathtt{w}$. A word $\mathtt{v}$ is a *factor* of a word $\mathtt{w}$ if there exists two other words $\mathtt{p}$ and $\mathtt{s}$, possibly empty, such that $\mathtt{w} = \mathtt{pvs}$. If $\mathtt{p}$ (resp. $\mathtt{s}$) is empty, $\mathtt{v}$ is a prefix (resp. a suffix) of $\mathtt{w}$. If $|\mathtt{p}| = k-1$ and $|\mathtt{pv}| = \ell$, $\mathtt{w}_{[k,\ell]}$ denotes the factor $\mathtt{v}$ of the word $\mathtt{w}$.

The total order, $\prec$, on the alphabet $\mathcal{A}$, induces a corresponding *lexicographic order*, again denoted by $\prec$, on $\mathcal{A}^+$: the word $\mathtt{v}$ is smaller than the word $\mathtt{w}$ (for the lexicographic order, $\mathtt{v} \prec \mathtt{w}$) at one of the following conditions: either $\mathtt{v}$ is a proper prefix of $\mathtt{w}$, or there exist words $\mathtt{p}$, $\mathtt{s}_1$, $\mathtt{s}_2$ in $\mathcal{A}^\star$ and letters $\mathtt{a} \prec \mathtt{b}$ in $\mathcal{A}$, such that $\mathtt{v} = \mathtt{pas}_1$ and $\mathtt{w} = \mathtt{pbs}_2$. For any factorization $\mathtt{w} = \mathtt{uv}$ of $\mathtt{w}$, $\mathtt{vu}$ is called a rotation of $\mathtt{w}$, and the set $\langle \mathtt{w} \rangle$ of rotations of $\mathtt{w}$ is called the *necklace* of $\mathtt{w}$. A word $\mathtt{w}$ is primitive if $|\mathtt{w}| = \#\langle \mathtt{w} \rangle$. In this case the necklace is said to be *aperiodic*. A word $\mathtt{v}$ is a factor of a necklace $\langle \mathtt{w} \rangle$ if $\mathtt{v}$ is a factor of some word $\mathtt{w}' \in \langle \mathtt{w} \rangle$.

The notion of *Lyndon word* has many equivalent definitions, to be found, for instance, in [Lot97].

**Definition 1** (Lyndon word)**.** *A word* $\mathtt{w} \in \mathcal{A}^+$ *is a Lyndon word if one of the 2 following (equivalent) conditions is satisfied :*

1. $\mathtt{w}$ *is primitive and is the smallest element of* $\langle \mathtt{w} \rangle$ *;*
2. $\mathtt{w}$ *is smaller than any proper suffix of* $\mathtt{w}$*.*

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

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

*but is not Lyndon;* $\mathtt{baac}$ *is not Lyndon, nor* $\mathtt{acba}$ *or* $\mathtt{cbaa}$*, but* $\mathtt{aacb}$ *is Lyndon. Here is an ordered list of Lyndon words* $\mathtt{w} \in \{\mathtt{a}, \mathtt{b}\} \cup \{\mathtt{a}, \mathtt{b}\}^2 \cup \{\mathtt{a}, \mathtt{b}\}^3$*:*

$$\mathtt{a} \prec \mathtt{aab} \prec \mathtt{ab} \prec \mathtt{abb} \prec \mathtt{b}.$$

Let $\mathfrak{L}$ denote the set of Lyndon words on the alphabet $\mathcal{A}$. Note that $\mathfrak{L} \subset \mathcal{A}^+$. A recursive characterization of Lyndon words is as follows:

**Definition 2** (and Proposition)**.** *One-letter words are Lyndon. A word* $\mathtt{w}$ *with length* $n \geq 2$ *is a Lyndon word if and only if there exists two Lyndon words* $u$ *and* $v$ *such that* $\mathtt{w} = \mathtt{uv}$ *and* $\mathtt{u} \prec \mathtt{v}$*. Among such factorizations of a Lyndon word* $\mathtt{w}$*, the factorization with the smallest* [1] *suffix* $\mathtt{v}$ *is called the* standard factorization*.*

**Example 2.** $\mathtt{0011} = (\mathtt{001})(\mathtt{1}) = (\mathtt{0})(\mathtt{011})$ *is a Lyndon word with two such factorizations. The latter is the standard factorization. Others examples of standard factorizations:* $\mathtt{aaabaab} = \mathtt{aaab.aab}$, $\mathtt{aaababb} = \mathtt{a.aababb}$, $\mathtt{aabaabb} = \mathtt{aab.aabb}$*.*

The *standard right factor* $v$ of a several letters-Lyndon word $\mathtt{w}$ is its smallest proper Lyndon suffix, but also its smallest proper suffix. The standard factorization of a Lyndon word is the first step in the construction of some basis of the free Lie algebra over $\mathcal{A}$, due to Lyndon [Lyn54] (see for instance [Lot97] or [Reu93]).

While the standard factorization has always 2 factors and applies to Lyndon words, the Lyndon factorization, useful for instance in data compression, see [GS12], has a variable number of factors, and applies to any word in $\mathcal{A}^+$. It is defined as follows :

**Theorem 1.1. and Definition** *[Chen, Fox and Lyndon, cf. [Lot97] Theorem 5.1.5] Any word* $\mathtt{w} \in \mathcal{A}^+$ *has a unique factorization as a nonincreasing product of Lyndon words, called the* Lyndon factorization *of* $\mathtt{w}$ *:*

$$\mathtt{w} = \mathtt{w}_\ell \mathtt{w}_{\ell-1} \dots \mathtt{w}_2 \mathtt{w}_1, \qquad \mathtt{w}_i \in \mathfrak{L}, \qquad \mathtt{w}_\ell \succeq \mathtt{w}_{\ell-1} \succeq \dots \succeq \mathtt{w}_2 \succeq \mathtt{w}_1.$$

Here, as opposed to the standard factorization, a Lyndon word factors trivially (it has only one factor).

[1] "Smallest" does not mean "shortest", actually it means "longest" here.

$\mathtt{w}$ = aabb.aaabbab.aaabb.a.a.a.a

**Fig. 1** A word $\mathtt{w}$ with 7 factors and a sequence of lengths :
$\rho^{(20)}(\mathtt{w}) = \frac{1}{20}\,(1,1,1,1,5,7,4,0,0\dots)$

### 1.1 Random words and random Lyndon words

In this paper, we study the asymptotic behavior of some natural statistics related to :

- the Lyndon factorization of a $n$-letters long random word chosen in $\mathcal{A}^n$ according to the probability distribution $\mathbb{P}_n$ ;
- the standard factorization of a $n$-letters long random Lyndon word chosen in $\mathfrak{L}_n$ according to the probability distribution $\mathbb{L}_n$.

Both $\mathbb{P}_n$ and $\mathbb{L}_n$ depend on a general probability distribution $(p_i)_{i\geq 1}$ on the finite or infinite alphabet $\mathcal{A} = \{\mathtt{a}_1 \prec \mathtt{a}_2 \prec \dots\}$, and we assume, without loss of generality, that $0 < p_1 < 1$. On the corresponding language $\mathcal{A}^\star$, we define the weight $p(\mathtt{w})$ of a word $\mathtt{w} = \mathtt{a}_{\ell_1}\mathtt{a}_{\ell_2}\dots\mathtt{a}_{\ell_n}$ as

$$p(\mathtt{w}) = p_{\ell_1}p_{\ell_2}\dots p_{\ell_n}.$$

The weight $p(.)$ defines a probability measure $\mathbb{P}_n$ on the set $\mathcal{A}^n$, through

$$\mathbb{P}_n(\{\mathtt{w}\}) = p(\mathtt{w}).$$

$\mathcal{P}_n$ (resp. $\mathcal{N}_n$, $\mathfrak{L}_n$) denotes the set of $n$-letters long primitive words (resp. its complement, resp. the set of $n$-letters long Lyndon words). Then we define a probability measure $\mathbb{L}_n$ on $\mathfrak{L}_n$, as follows

$$\mathbb{L}_n(\{\mathtt{w}\}) = \lambda_n p(\mathtt{w}),$$

in which $\lambda_n = 1/\mathbb{P}_n(\mathfrak{L}_n) = n/\mathbb{P}_n(\mathcal{P}_n)$. The probability measure $\mathbb{L}_n$ has a trivial extension to $\mathcal{A}^n$ (setting $\mathbb{L}_n\left(\mathfrak{L}_n^c\right) = 0$).

### 1.2 Main results

The sequence $\rho^{(n)}(\mathtt{w}) = (\rho_{i,n}(\mathtt{w}))_{i\geq 1}$ of normalized lengths of the Lyndon factors of a word $\mathtt{w} \in \mathcal{A}^n$, with Lyndon factorization $\mathtt{w} = \mathtt{w}_\ell\mathtt{w}_{\ell-1}\dots\mathtt{w}_1$, is defined as follows:

$$\rho_{i,n}(\mathtt{w}) = \begin{cases} \frac{|\mathtt{w}_i|}{n} & \text{if } \ 1 \leq i \leq \ell \\ 0 & \text{if } \ i > \ell. \end{cases}$$

Thus $\rho_{i,n}(\mathtt{w})$ denotes the normalized length of the $i$-th smallest Lyndon factor of $\mathtt{w}$ (remark that in this paper, when applied to words, e.g. to factors of words and necklaces, the adjectives "small" and "large" refer to the *lexicographic order* on words, while "short" and "long" refer to the *size*, or number of letters). Our first result describes the limit distribution, as $n$ grows, of the sequence $\rho^{(n)}(\mathtt{w}) = (\rho_{i,n}(\mathtt{w}))_{i\geq 1}$, seen as a random variable on $(\mathcal{A}^n, \mathbb{P}_n)$. We have:

**Theorem 1.2.** *For a totally ordered alphabet with probability distribution $p$ on its letters, $\rho^{(n)}$ converges in law, when $n \to \infty$, to the random sequence $\rho = (\rho_i)_{i\geq 1}$ whose law is defined by the law of $\rho_1$:*

$$\mu(dx) = p_1\delta_0(dx) + (1-p_1)\mathbf{1}_{(0,1]}(x)dx,$$

*and by the conditional distribution $\mu^y$ of $\rho_i$ given $(\rho_1, \rho_2, \ldots, \rho_{i-1})$, that only depends on $y = \rho_1 + \rho_2 + \cdots + \rho_{i-1}$, and is as follows :*

$$\mu^y(dx) = \begin{cases} p_1\delta_0(dx) + (1-p_1)\mathbf{1}_{(0,1]}(x)dx & \text{if } y = 0 \\ \\ \frac{1}{1-y}\mathbf{1}_{(0,1-y]}(x)dx & \text{if } y > 0. \end{cases}$$

In other words, if we set $s_i = 1 - (\rho_1 + \rho_2 + \cdots + \rho_i)$, then $s = (s_i)_{i\geq 1}$ is a Markov chain starting from 1 at time 0, with transition probability

$$p(y, dx) = \begin{cases} p_1\delta_1(dx) + (1-p_1)\mathbf{1}_{(0,1]}(x)dx \ ; & y = 1 \\ \\ \frac{1}{y}\mathbf{1}_{(0,y]}(x)dx \ ; & y < 1. \end{cases}$$

The process $s$ is a variant of the *stickbreaking process* [McC65, PPY92] related to the Poisson-Dirichlet (0,1) distribution, in which the first attempts to break the stick would fail (with probability $p_1$) and would produce a geometric number of fragments with size 0 at the beginning of the process. For the stickbreaking process the transition probability $\tilde{p}(y, dx)$ is $\frac{1}{y}\mathbf{1}_{(0,y]}(x)dx$ for any $y \in [0, 1]$ : the stickbreaking process can be seen as the sequence of low records of an i.i.d. sequence $U$ of uniform random variables on $[0, 1]$. Of course, whence $\rho^{(n)}$ and $\rho$ are rearranged in decreasing order, the small initial fragments are rejected at the end or, in the case of $\rho$, they disappear. Thus

**Corollary 1.3.** *The decreasing rearrangement of $\rho^{(n)}$ converges in law to the Poisson-Dirichlet (0,1) distribution.*

As regards the second result, for any Lyndon word $\mathtt{w} \in \mathfrak{L}_n$, let $R_n(\mathtt{w})$ denotes the length of its standard right factor, and set $r_n = R_n/n$. We have:

**Theorem 1.4.** *For a totally ordered alphabet with probability distribution $p$ on its letters, the sequence of normalized lengths $r_n$ of standard right factors of a random $n$-letters long Lyndon word converges in law, when $n \to \infty$, to*

$$\mu(dx) = p_1\delta_1(dx) + (1-p_1)\mathbf{1}_{[0,1)}(x)dx,$$

*where $\delta_1$ denotes the Dirac mass on the point 1 and $dx$ the Lebesgue measure on $\mathbb{R}$. As a consequence the moments of $r_n$ converge to the corresponding moments of $\mu$.*

For instance, if $p$ is the uniform distribution on $q$ letters, then the limit law of the normalized length of the standard right factor of a random Lyndon word, is

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

### 1.3 Context

The Poisson-Dirichlet family of distributions was introduced by Kingman [Kin75]. This distribution arises as a limit for the size of components of decomposable structures in a variety of settings, as shown by Hansen [Han94] or Arratia et al. [ABT99].

When the distribution $p$ is uniform on $q$ letters, i.e.

$$p_k = \frac{1}{q} \mathbb{1}_{1 \leq k \leq q},$$

the combinatorics of the Lyndon factorization have connections with that of $q$-shuffles [BD92] and of monic polynomials of degree $n$ over the finite field $GF(q)$, as explained in [GR93, DMP95]. When $p$ is uniform, Corollary 1.3 is well known (cf. [ABT93, Han94]), through the Golomb correspondance between polynomials and words. Actually, for a uniform $p$, a precise description of the size of Lyndon factors in term of the standard Brownian motion is given in [Han93, ABT93]. Our contribution is twofold :

- Theorem 1.2 provides a description of the sizes of factors in the Lyndon factorization of random words *depending on their rank* in the factorization. Obviously, the order of factors matters in the Lyndon factorization of words, while it has no meaning in the previously cited papers about polynomials or permutations ;
- even if we sort the factors' lengths in decreasing order, as in Corollary 1.3, a proof along the lines of [ABT93, Han94] seems out of reach, for we use a perfectly general distribution $p$ on the alphabet (we only require more than one letter): thus, for a combinatorial proof of our result, a precise description for the distribution of sizes of factors jointly with a count of each letter of the alphabet in each factor would be needed, most likely. For instance, for a general $p$, we were not able to prove, or to disprove, the *conditioning relation* (cf. [ABT03, p. 2]) that is usually required for convergence to the Poisson-Dirichlet distribution in such settings.

As explained in the next section, we circumvent the combinatorial complexity of the problem with the help of a shuffle trick, Lemma 2.1, a multivariate extension of the univariate result that was used in [BCN05, MZA07]. In these two papers ([BCN05, MZA07]), an invariance by shuffle is used to analyze the lengths of the 2 factors in the *standard factorization* of random *Lyndon* words with 2 equiprobable letters (an average case analysis is given in [BCN05], and the limit distribution is obtained in [MZA07]), while our paper uses a multivariate extension of the shuffle trick to obtain the asymptotic Poisson-Dirichlet behaviour in the *Lyndon decomposition* of random words.

Incidentally, because the path has already been cleared by our work on the asymptotic Poisson-Dirichlet behaviour, we provide, in Theorem 1.4, a full generalisation of the result given in [MZA07], that is, the asymptotic distribution of the lengths of the two factors of the standard factorization of random Lyndon words is given here for a general probability distribution on an eventually infinite alphabet.

## 2 Sketch of proofs

### 2.1 Lyndon factorization and runs

As seen on Figure 1, the smallest factors in the Lyndon factorization of some word $\mathtt{w} \in \mathcal{A}^n$ are usually several monoletters words $\mathtt{a}_1$ at the end of $\mathtt{w}$, but besides that, almost surely, a factor of the Lyndon factorization of a random word $\mathtt{w}$ begins with *long run* of the letter $\mathtt{a}_1$ : the long runs mark the beginnings of the smallest words in the necklace $\langle \mathtt{w} \rangle$, which are also the places where $\mathtt{w}$ is split into its Lyndon factors, according to the following rule : for instance, in a word containing 9 long runs with lexicographic ranks going from 1 to 9, without ties, the runs could be placed along the word in the following way :

$$\ldots 4 \ldots 8 \ldots 3 \ldots 5 \ldots 7 \ldots 1 \ldots 9 \ldots 2 \ldots 6 \ldots .$$

In this example, run 1, both the longest, and the smallest lexicographically, marks the beginning of the first Lyndon factor so that the position of run 1 determines the length of the first Lyndon factor, that englobes runs 9, 2 and 6. The second Lyndon factor begins with run 3 and englobes runs 5 and 7, its length is given by the positions of runs 3 and 1, then Lyndon factors 3 and 4 split at the beginning of run 4, and so on ... Note that this argument breaks down if some of these runs are tied. Runs 1, 3 and 4, are *records* of the sequence $\underline{4}8\underline{3}57\underline{1}926$ : with the help of Lemma 2.1, we plan to prove that if $V_i$ denote the (normalized) position of the beginning of run $i$ inside the random word $\mathtt{w}$, then the sequence $V = (V_i)_{i \geq 1}$ is asymptotically i.i.d. uniform on $[0, 1]$, and the Lyndon factors split $\mathtt{w}$ at positions distributed as the successive minimums (or records) of the sequence $V$. That would be the stick-breaking construction of the Dirichlet process [McC65].

However there are catches : to provide the asymptotic behaviour of the complete sequence of Lyndon factors, we need an unlimited supply of long runs, and the lexicographic ranking of these runs must be unambiguous, without any tie. For the first point, let $H_n(\mathtt{w})$ denote the number of runs longer than

$$r = \left\lceil (1 - \varepsilon) \log_{1/p_1} n \right\rceil , \tag{1}$$

in $\mathtt{w} \in \mathcal{A}^n$ (such runs are called *long runs*). According to Lemma 3.3, the sequence $(H_n)_{n \geq 1}$ is an unbounded sequence of random variables if $1 > \varepsilon > 0$. However, for the second point, the probability that there are several longest runs, tied, and also that there are ties at other positions than the first, is non vanishing. Set

$$\beta = \max \{p_1, 1 - p_1\} , \quad \tilde{r} = 1 + \lceil 3 \log_{1/\beta} n \rceil .$$

In order to break the ties, each long run of $\mathtt{w}$ has to be appended with a suffix to form a $\tilde{r}$-letters long factor of $\mathtt{w}$, called *long block*. According to Lemma 4.2, but for a vanishing probability, all these factors of $\mathtt{w}$ are different, thus they are not tied in the lexicographic order, and they are strictly smaller than the

other factors of the same length, since they begin with a long run of $\mathtt{a}_1$. Thus their rank in the lexicographic order, with their positions, give the Lyndon factorization of $\mathtt{w}$, as explained at the beginning of this section.

Let us give a more formal definition of *long runs* : in this paper $\varepsilon$ denotes a real number in $(0, 1/2)$, and a maximal run of the letter $\mathtt{b}$ in the word $\mathtt{w}$ is a factor $\mathtt{w}_{[k+1,\ell]}$ of $\mathtt{w}$ of the form $\mathtt{b}^{\ell-k}$, such that no factor $\mathtt{w}_{[s,t]}$ with $s \leq k+1 \leq \ell \leq t$ is a run of $\mathtt{b}$, unless $s = k+1 \leq \ell = t$. We usually call $k$ the *position* of the factor $\mathtt{w}_{[k+1,\ell]}$.

**Definition 3.** (Long runs and short runs) *We call* long run *(resp. short run) of* $\mathtt{w} \in \mathcal{A}^n$ *a maximal run of the letter* $\mathtt{a}_1$ *with length at least (resp. smaller than)* $r = \left\lceil (1-\varepsilon)\log_{1/p_1} n \right\rceil$. *We denote by* $H_n(\mathtt{w})$ *the number of long runs of "*$\mathtt{a}_1$*" in* $\mathtt{w}$.

## 2.2 A refined factorization

We introduce 2 refinements of the Lyndon factorization, in smaller factors, the first one according to a simple code. For that, we need the following definition:

**Definition 4.** *Set* $\mathcal{B} = \{0, 1\}$. *Now, let* $\varphi$ *denote the morphism, from* $\mathcal{A}^\star$ *to* $\mathcal{B}^\star$, *that sends the letter* $\mathtt{a}_1$ *on the digit* $\mathtt{0}$, *any other letter of* $\mathcal{A}$ *on the digit* $\mathtt{1}$, *and for any* $k$ *and any word* $\mathtt{w} \in \mathcal{A}^k$, *let* $\varphi$ *send* $\mathtt{w}$ *on the word* $\varphi(\mathtt{w}) = \varphi(\mathtt{w}_1)\varphi(\mathtt{w}_2)\ldots\varphi(\mathtt{w}_k) \in \mathcal{B}^k$. *Let* $\varphi_n$ *denote the restriction of* $\varphi$ *to* $\mathcal{A}^n$.

Let $\mathfrak{S}_m$ denotes the set of permutations of $\{1, \ldots, m\}$, and let $\tau \in \mathfrak{S}_m$. Consider, in $\mathcal{A}^\star$, the monoid $\mathcal{M}$ of words that begin, but do not end, with letter $\mathtt{a}_1$ ; $\mathcal{M}$ is stable, thus free, and its minimal set of generators

$$\mathcal{X} = \varphi^{-1}\left(\left\{\mathtt{0}^k\mathtt{1}^\ell \mid k, \ell \geq 1\right\}\right)$$

is a code, according to [BP85, Ch. 1.2]. As a consequence, the factorization in $\mathcal{X}$ of some word $\mathtt{w} \in \mathcal{M}$ is unique. It follows that the action of $\mathfrak{S}_m$ on such a factorization $\mathtt{w} = \mathtt{x}_1\mathtt{x}_2\ldots\mathtt{x}_m$ is well defined, if $\tau.\mathtt{w}$ is defined as the word of $\mathcal{M}$ whose unique factorization is $\mathtt{x}_{\tau(1)}\mathtt{x}_{\tau(2)}\ldots\mathtt{x}_{\tau(m)}$. The *orbit* $\mathfrak{O}(\mathtt{w})$ of $\mathtt{w}$ under the action of $\mathfrak{S}_m$ is the set $\{\tau.\mathtt{w} \mid \tau \in \mathfrak{S}_m\}$.

Thus, if $\tau$ is a *random permutation*, then $\tau.\mathtt{w}$ is uniformly distributed on the orbit $\mathfrak{O}(\mathtt{w})$. But the conditional distribution of $\mathtt{w}$ under $\mathbb{P}_n$, given that $\mathtt{w}$ belongs to some orbit $\mathfrak{O}$ in $\mathcal{M}$, is also the uniform distribution on $\mathfrak{O}$, since the number of occurences of any letter of $\mathcal{A}$ is the same in $\tau.\mathtt{w}$ and in $\mathtt{w}$, so that $\mathbb{P}_n(\{\tau.\mathtt{w}\}) = \mathbb{P}_n(\{\mathtt{w}\})$.

When $\mathtt{w} \notin \mathcal{M}$, we consider the longest factor $\mathrm{pr}_{\mathcal{M}}(\mathtt{w})$ of $\mathtt{w}$ that belongs to $\mathcal{M}$ ; $\mathrm{pr}_{\mathcal{M}}(\mathtt{w})$ is obtained by erasing eventually a run of non-$\mathtt{a}_1$ letters at the beginning of $\mathtt{w}$, and a run of $\mathtt{a}_1$ at the end. The lengths of these 2 runs are $\mathcal{O}(1)$ with a probability close to 1, thus $\mathrm{pr}_{\mathcal{M}}(\mathtt{w})$ and $\mathtt{w}$ are close in terms of the positions of their long runs, once they are rescaled by a factor $1/|\mathtt{w}|$. A random permutation of the factors of $\mathtt{w}$ will then be defined as a permutation of the factors of $\mathrm{pr}_{\mathcal{M}}(\mathtt{w})$, see the next section.

However, the probability that the smallest factors of $\mathtt{w}$ in $\mathcal{X}$ are tied in the lexicographic order is non vanishing when $|\mathtt{w}|$ grows, and that prevents us from reading the Lyndon factorization on each of these factors : we would need to consider sequences of factors to break the ties, and these sequences are not preserved by permutations of factors. The shuffle argument of the next section would then break.

To circumvent this problem, consider a new factorization : each factor $\mathtt{x}$ of $\mathtt{w}$ in $\mathcal{X}$ belongs to some subset $\varphi^{-1}\left(\mathtt{0}^k\mathtt{1}^\ell\right)$ for some $k, \ell \in \mathbb{N}$, so set $|\mathtt{x}|_\mathtt{0} = k$ (while $|\mathtt{x}| = k + \ell$) and consider the factorization $\mathtt{x}_1\mathtt{x}_2 \ldots \mathtt{x}_m$ of $\mathtt{w}$ in $\mathcal{X}$. In the new factorization, each factor $\mathtt{x}_i$ such that $|\mathtt{x}_i|_\mathtt{0} \geq r$ is merged with its successors $\mathtt{x}_{i+1}$, $\mathtt{x}_{i+2}$, etc ... to form a new factor $\mathtt{y} \in \mathcal{M}$, that we call *long block*, such that $|\mathtt{y}| \geq \tilde{r}$, but such that $\mathtt{y}$ is minimal in the sense that $\mathtt{y}$ has no factorization $\mathtt{uv}$ (in $\mathcal{M}$) with $|\mathtt{u}| \geq \tilde{r}$. In this way, one obtains a *new factorization* of $\mathtt{w}$, provided that there is enough space between the last long run and the end of the word, or between two long runs, so that the associated long blocks do not overlap. The subset $\tilde{\mathcal{M}} \subset \mathcal{M}$ of words with a such a factorization is again a stable, thus free, monoid, with a new minimal set of generators $\tilde{\mathcal{X}}$, again a code, that includes the long blocks plus the elements $\mathtt{x} \in \mathcal{X}$ such that $|\mathtt{x}|_\mathtt{0} < r$. In Section 4, we prove that

$$\lim_n \mathbb{P}_n \left(\mathrm{pr}_{\mathcal{M}}(\mathtt{w}) \in \tilde{\mathcal{M}}\right) = 1,$$

and more.

## 2.3 The factorization shuffle

When $\mathrm{pr}_{\mathcal{M}}(\mathtt{w}) \in \tilde{\mathcal{M}}$, there exists a unique factorization

$$\mathtt{x}_0 \mathrm{pr}_{\mathcal{M}}(\mathtt{w})\, \mathtt{x}_{m+1} = \mathtt{x}_0\mathtt{y}_1\mathtt{y}_2 \ldots \mathtt{y}_m\mathtt{x}_{m+1}$$

of $\mathtt{w}$, in which $\mathtt{y}_i \in \tilde{\mathcal{X}}$ for $1 \leq i \leq m$, $\mathtt{x}_0$ is a run of non-$\mathtt{a}_1$ letters at the beginning of $\mathtt{w}$, and $\mathtt{x}_{m+1}$ is the final run of the letter $\mathtt{a}_1$ ; if $\mathtt{w} \in \tilde{\mathcal{M}}$, $\mathtt{x}_0 = \mathtt{x}_{m+1} = \emptyset$. For $\omega \in \mathfrak{S}_m$, set

$$\omega.\mathtt{w} = \mathtt{x}_0\mathtt{y}_{\omega(1)}\mathtt{y}_{\omega(2)} \cdots \mathtt{y}_{\omega(m)}\mathtt{x}_{m+1},$$

and let

$$\mathtt{x}_1 \preceq \mathtt{x}_2 \preceq \cdots \preceq \mathtt{x}_m$$

denote the sequence of factors of $\mathrm{pr}_{\mathcal{M}}(\mathtt{w})$ sorted in increasing lexicographic order : almost surely, the first $\Theta\left(n^\varepsilon\right)$ terms at the beginning of the sequence form a *strictly* increasing subsequence, according to Lemmas 3.3 and 4.2. The orbit of $\mathtt{w}$ under $\mathfrak{S}_m$ is

$$\mathfrak{O}(\mathtt{w}) = \left\{\mathtt{x}_0\mathtt{x}_{\omega(1)}\mathtt{x}_{\omega(2)} \ldots \mathtt{x}_{\omega(m)}\mathtt{x}_{m+1} \mid \omega \in \mathfrak{S}_m\right\}.$$

This is an extension, to $\mathrm{pr}_{\mathcal{M}}^{-1}\left(\tilde{\mathcal{M}}\right)$, of the definition of $\mathfrak{O}(\mathtt{w})$ when $\mathtt{w} \in \tilde{\mathcal{M}}$. More generally, any such orbit $\mathfrak{O}$ is characterized by $m$ and by the sequence

$(\mathtt{x}_i)_{0\le i\le m+1}$, in which $(\mathtt{x}_i)_{1\le i\le m}$ is a sorted sequence of elements of $\tilde{\mathcal{X}}$, $\mathtt{x}_0$ is a run of non-$\mathtt{a}_1$ letters at the beginning of any element $\mathtt{v}\in\mathfrak{O}$, and $\mathtt{x}_{m+1}$ is the final run of the letter $\mathtt{a}_1$.

Each element $\omega.\mathfrak{O}=\mathtt{x}_0\mathtt{x}_{\omega(1)}\mathtt{x}_{\omega(2)}\dots\mathtt{x}_{\omega(m)}\mathtt{x}_{m+1}$ of $\mathfrak{O}$ yields a partition of $[0,1)$ into a (large) number $(m+2)$ of subintervals with small widths

$$x_j=\frac{|\mathtt{x}_j|}{|\mathtt{w}|},\quad 0\le j\le m+1,$$

in which the interval $[V_i(\omega),V_i(\omega)+x_i)$ filled by the factor $\mathtt{x}_i$ depends on $\omega$ through its position : this position $V_i(\omega)$ is the sum of lengths of the factors $\mathtt{x}_j$ on the left of $\mathtt{x}_i$,

$$V_i(\omega)=x_0+\sum_{j=1}^{m}x_j\mathbb{1}_{\omega^{-1}(j)<\omega^{-1}(i)},$$

and the factor $\mathtt{x}_j$ is on the left of $\mathtt{x}_i$ iff $\omega^{-1}(j)<\omega^{-1}(i)$. When convenient, set $V_i(\omega)=1$ for $i>m$. Then the successive records to the left, for the sequence $(V_i)_{i\ge1}$, give the positions of the first factors in the Lyndon decomposition, as explained at Section 2.1. Set :

$$\|\mathfrak{O}\|_2=\left\|(x_j)_{0\le j\le m+1}\right\|_2,\quad \|\mathfrak{O}\|_\infty=\max\{x_j,\ 0\le j\le m+1\}.$$

As noted previously for the factorization in $\mathcal{X}$, the conditional distribution of $\mathtt{w}$ under $\mathbb{P}_n$, given that $\mathtt{w}$ belongs to some orbit $\mathfrak{O}$ with $m$ factors in $\tilde{\mathcal{X}}$, $\mathbb{P}_n(.|\mathfrak{O})$, is uniform, but, for a nonrandom word $\mathtt{w}\in\mathfrak{O}$, if $\omega$ is a *random permutation* in $\mathfrak{S}_m$, then $\omega.\mathtt{w}$ is uniformly distributed on the orbit $\mathfrak{O}$ too. We prove in this Section that, under $\mathbb{P}_n(.|\mathfrak{O})$, $V_{[d]}=(V_i)_{1\le i\le d}$ is close to uniform on $[0,1]^d$, under mild conditions on $\mathfrak{O}$. More specifically, let $\mathbb{U}_A$ (resp. $\mathbb{U}_d$) denote the uniform probability distribution on a finite set $A$ (resp. the uniform distribution on $[0,1]^d$, for a given integer $d$). Let $U=(U_i)_{i\ge1}$ denote a sequence of i.i.d. random variables uniform on $[0,1]$, and let $U_{[d]}$ denote the sequence of its $d$ first terms, distributed according to $\mathbb{U}_d$. Recall that the $L_2$-Wasserstein metric $\mathcal{W}_2(.,.)$ is defined by

$$\mathcal{W}_2(\mu,\nu)=\inf_{\substack{\mathcal{L}(X)=\mu\\ \mathcal{L}(Y)=\nu}}\mathbb{E}\left[\|X-Y\|_2^2\right]^{1/2},\tag{2}$$

in which $\mu$ and $\nu$ are probability distributions on $\mathbb{R}^d$, and $\|.\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$. Convergence of $\mathcal{L}(X_n)$ to $\mathcal{L}(X)$ with respect to $\mathcal{W}_2(.,.)$ entails convergence of $X_n$ to $X$ in distribution (see [Rac91]). The *shuffle lemma* asserts that the Wasserstein distance between $V_{[d]}$ and $U_{[d]}$ is bounded by a simple expression depending on the maximal width of the subintervals in $\mathfrak{O}$.

**Lemma 2.1** (Shuffle lemma)**.** *For $m\ge d\ge1$, and for any orbit $\mathfrak{O}$ with sequence of factors $(\mathtt{x}_i)_{0\le i\le m+1}$,*

$$\mathcal{W}_2(V_{[d]},\mathbb{U}_d)\le\sqrt{d/3}\,\|\mathfrak{O}\|_2\le\sqrt{d\,\|\mathfrak{O}\|_\infty/3}.$$

As a consequence of Propositions 4.1 and 4.5, a.s. under $\mathbb{P}_n$, $\|\mathfrak{O}\|_\infty = \mathcal{O}(\ln(n)/n)$ almost surely.

*Proof.* The proof is similar to the proof of [MZA07, Lemma 6.3], which is the special case $k = 1$ of Lemma 2.1. As in the proof of [MZA07, Lemma 6.3], we set

$$\tilde{V}_i = x_0 + \sum_{\substack{j\colon 1\le j\le m,\\ \text{and } U_j<U_i}} x_j,$$

and we note that $\tilde{V}_{[d]} = (\tilde{V}_i)_{1\le i\le d}$ has the same distribution as $V_{[d]}$. Among the many couplings between $V_{[d]}$ and $U_{[d]}$, this special one provides the desired bound on the Wasserstein distance. Actually, for some $j \in [\![1, d]\!]$, conditioning given $U_j$ and using $\sum x_j = 1$, we obtain

$$\begin{aligned}
\mathbb{E}\left[(U_j - \tilde{V}_j)^2\right] &= \mathbb{E}\left[\left(x_0(U_j - 1) + \sum_{i=1}^m x_i\left(U_j - 1_{\{U_i<U_j\}}\right) + x_{m+1}U_j\right)^2\right] \\
&= \mathbb{E}\left[(1-U_j)^2\right]\ x_0^2 + \mathbb{E}\left[U_j^2\right]\ x_{m+1}^2 + \mathbb{E}\left[U_j(1-U_j)\right]\ \sum_{i=1}^m x_i^2 \\
&= \tfrac{1}{3}\ \left(x_0^2 + x_{m+1}^2\right) + \tfrac{1}{6}\ \sum_{j=1}^m\ x_j^2 \ \le\ \tfrac{1}{3}\ \|\mathfrak{O}\|_2^2\,.
\end{aligned}$$

The result follows from

$$\mathcal{W}_2(V_{[d]}, \mathbb{U}_d)^2 \le \sum_{j=1}^d \mathbb{E}\left[(U_j - \tilde{V}_j)^2\right].$$

We struggle with the idea that such computations are new. Actually the argument can be adapted (taking the $x_i$'s in $\{0, 1/n\}$) to compute the $L^2$ distance $(\sum t(1-t))/n$ between an evaluation $F_n(t)$ of the empirical distribution function and $t$, cf. [SW09, Ch. 3.1, p.85, display (3)]. □

The shuffle lemma delivers the expected result, provided that the factorization of $\mathtt{w}$ in $\tilde{\mathcal{X}}$ behaves as follows, under $\mathbb{P}_n$ or under $\mathbb{L}_n$ :

1. as explained before, the distribution of the random word is invariant under a random uniform shuffle of the factors (subintervals) ;
2. the length of the factors of $\mathtt{w}$ is typically $o(n)$ (actually, $\mathcal{O}(\ln n)$), while the Lyndon factors are $\Theta(n)$, so that

$$\|\mathfrak{O}\|_\infty = \mathcal{O}(\ln n/n)\,;$$

3. almost surely under $\mathbb{P}_n$ or under $\mathbb{L}_n$, $\mathrm{pr}_{\mathcal{M}}(\mathtt{w}) \in \tilde{\mathcal{M}}$.

Section 3 is devoted to preliminary results on some statistics on runs. Specially useful is the observation that the length of the longest run of the letter $\mathtt{a}_1$ is typically of order $\log_{1/p_1} n$. In Section 4, we introduce the set $\mathcal{G}_n \subset \mathcal{A}^n$ of *good words* that satisfy points 2 and 3, and we prove that the set $\mathcal{G}_n$ is almost sure.

Once these preliminary tasks are performed, we use the Shuffle Lemma 2.1, to prove the main results, Theorem 1.2 in Section 5, and Theorem 1.4 in Section 6.

## 3 Statistical properties of runs, under $\mathbb{P}_n$ or under $\mathbb{L}_n$

For $\mathtt{w} \in \mathcal{P}_n$, let $\pi(\mathtt{w})$ denote the unique Lyndon word in the necklace of $\mathtt{w}$. For $s \geq 1$, we set

$$\|p\|_s = \left(\sum_i p_i^s\right)^{1/s}.$$

The next Lemma allows to translate bounds on $\mathbb{P}_n$ into bounds on $\mathbb{L}_n$:

**Lemma 3.1.** *For $A \subset \mathcal{A}^n$, we have:*

$$\mid \mathbb{L}_n(A) - \mathbb{P}_n(\pi^{-1}(A)) \mid = \mathcal{O}\left(\|p\|_2^n\right).$$

Note that $\|p\|_1 = 1$, and that, under the assumption $\{0 < p_1 < 1\}$, $\|p\|_s$ is strictly decreasing in $s$. Among other well known inequalities, we shall make use of $\|p\|_2 \leq \sqrt{\max p_i} \leq \sqrt{\beta}$. For instance, the choice $A = \mathcal{A}^n$ leads to

$$\mid 1 - \mathbb{P}_n(\mathcal{P}_n) \mid = \mathcal{O}\left(\|p\|_2^n\right) = \mathcal{O}\left(\beta^{n/2}\right).$$

Due to Lemma 3.1, some properties of statistics, such as the number of runs and the length of the longest runs, that behave nicely under cyclic permutations, hold true a.s. for random Lyndon words as soon as they hold true a.s. for random words. That is, the next Lemmas hold true under $\mathbb{L}_n$ as well as under $\mathbb{P}_n$. Thus they prepare simultaneously the proofs of Theorems 1.2 and 1.4. The proofs of the results in this Section are tedious, and thus they are postponed to Section 7.

**Definition 5.** *For any word $\mathtt{w} \in \mathcal{A}^n$, let $N_n(\mathtt{w})$ denote the number of runs in $\varphi(\mathtt{w})$, let $W_1(\mathtt{w}), W_2(\mathtt{w}), \ldots, W_{N_n}(\mathtt{w})$ be their lengths, let $N_n^{(\mathtt{e})}(\mathtt{w})$ (resp. $M_n^{(\mathtt{e})}(\mathtt{w})$) denote the number of runs of the letter $\mathtt{e}$ in the word $\mathtt{w}$ (resp. the maximal length of such runs). Thus, for $\mathtt{w} \in \mathcal{A}^\star$, $N_n^{(\mathtt{a}_1)}(\mathtt{w}) = N_n^{(0)}(\varphi(\mathtt{w}))$.*

First, since Theorem 1.2 deals with a sequence of factors that begin with a long run of $\mathtt{a}_1$, a large number of such runs is required:

**Lemma 3.2** (Number of runs of the letter $\mathtt{a}_1$)**.** *Set $\sigma^2 = p_1(1-p_1)$. For $(a,b) \in \mathbb{R}^2$ such that $a < \sigma^2$, we have*

$$\mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} < an + b\right) = \mathcal{O}\left(n^{-1}\right),$$

*and*

$$\mathbb{L}_n\left(N_n^{(\mathtt{a}_1)} < an + b\right) = \mathcal{O}\left(n^{-1}\right).$$

We also need some information about the length of the longest runs of "$\mathtt{a}_1$" in a word $\mathtt{w} \in \mathcal{A}^n$ and in its necklace $\langle \mathtt{w} \rangle$, for, among these long runs, the longest is expected to be the prefix of the smallest Lyndon factor of $\mathtt{w}$, or the prefix of the unique Lyndon word in $\langle \mathtt{w} \rangle$. Also, the second longest is expected to be the prefix of the second smallest Lyndon factor of $\mathtt{w}$ or the prefix of the standard right factor of the Lyndon word in $\langle \mathtt{w} \rangle$. Furthermore, if Theorem 1.4 is to be true, there should exist at least two long runs and, if Theorem 1.2 is to be true, the number of these long runs should grow indefinitely with $n$, like the number of Lyndon factors of the random word. These points are consequences of Theorems 1.2 and 1.4, but they are also some of the steps of the proofs of these Theorems. They are addressed by the next Lemmas. According to Definition 3, $H_n(\mathtt{w})$ is the number of long runs in the word $\mathtt{w} \in \mathcal{A}^n$.

**Lemma 3.3** (Number of long runs)**.**

$$\mathbb{P}_n\left(H_n \geq \alpha n^{\varepsilon}\right) = 1 - \mathcal{O}\left(n^{-1}\right),$$

*and*

$$\mathbb{L}_n\left(H_n \geq \alpha n^{\varepsilon}\right) = 1 - \mathcal{O}\left(n^{-1}\right),$$

*in which $\alpha$ is a positive constant smaller than $\sigma^2$.*

Recall that $M_n^{(1)}$ denote the length of the largest run of non-$\mathtt{a}_1$ letters. We have:

**Lemma 3.4** (Large values of the longest runs)**.** *Under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probabilities of the events $\left\{M_n^{(0)} \geq 2\log_{1/p_1} n\right\}$ and $\left\{M_n^{(1)} \geq 2\log_{1/(1-p_1)} n\right\}$ are $\mathcal{O}\left(n^{-1}\right)$.*

The asymptotic behaviour of the Lyndon (resp. standard) factorization depends on $p$ only through $p_1$, and the reason appears in the proofs of the previous Lemmas, to be found in Section 7 : only the lengths and positions of the runs of $\mathtt{a}_1$ matter.

## 4 Long blocks of words and good words

We mentioned in Section 2.2 the need to break ties between the long runs: in order to do just that, we introduced the stable monoid $\tilde{\mathcal{M}}$ and its minimal set of generators, the code $\tilde{\mathcal{X}} = (\tilde{\mathcal{M}}/\{\emptyset\})^2 \backslash (\tilde{\mathcal{M}}/\{\emptyset\})$.

**Definition 6.** *The elements of $\tilde{\mathcal{X}}$ are of two sorts, that we call* long *and* short blocks *:*

- *the* short blocks *are the elements* $\mathtt{x}$ *of* $\mathcal{X}$ *such that* $|\mathtt{x}|_0 < r$ *;*

– *the* long blocks *are the elements* $\mathtt{y}$ *of* $\mathcal{M}$ *that satisfy*
  – $\mathtt{y}$ *begin with a long run,*
  – $|\mathtt{y}| \geq \tilde{r}$,
  – $\mathtt{y}$ *is minimal in that* $\mathtt{y} \notin \{\mathtt{uv} \mid \mathtt{u}, \mathtt{v} \in \mathcal{M} \setminus \{\emptyset\}$ *and* $|\mathtt{u}| \geq \tilde{r}\}$.

When $\mathtt{w}$ belongs to the set $\mathcal{G}_n \subset \mathcal{A}^n$ of *good words*, defined below,

$$\|\mathfrak{D}(\mathtt{w})\|_\infty = \mathcal{O}\left(\ln n/n\right)$$

and Lemma 2.1 provides the desired asymptotically uniform distribution for $V_{[d]}$, conditionally given $\mathfrak{D}(\mathtt{w})$, for any $d \geq 1$.

**Definition 7.** *A word* $\mathtt{w} \in \mathcal{A}^n$ *is a* good word *if it satisfies the following conditions:*

***i.*** $\mathtt{w}$ *has at least* $\lfloor \alpha n^\varepsilon \rfloor$ *long blocks ;*
***ii.*** $pr_{\mathcal{M}}(\mathtt{w}) \in \tilde{\mathcal{M}}$ *;*
***iii.*** *if two long blocks have a common factor, its length is not larger than* $\tilde{r} - 1$,
***iv.*** $M_n^{(0)}(\varphi(\mathtt{w})) \leq 2\log_{1/p_1} n$, *and* $M_n^{(1)}(\varphi(\mathtt{w})) \leq 2\log_{1/(1-p_1)} n$.

In this section, we prove that $\mathcal{G}_n$ has a large probability:

**Proposition 4.1.** *Under* $\mathbb{P}_n$ *or* $\mathbb{L}_n$, *the probability of* $\mathcal{G}_n^c$ *is* $\mathcal{O}\left(n^{2\varepsilon-1} \ln^2 n\right)$.

Recall that $\tilde{r} = 1 + \lceil 3\log_{1/\beta} n \rceil$. For the proof of Proposition 4.1, we need a few lemmas:

**Lemma 4.2.** *Denote by* $\mathcal{E}_n$ *the set of words* $\mathtt{w} \in \mathcal{A}^n$ *in which some* $\tilde{r}$*-letters long factor appears twice in the necklace* $\langle \mathtt{w} \rangle$, *at two non-overlapping positions:*

$$\mathcal{E}_n = \left\{\mathtt{w} \in \mathcal{A}^n \mid \exists (\mathtt{w}', v, a, b) \in \langle \mathtt{w} \rangle \times \mathcal{A}^{\tilde{r}} \times (\mathcal{A}^\star)^2 \ s.t.\ \mathtt{w}' = vavb\right\}.$$

*Then, under* $\mathbb{P}_n$ *or* $\mathbb{L}_n$, *the probability of* $\mathcal{E}_n$ *is* $\mathcal{O}\left(n^{-1}\right)$.

A key argument of the proof of the main results breaks down if some long block of the factorization of a random word is a prefix of another long block, somewhere else in the word. In order to preclude that, we shall consider blocks with at least $\tilde{r}$ letters (at least thrice the length of the longest run(s)[2] of the letter $\mathtt{a}_1$), and we shall use Lemma 4.2.

*Proof.* We have

$$\mathbb{P}_n(\mathcal{E}_n) = \mathcal{O}\left(n^2\ \beta^{\tilde{r}}\right) = \mathcal{O}\left(n^{-1}\right), \tag{3}$$

in which $n^2$ is a bound for the number of positions of the pair of factors of $\mathtt{w}$, and $\beta^{\tilde{r}}$ is a bound for the conditional probability that the second factor is equal to the first factor, given the value of the first factor and the positions of the factors. Due to Lemma 3.1, $\mathbb{L}_n(\mathcal{E}_n)$ satisfies

$$\mid \mathbb{L}_n(\mathcal{E}_n) - \mathbb{P}_n\left(\pi^{-1}(\mathcal{E}_n)\right) \mid = \mathcal{O}\left(\beta^{n/2}\right),$$

and $\pi^{-1}(\mathcal{E}_n) = \mathcal{E}_n \cap \mathcal{P}_n \subset \mathcal{E}_n$. □

[2] The probability that there exists several runs with the same maximal length inside a $n$-letters long random word is non vanishing with $n$ large, so $\log_{1/p_1} n$ characters would be too short.

As mentioned at the very beginning of the paper, a word $\mathtt{v}$ is a factor of $\langle \mathtt{w} \rangle$ as soon as it is a factor of some element $\mathtt{w}' \in \langle \mathtt{w} \rangle \subset \mathcal{A}^n$. As a consequence, an $\ell$-letters long factor $\mathtt{v}$ of $\mathtt{w}$ can be found at $n - \ell + 1$ positions, while such a factor $\mathtt{v}$ of $\langle \mathtt{w} \rangle$ has $n$ possible positions.

**Lemma 4.3** (Overlap of long blocks)**.** *Let $\mathcal{F}_n$ denote the set of words $\mathtt{w} \in \mathcal{A}^n$ such that some factor of $\langle \mathtt{w} \rangle$, $\lceil 7 \log_{1/\beta} n \rceil$-letters long, contains two disjoint long runs. Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of $\mathcal{F}_n$ is $\mathcal{O}\left(n^{2\varepsilon-1} \ln^2 n\right)$.*

*Proof.* The bound for $\mathbb{P}_n(\mathcal{F}_n)$ has a factor $n$ for the position of the $\lceil 7 \log_{1/\beta} n \rceil$-letters long factor, a factor $49(\log_{1/\beta} n)^2$ (a crude bound) for the positions of the 2 runs inside this factor, and a factor $n^{2\varepsilon-2} \geq p_1^r \times p_1^r$ for the probability of 2 disjoint $r$-letters long runs at 2 specified positions. The proof extends to $\mathbb{L}_n$ according to Lemma 3.1. □

**Lemma 4.4.** *Let $\mathcal{I}_n$ denote the set of words $\mathtt{w} \in \mathcal{A}^n$ whose suffix of length $2\tilde{r}$ contains a long run of "$\mathtt{a}_1$". Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of $\mathcal{I}_n$ is $\mathcal{O}\left(n^{2\varepsilon-1} \ln^2 n\right)$.*

*Proof.* We have

$$\mathbb{P}_n(\mathcal{I}_n) \leq 2n^{\varepsilon-1}\ \tilde{r}.$$

The factor $2\tilde{r}$ bounds the number of positions where such a long run could begin. The factor $n^{\varepsilon-1} = {p_1}^{(1-\varepsilon)\log_{1/p_1} n}$ is the probability that a long run begins at some given position. The result for $\mathbb{L}_n$ follows from Lemma 3.1, Lemma 4.3 and

$$\mathbb{P}_n\left(\pi^{-1}(\mathcal{I}_n)\right) \leq \mathbb{P}_n(\mathcal{F}_n).$$

□

*Proof of Proposition 4.1.* Consider the sets

$$\mathcal{V}_n = \{\mathtt{w} \in \mathcal{A}^n \mid \mathtt{w} \text{ satisfies iv. and } H_n(\mathtt{w}) \geq \alpha n^{\varepsilon}\}$$

and

$$\tilde{\mathcal{G}}_n = \mathcal{V}_n \setminus (\mathcal{E}_n \cup \mathcal{F}_n \cup \mathcal{I}_n).$$

Then, under $\mathbb{P}_n$ or $\mathbb{L}_n$, the probability of $\tilde{\mathcal{G}}_n^c$ is $\mathcal{O}\left(n^{2\varepsilon-1} \log^2 n\right)$, due to Lemmas 3.3, 3.4, 4.2 and 4.4, since, for instance,

$$\mathbb{P}_n\left(\tilde{\mathcal{G}}_n^c\right) \leq \mathbb{P}_n\left(\mathcal{V}_n^c\right) + \mathbb{P}_n\left(\mathcal{E}_n\right) + \mathbb{P}_n\left(\mathcal{F}_n\right) + \mathbb{P}_n\left(\mathcal{I}_n\right).$$

Let us prove that $\tilde{\mathcal{G}}_n \subset \mathcal{G}_n$ : consider a word $\mathtt{w} \in \tilde{\mathcal{G}}_n$, and in order to prove that $\mathtt{w}$ satisfies conditions **i.** and **ii.** in Definition 7, consider a long run $\mathtt{u} = \mathtt{a}_1^k$ of $\mathtt{w} = \mathtt{tus}$, with $k \in [\![r, 2\log_{1/p_1} n]\!]$ : is it the prefix of a long block $\mathtt{uv}$ ?

The eventual long block beginning with $\mathtt{u}$ ends with the run of non-$\mathtt{a}_1$ letters containing the character $\mathtt{w}_{|\mathtt{t}|+\tilde{r}}$ of the word $\mathtt{w}$, if $\mathtt{w}_{|\mathtt{t}|+\tilde{r}} \neq \mathtt{a}_1$, or with the

next run of non-$\mathtt{a}_1$ letters, if $\mathtt{w}_{|\mathtt{t}|+\tilde{r}} = \mathtt{a}_1$. In this last case, $\mathtt{w}_{|\mathtt{t}|+\tilde{r}}$ is part of a short run, else $\mathtt{w}$ would belong to $\mathcal{F}_n$. Thus, in any case,

$$|\mathtt{uv}| \leq \tilde{r} + r + 2\log_{1/(1-p_1)} n < 2\tilde{r}.$$

Since $\mathtt{w} \notin \mathcal{I}_n$, we have

$$|\mathtt{t}| + |\mathtt{uv}| \leq n,$$

thus there is room enough to build a long block beginning with $\mathtt{v}$, by merging $\mathcal{X}$-factors of $\mathtt{w}$. Also $\mathtt{w} \notin \mathcal{F}_n$, thus $\mathtt{v}$ does not contain a second long run, and long blocks overlap does not happen. Finally, $\mathtt{w}$ satisfies **ii.**, so there exists a long block for each long run, and since $\mathtt{w} \in \mathcal{V}_n$, $\mathtt{w}$ satisfies **i.**. Condition **iv.** is satisfied by definition of $\mathcal{V}_n$ and **iii.** is satisfied because $\mathtt{w} \notin \mathcal{E}_n$. □

**Proposition 4.5.** *If* $\mathtt{w} \in \mathcal{G}_n$*, then* $\|\mathfrak{O}(\mathtt{w})\|_\infty \leq 2\tilde{r}/n$.

*Proof.* Note that by Definition 6, the short blocks of some word $\mathtt{w} \in \mathcal{G}_n$ begin with a short run of less than $r$ letters $\mathtt{a}_1$, and due to point **iv.** of Definition 7, a run of less than $2\log_{1/(1-p_1)} n$ "letters" $\bar{\mathtt{a}}_1$, thus short blocks have less than

$$r + 2\log_{1/(1-p_1)} n \leq \tilde{r}$$

letters, while its long blocks are not longer than

$$\tilde{r} - 1 + r - 1 + \lfloor 2\log_{1/(1-p_1)} n \rfloor \leq 2\tilde{r}.$$

For a long block, count $\tilde{r}$ letters for the minimal size of a long block, plus eventually a run of "$\mathtt{a}_1$" (a short one, due to point **ii.** of Definition 7, at most $r-1$-letters long starting before the $\tilde{r}$-limit) and a run of "$\bar{a}_1$", at most $\lfloor 2\log_{1/(1-p_1)} n \rfloor$ letters, due to point **iv.** of Definition 7. □

In the two following sections, we prove separately the main theorems, Theorem 1.2 and Theorem 1.4.

## 5 Proof of Theorem 1.2

Set

$$s_{n,i}(\mathtt{w}) = 1 - (\rho_{n,1}(\mathtt{w}) + \rho_{n,2}(\mathtt{w}) + \cdots + \rho_{n,i}(\mathtt{w}));$$

$s_{n,i}(\mathtt{w})$ is the normalized position of the $i$th factor of the Lyndon decomposition of $\mathtt{w}$, meaning that the $i$th factor of $\mathtt{w}$ is $\mathtt{w}_{[n\,s_{n,i}, n\,s_{n,i-1}-1]}$. The correspondance between $s^{(n)} = (s_{n,i})_{i\geq 1}$ and $\rho^{(n)}$ is bicontinuous on $[0,1]^{\mathbb{N}}$, thus Theorem 1.2 is equivalent to the convergence in distribution of $s^{(n)}$ to $s$, $s$ being the variant of the stickbreaking process defined at Section 1.2.

For any word $\mathtt{w} \in \mathcal{G}_n$, according to condition **ii.** of Definition 7, $\mathrm{pr}_{\mathcal{M}}(\mathtt{w})$ belongs to $\tilde{\mathcal{M}}$ and, as such, it has a unique factorization in $\tilde{\mathcal{X}}$, thus we can write :

$$\mathtt{w} = \bar{\mathtt{a}}_1^{k(\mathtt{w})} Y_1(\mathtt{w}) \ldots Y_{M_n(\mathtt{w})-1}(\mathtt{w}) Y_{M_n(\mathtt{w})}(\mathtt{w}) \mathtt{a}_1^{L_n(\mathtt{w})}, \tag{4}$$

in which $k \geq 0$, $L_n(\mathtt{w}) \geq 0$ and the $Y_i$'s are elements of $\tilde{\mathcal{X}}$, either long blocks, or short blocks. Let $J_{i,n}(\mathtt{w})$, $1 \leq i \leq H_n(\mathtt{w})$, denote the index of the $i$-th smallest block of $\mathtt{w} \in \mathcal{G}_n$: since $i \leq H_n(\mathtt{w})$, $Y_{J_{i,n}}$ has to be a long block, and there are no ties among long blocks, so that $J_{i,n}$ is well defined on $\mathcal{G}_n$. Let $V_{i,n}(\mathtt{w})$, $1 \leq i \leq H_n(\mathtt{w})$, denote the normalized position of $Y_{J_{i,n}}(\mathtt{w})$ in $\mathtt{w}$, defined as the ratio $|\mathtt{u}|/|\mathtt{w}|$, in which $\mathtt{w}$ has the factorization $\mathtt{w} = \mathtt{u}Y_{J_{i,n}}(\mathtt{w})\mathtt{v}$. The normalized position $V_{i,n}(\mathtt{w})$ is given by the formula :

$$V_{i,n}(\mathtt{w}) = \frac{1}{|\mathtt{w}|}\left(k(\mathtt{w}) + \sum_{j=1}^{J_{i,n}(\mathtt{w})-1} |Y_j(\mathtt{w})|\right) \quad ; \qquad i = 1, \ldots, H_n(\mathtt{w}). \tag{5}$$

For a word $\mathtt{w} \in \mathcal{G}_n$, it is convenient to complete the sequence $(V_{i,n}(\mathtt{w}))_{1 \leq i \leq H_n(\mathtt{w})}$ by an infinite sequence of 0's, in order to form an infinite sequence

$$V^{(n)}(\mathtt{w}) = (V_{i,n}(\mathtt{w}))_{i \geq 1},$$

and for a word $\mathtt{w} \in \mathcal{A}^n \backslash \mathcal{G}_n$, let $V^{(n)}(\mathtt{w})$ be an infinite sequence of 0's. For a word $\mathtt{w} \in \mathcal{G}_n$, this is not much of a perturbation, since the original sequence is very long : according to Lemma 3.3, the probability that $H_n(\mathtt{w})$ is smaller than $\alpha n^{\varepsilon}$ vanishes.

Now let us address the $L_n$ first terms of $s^{(n)}$ : they form the sequence

$$\tfrac{1}{n}\ (n-1, n-2, \ldots, n-L_n-1, n-L_n),$$

If $L_n(\mathtt{w}) = \ell \geq 1$, the first $\ell$ factors of the Lyndon factorization are $\ell$ words reduced to one letter "$a_1$". Thus, for $1 \leq i \leq \ell$,

$$s_{i,n} = 1 - \frac{i}{n}.$$

Let $L$ denote a geometric random variable with parameter $1 - p_1$, such that, for $\ell \geq 0$,

$$\mathbb{P}(L = \ell) = p_1^{\ell}(1 - p_1),$$

and let $\xi$ denote its probability distribution. It turns out that $L_n$ converges in distribution to $L$. Let $\Lambda_n$ denote the number of low records of the sequence $(V_{i,n}(\mathtt{w}))_{1 \leq i \leq H_n(\mathtt{w})}$. Then, as explained at Section 2.1, the next $\Lambda_n$ terms of the sequence $s^{(n)}$, i.e. $s_{k+1,n}, s_{k+2,n}, \ldots, s_{k+\Lambda_n,n}$, are the low records of the sequence $(V_{i,n}(\mathtt{w}))_{1 \leq i \leq H_n(\mathtt{w})}$. In light of this, the proof has 4 steps :

1. $s$ can be described as the sequence of low records of a sequence $U = (U_i)_{i \geq 1}$ of i.i.d. random variables uniform on $[0, 1]$, appended with a prefix sequence of $L$ 1's, $L$ and $U$ independent, an operation that we denote $\mathfrak{A}(L, U)$, and that we define more formally below ;
2. $(L_n, V^{(n)})$ converges in distribution to $(L, U)$ ;
3. provided that $\mathfrak{A}$ has some regularity properties, $\mathfrak{A}(L_n, V^{(n)})$ converges in distribution to $\mathfrak{A}(L, U) = s$ ;
4. since $\lim_n \Lambda_n = +\infty$ in some sense, cf. Proposition 5.1 below, $s^{(n)}$ and $\mathfrak{A}(L_n, V^{(n)})$ are close in some sense.

The following bound is proven at Section 7.6 :

**Proposition 5.1.** *For $\varepsilon$ previously chosen in* $(0,1)$,

$$\mathbb{P}_n(\Lambda_n \leq \varepsilon \log n/3) = \mathcal{O}\left(\frac{1}{\log n}\right).$$

For points 1 and 3, let $T$ be the functional that shifts a sequence $u$ as follows:

$$T(u) = T(u_1, u_2, \dots) = (1, u_1, u_2, \dots).$$

Let $S$ be the functional that keeps track of the sequence of low records (in the broad sense) of a sequence $u$ of real numbers. The functional $S$ is well defined and is continuous on a set of measure 1 of $[0,1]^{\mathbb{N}}$, for instance on the set $\mathcal{R}$ of sequences $u$ without repetition such that $\liminf u = 0$. Then the functional $\mathfrak{A}$ defined on $\mathbb{N} \times \mathcal{R}$ by

$$\mathfrak{A}(k, u) = T^k \circ S(u)$$

is continuous as well. Now $s$ has the same distribution as $\mathfrak{A}(L, U)$ as a consequence of the Markov property of $s$ and of the particular form of its transition kernel.

For point 4, note that the difference between the two sequences $s^{(n)}$ and $\mathfrak{A}(L_n, V^{(n)})$ is

$$\mathfrak{A}(L_n, V^{(n)}) - s^{(n)} = \left(\tfrac{1}{n}, \tfrac{2}{n}, \dots, \tfrac{k}{n}, 0, 0, \dots, 0, s_{k+\Lambda_n+1,n}, s_{k+\Lambda_n+2,n}, \dots\right).$$

Endowing $[0,1]^{\mathbb{N}}$ with the distance

$$d(u, v) = \sum_{k \geq 1} 2^{-k} |u_k - v_k|,$$

we obtain

$$d(s^{(n)}, \mathfrak{A}(L_n, V^{(n)})) \leq \frac{8}{n} + 2^{-L_n - \Lambda_n} \leq \frac{8}{n} + 2^{-1-\Lambda_n}.$$

This inequality and Proposition 5.1 entail that

$$\mathbb{E}\left[d(s^{(n)}, \mathfrak{A}(L_n, V^{(n)}))\right] \leq \frac{8}{n} + \mathcal{O}\left(\frac{1}{\log n}\right) + n^{-\varepsilon \ln(2)/3}.$$

According to [Bil99][Th. 3.1], the previous bound entails that if $\mathfrak{A}(L_n, V^{(n)})$ converges in distribution to $\mathfrak{A}(L, U)$, then $s^{(n)}$ converges in distribution to $\mathfrak{A}(L, U)$ too. Finally, for point 2, note that, under $\mathbb{P}_n$, $L_n$ has the same distribution $\xi_n$ as $L \wedge n$. This is perhaps clearer when one considers the word $\overline{\mathtt{w}}$ obtained by reading the word $\mathtt{w}$ from right to left: clearly, under $\mathbb{P}_n$, $\overline{L}_n$ defined by

$$\overline{L}_n(\mathtt{w}) = L_n(\overline{\mathtt{w}})$$

has the same distribution as $L_n$, for $\mathtt{w}$ and $\overline{\mathtt{w}}$ have the same weight. But $\overline{L}_n$ has the same distribution as $L \wedge n$. Thus $L^2$ convergence of $L \wedge n$ to $L$ entails that

$$\mathcal{W}_2\left(\xi_n, \xi\right) \left(= \mathcal{W}_2\left(\xi_n \otimes \mathbb{U}_d, \xi \otimes \mathbb{U}_d\right)\right) = \mathcal{O}\left(n^2 e^{-n}\right). \tag{6}$$

Let $\mathbb{G}_n$ denote the conditional probability given $\mathcal{G}_n$:

$$\mathbb{G}_n\left(A\right) = \frac{\mathbb{P}_n\left(A \cap \mathcal{G}_n\right)}{\mathbb{P}_n\left(\mathcal{G}_n\right)}.$$

**Theorem 5.2.** *Under $\mathbb{P}_n$ or under $\mathbb{G}_n$, $V^{(n)}$ converge in distribution to $U$.*

*Proof.* Set

$$V_{[d]}^{(n)} = \left(V_{i,n}\right)_{1 \le i \le d},$$

and let $\nu_{d,n}$ be the the distribution of $V_{[d]}^{(n)}$ under $(\mathcal{G}_n, \mathbb{G}_n)$. Due to [Kal97, Theorem 3.29], we need to prove that, for any $d \ge 1$, the sequence $V_{[d]}^{(n)}$ is, under $\mathbb{G}_n$, asymptotically uniform on $[0,1]^d$, which results from considerations in Sections 2.2 and 2.3 :

**Lemma 5.3** (Positions of the $d$ first smallest blocks)**.** *We have*

$$\mathcal{W}_2\left(\nu_{d,n}, \mathbb{U}_d\right) \le \sqrt{2d\tilde{r}/3n}.$$

*Proof.* For $\mathtt{w} \in \mathcal{G}_n$, let $\tau \in \mathfrak{S}_{M_n(\mathtt{w})}$ act on $\mathtt{w}$ by permutation of blocks :

$$\tau.\mathtt{w} = \overline{\mathtt{a}}_1^k Y_{\tau(1)}(\mathtt{w}) \dots Y_{\tau(M_n(\mathtt{w}))}(\mathtt{w}) \mathtt{a}_1^{L_n(\mathtt{w})}.$$

As in Section 2, let $\mathfrak{O}(\mathtt{w})$ denote its orbit under that action. Let $\mathfrak{C}_n$ denote the $\sigma$-algebra generated by $\mathcal{C}_n = \{\mathfrak{O}(\mathtt{w})\,;\ \mathtt{w} \in \mathcal{G}_n\}$. For $\mathfrak{O} \in \mathcal{C}_n$, let $\nu_{\mathfrak{O},n}$ denote the conditional distribution of $V_{[d]}^{(n)}(\mathtt{w})$ given that $\mathtt{w} \in \mathfrak{O}$. Let $X(\mathtt{w}) = \left(X_i(\mathtt{w})\right)_{i \ge 0}$ denote the sequence of long blocks of $\mathrm{pr}_{\mathcal{M}}\left(\mathtt{w}\right)$ sorted in increasing lexicographic order, ended by an infinite sequence of empty words, i.e. $X_i(\mathtt{w}) = Y_{J_{i,n}}(\mathtt{w})$, if $1 \le i \le H_n(\mathtt{w})$, else $X_i(\mathtt{w}) = \emptyset$. Let $\Xi(\mathtt{w}) = \left(\Xi_i(\mathtt{w})\right)_{i \ge 0}$ be the corresponding sequence of lengths. We have :

**Lemma 5.4.** *The weight $p(.)$, $X$, $\Xi$, $H_n$, $L_n$ and $M_n$ are $\mathfrak{C}_n$-measurable, and*

$$\mathbb{G}_n = \sum_{\mathfrak{O} \in \mathcal{C}_n} \frac{\mathbb{P}_n(\mathfrak{O})}{\mathbb{P}_n(\mathcal{G}_n)}\, \mathbb{U}_{\mathfrak{O}}.$$

*Also, $\nu_{\mathfrak{O}(\mathtt{w}),n}$ is the image of the uniform probability on $\mathfrak{S}_{K_n(\mathtt{w})}$ by the application $\tau \longmapsto V_d^{(n)}(\tau.\mathtt{w})$.*

*Proof.* The weight $p(\mathtt{w})$ depends only on the number of letters $a_1$, $a_2$, ... that compose the word $\mathtt{w}$, not on the order of these letters in $\mathtt{w}$, thus $p(.)$ is constant on each $\mathfrak{O} \in \mathcal{C}_n$: as a consequence, under $\mathbb{G}_n$, the conditional distribution of $\mathtt{w}$ given that $\mathtt{w} \in \mathfrak{O}$ is $\mathbb{U}_{\mathfrak{O}}$. The statement about $X$, $\Xi$, $H_n$, $L_n$ and $M_n$ holds true because $\tilde{\mathcal{X}}$ is a code. Since $\mathcal{C}_n$ is a partition of $\mathcal{G}_n$, the relation in Lemma 5.4 is the desintegration of $\mathbb{G}_n$ according to its conditional distributions given $\mathcal{C}_n$. The last part holds true because the distribution of $\tau \longmapsto \tau.\mathtt{w}$ is $\mathbb{U}_{\mathfrak{O}(\mathtt{w})}$. □

According to Lemma 2.1 and (5),

$$\mathcal{W}_2\left(\nu_{\mathfrak{O},n}, \mathbb{U}_d\right) \leq \sqrt{d \left\|\mathfrak{O}\right\|_\infty /3}. \tag{7}$$

Then, Proposition 4.5, with the desintegration in Lemma 5.4, give the desired result. □

This is for the proof under $\mathbb{G}_n$. Note that the conditional law $\tilde{\nu}$, given $A$, of a $[0,1]^d$-valued random variable $X$, defined on a probabilistic space $\Omega$, is Wasserstein-close to its unconditional law $\nu$, if $A$ is close to $\Omega$ :

$$\mathcal{W}_2\left(\nu, \tilde{\nu}\right) \leq \sqrt{d\ \mathbb{P}\left(\Omega \backslash A\right)}. \tag{8}$$

As a consequence, Lemma 5.3, together with Proposition 4.1, entails that, under $\mathbb{P}_n$,

$$\mathcal{W}_2\left(V_{[d]}^{(n)}, \mathbb{U}_d\right) = \mathcal{O}\left(n^{-1/2+\varepsilon}\ \log n\right),$$

for any $d \geq 1$, which ensures the convergence of $V^{(n)}$ to $U$ under $\mathbb{P}_n$ too. □

The proofs of the weak convergence of $L_n$ and $V^{(n)}$, respectively, are complete, now the weak convergence of $(L_n, V^{(n)})$ is a consequence of the asymptotic independence between $L_n$ and $V^{(n)}$ : all the conditional distributions $\nu_{\mathfrak{O},n}$ have the same limit $\mathbb{U}_d$, thus $V_{[d]}^{(n)}$ is asymptotically independent of $\mathfrak{C}_n$, while, according to Lemma 5.4, $L_n$ is $\mathfrak{C}_n$ measurable, i.e. constant on each $\mathfrak{O}$ (equal to $L_n(\mathfrak{O})$). More precisely, step by step,

- the distribution $\chi_n$ of $(L_n, V_{[d]}^{(n)})$, under $\mathbb{G}_n$, has the desintegration

$$\chi_n = \sum_{\mathfrak{O} \in \mathcal{C}_n} \frac{\mathbb{P}_n(\mathfrak{O})}{\mathbb{P}_n(\mathcal{G}_n)}\ \delta_{L_n(\mathfrak{O})} \otimes \nu_{\mathfrak{O},n}\ ;$$

- relation (7) has the straightforward extension

$$\mathcal{W}_2\left(\delta_{L_n(\mathfrak{O})} \otimes \nu_{\mathfrak{O},n}, \delta_{L_n(\mathfrak{O})} \otimes \mathbb{U}_d\right) \leq \sqrt{d \left\|\mathfrak{O}\right\|_\infty /3}\ ;$$

- due the desintegration of $\chi_n$, the previous bound entails that, under $\mathbb{G}_n$,

$$\mathcal{W}_2\left(\chi_n, \xi_n \otimes \mathbb{U}_d\right) \leq \sqrt{2d\tilde{r}/3n}.$$

- under $\mathbb{P}_n$, $\mathcal{W}_2\left(\chi_n, \xi_n \otimes \mathbb{U}_d\right)$ still vanishes due to (8), and (6) completes the proof of point 2.

Note that the largest (and shortest) Lyndon factors, that begin with short blocks, or that do not begin with letter $\mathtt{a}_1$, do not appear in this list of $H_n$ factors, but, as a consequence of Theorem 1.2, the total length of these largest factors is $o(n)$ : once normalized by $n$, their lengths do not contribute to the asymptotic behavior of the factorization.

## 6 Proof of Theorem 1.4

In this section, we extend [MZA07, Theorem 6.4] to a general distribution on an infinite alphabet. The proof is similar to that of Theorem 1.2 or of [MZA07, Theorem 6.4]. For $n \geq 2$, $\mathtt{w} \in \mathfrak{L}_n$ entails that $\mathrm{pr}_{\mathcal{M}}(\mathtt{w}) = \mathtt{w}$, thus $\mathfrak{L}_n \subset \mathcal{M}$, and, according to Definition 7, $\mathcal{G}_n \cap \mathfrak{L}_n \subset \tilde{\mathcal{M}}$. As a consequence, any $\mathtt{w} \in \mathcal{G}_n \cap \mathfrak{L}_n$ has a unique factorization according to the code $\tilde{\mathcal{X}}$:

$$\mathtt{w} = Y_0(\mathtt{w})Y_1(\mathtt{w})\dots Y_{K_n(\mathtt{w})-1}(\mathtt{w})Y_{K_n(\mathtt{w})}(\mathtt{w}),$$

in which the $Y_i$'s stand either for a long block or for a short block. Moreover, $Y_0(\mathtt{w})$ is the smallest block.

Let $\widehat{\mathbb{G}}_n$ denote the conditional distribution given that $\mathtt{w} \in \mathcal{G}_n \cap \mathfrak{L}_n$. As in the previous section, let $V_{2,n}(\mathtt{w})$ denote the normalized position of the second smallest block in the factorization of $\mathtt{w}$ according to the code $\tilde{\mathcal{X}}$, and let $\nu_n$ denotes the distribution of $V_{2,n}$ under $(\mathcal{G}_n \cap \mathfrak{L}_n, \widehat{\mathbb{G}}_n)$. We first prove that:

**Theorem 6.1.** *For the distribution of the position of the second smallest block, it holds that:*

$$\mathcal{W}_2(\nu_n, \mathbb{U}_1) = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right).$$

*As a consequence, under $\widehat{\mathbb{G}}_n$, the moments of $V_{2,n}$ converge to the corresponding moments of $\mathbb{U}_1$.*

*Proof.* The proof of Theorem 1.2 holds step by step: for $\mathtt{w} \in \mathcal{G}_n \cap \mathfrak{L}_n$ and $\tau \in \mathfrak{S}_{K_n(\mathtt{w})}$, set

$$\tau.\mathtt{w} = Y_0(\mathtt{w})Y_{\tau(1)}(\mathtt{w})Y_{\tau(2)}(\mathtt{w})\dots Y_{\tau(K_n(\mathtt{w})-1)}(\mathtt{w})Y_{\tau(K_n(\mathtt{w}))}(\mathtt{w}),$$

and note that $\tau.\mathtt{w}$ still belongs to $\mathcal{G}_n \cap \mathfrak{L}_n$. Let $\mathfrak{O}(\mathtt{w})$ denote the orbit of $\mathtt{w}$ and let $\mathcal{C}_n$ be the class of orbits of elements of $\mathcal{G}_n \cap \mathfrak{L}_n$. If $\mathfrak{O} \in \mathcal{C}_n$ and if $\nu_{\mathfrak{O}}$ is the conditional distribution of $V_{2,n}(\mathtt{w})$ given that $\mathtt{w} \in \mathfrak{O}$, then $\nu_{\mathfrak{O}}$ is also the image of the uniform probability on $\mathfrak{S}_{K_n(\mathtt{w})}$ by the application $\tau \longmapsto V_{2,n}(\tau.\mathtt{w})$. Thus Lemma 2.1 leads to

$$\mathcal{W}_2(\nu_{\mathfrak{O}}, \mathbb{U}_1) \leq \sqrt{\|\mathfrak{O}\|_\infty / 3}.$$

Then, Proposition 4.5, with the desintegration of $\nu_n$ along $\mathcal{C}_n$, entails

$$\mathcal{W}_2(\nu_n, \mathbb{U}_1) \leq \sqrt{2\tilde{r}/n}.$$

□

Now, let us draw some additional consequences, for good *Lyndon* words, of Definition 7.

**Proposition 6.2.** *A good Lyndon word* $\mathtt{w} \in \mathcal{G}_n \cap \mathcal{L}_n$ *satisfies the following points:*

1. *each long block, by definition a factor of* $\langle \mathtt{w} \rangle$*, is also a factor of* $\mathtt{w}$ *;*
2. *if* $\lfloor \alpha n^{\varepsilon} \rfloor \geq 2$*, there exists a smallest (resp. a second smallest) long block ;*
3. *given a sequence of long blocks of* $\mathtt{w}$*,* $(\zeta_i)_{1 \leq i \leq k}$*,* sorted in increasing lexicographic order*, the sequence* $(\zeta_i \mathtt{v}_i)_{1 \leq i \leq k}$ *is also sorted in increasing lexicographic order, for any sequence of words,* $(\mathtt{v}_i)_{1 \leq i \leq k}$ *;*
4. *the smallest of the long blocks is a prefix of* $\mathtt{w}$ *;*
5. *either the second smallest of the long blocks is a prefix of the standard right factor of* $\mathtt{w}$*, or* $\mathtt{w} \in \mathtt{a}_1 \mathfrak{L}_{n-1}$ *and* $r_n(\mathtt{w}) = 1 - \frac{1}{n}$*.*

*Proof.* Item (1) follows from point **ii.** of Definition 7. Item (2) follows from points **i.** and **iii.** of Definition 7, as **iii.** insures that the prefixes $\mathtt{x}_{[1,\bar{r}]}$ and $\mathtt{y}_{[1,\bar{r}]}$ of two long blocks $\mathtt{x}$ and $\mathtt{y}$ are different. That item (3) holds true follows from **iii.** again, since **iii.** insures not only that $\mathtt{x}$ and $\mathtt{y}$ are not tied, but also that they are not prefixes of each other. Here it can be useful to remember a basic fact about the lexicographic order: if two words $\mathtt{t}_1$ and $\mathtt{t}_2$ have prefixes, respectively $\mathtt{s}_1$ and $\mathtt{s}_2$, such that $\mathtt{s}_1 \prec \mathtt{s}_2$, it does not insure that $\mathtt{t}_1 \prec \mathtt{t}_2$. However, under the additional condition that $\mathtt{s}_1$ is not a prefix of $\mathtt{s}_2$, $\mathtt{s}_1 \prec \mathtt{s}_2$ entails $\mathtt{t}_1 \prec \mathtt{t}_2$. Thus item (3) fails only if some $\zeta_i$ is a prefix of some $\zeta_j$, $i < j$. But this would violate point **iii.** of Definition 7. As a consequence of the definition of Lyndon words, $\mathtt{w}$ begins with one of the long runs of $\mathtt{a}_1$ in $\langle \mathtt{w} \rangle$. This long run is a prefix of some long block due to point **i.** of Definition 7. This, together with item (3), entails item (4).

For item (5), consider the two smallest long blocks, $\zeta_1 \prec \zeta_2$, in the necklace $\langle \mathtt{w} \rangle$, and let $k_1$ and $k_2$ be the lengths of the runs they begin with: $\zeta_i = \mathtt{a}_1^{k_i} \mathtt{u}_i$, $i \in \{1, 2\}$, in which the words $\mathtt{u}_i$ do not begin with the letter $\mathtt{a}_1$. We know that $\mathtt{w}$ begins necessarily with $\zeta_1$, see the considerations leading to item (3). Either the second smallest word in $\langle \mathtt{w} \rangle$, $\mathtt{w}_2$, begins with $\zeta_2$, or $\mathtt{w}_2$ begins with $\mathtt{a}_1^{k_1-1} \mathtt{u}_1$, but, since $\mathtt{a}_1^{k_1-1} \mathtt{u}_1$ or $\zeta_2$ are at least $\lceil 3 \log_{1/\beta} n \rceil$-letters long, they cannot be prefixes of each other, due to point **iii.** of Definition 7. Thus $r_n(\mathtt{w}) = 1 - \frac{1}{n}$ if $\mathtt{a}_1^{k_1-1} \mathtt{u}_1 \prec \zeta_2$, and $r_n(\mathtt{w}) = 1 - \frac{v}{n}$ if $\mathtt{a}_1^{k_1-1} \mathtt{u}_1 \succ \zeta_2$. Here $v$ denotes the position of $\zeta_2$ in $\mathtt{w}$. □

By Proposition 6.2, the smallest of all these factors is $Y_0(\mathtt{w})$. Let $J_{2,n}(\mathtt{w})$ denote the index of the second smallest factor, so that $V_{2,n}$ is given by

$$V_{2,n}(\mathtt{w}) = \frac{1}{n} \sum_{i=0}^{J_{2,n}(\mathtt{w})-1} |Y_i(\mathtt{w})|. \tag{9}$$

If $\mathtt{w} \in \mathtt{a}_1 \mathfrak{L}_{n-1}$,

$$r_n(\mathtt{w}) = 1 - 1/n,$$

(incidentally, we shall see later that this happens with probability $p_1 + o(1)$, according to (13)), while if $\mathtt{w} \in \mathcal{G}_n \cap (\mathfrak{L}_n \backslash \mathtt{a}_1 \mathfrak{L}_{n-1})$, the second smallest block $Y_{J_{2,n}(\mathtt{w})}$, also a long block, is a prefix of the standard right factor, by Proposition 6.2, and

$$r_n(\mathtt{w}) = 1 - V_{2,n}(\mathtt{w}).$$

When $\mathtt{w} \in \mathcal{G}_n$, both cases can be detected by inspection of the two smallest blocks.

Let $\widehat{\mathbb{G}}_n$ denote the conditional probability given $\mathcal{G}_n \cap \mathfrak{L}_n$:

$$\widehat{\mathbb{G}}_n(A) = \frac{\mathbb{P}_n(A \cap \mathcal{G}_n \cap \mathfrak{L}_n)}{\mathbb{P}_n(\mathcal{G}_n \cap \mathfrak{L}_n)} = \frac{\mathbb{L}_n(A \cap \mathcal{G}_n \cap \mathfrak{L}_n)}{\mathbb{L}_n(\mathcal{G}_n \cap \mathfrak{L}_n)}.$$

As in [MZA07], the key point is the invariance of $\widehat{\mathbb{G}}_n$ under uniform random permutation of the blocks $\{Y_1(\mathtt{w}), \ldots, Y_{K_n(\mathtt{w})}(\mathtt{w})\}$.

**Notations 6.3.** *For $\mathtt{w} \in \mathcal{G}_n \cap \mathfrak{L}_n$, and $\tau \in \mathfrak{S}_{K_n(\mathtt{w})}$, we set*

$$\tau.\mathtt{w} = Y_0(\mathtt{w}) Y_{\tau(1)}(\mathtt{w}) \ldots Y_{\tau(K_n(\mathtt{w}))}(\mathtt{w}),$$

*and*

$$\mathfrak{O}(\mathtt{w}) = \{\tau.\mathtt{w} \ : \ \tau \in \mathfrak{S}_{K_n(\mathtt{w})}\}.$$

**Proposition 6.4.** *Assume that $\mathtt{w} \in \mathcal{G}_n \cap \mathfrak{L}_n$, and $\mathtt{w}' \in \mathfrak{O}(\mathtt{w})$: then $\mathtt{w}' \in \mathcal{G}_n \cap \mathfrak{L}_n$ and $\mathtt{w}'$ has the same multiset of blocks as $\mathtt{w}$ (it has the same blocks, with the same multiplicity). As a consequence, for $\mathtt{w}, \mathtt{w}' \in \mathcal{G}_n \cap \mathfrak{L}_n$, either $\mathfrak{O}(\mathtt{w}) = \mathfrak{O}(\mathtt{w}')$ or $\mathfrak{O}(\mathtt{w}) \cap \mathfrak{O}(\mathtt{w}') = \emptyset$.*

This follows directly from Definition 7 and the definition of a code. Let $\mathcal{C}_n = \{\mathfrak{O}(\mathtt{w}) \, ; \ \mathtt{w} \in \mathcal{G}_n \cap \mathfrak{L}_n\}$, and let $\mathfrak{C}_n$ denote the $\sigma$-algebra generated by $\mathcal{C}_n$. Also, let $X(\mathtt{w}) = (X_i(\mathtt{w}))_{i \geq 0}$ be the sequence of blocks of $\mathtt{w}$ sorted in increasing lexicographic order, ended by an infinite sequence of empty words, and let $\Xi(\mathtt{w}) = (\Xi_i(\mathtt{w}))_{i \geq 0}$ be the corresponding sequence of lengths.

**Corollary 6.5.** *The weight $p(.)$, $X$, $\Xi$, $H_n$ and $K_n$ are $\mathfrak{C}_n$-measurable, and*

$$\widehat{\mathbb{G}}_n = \sum_{\mathfrak{O} \in \mathcal{C}_n} \frac{\mathrm{Card}(\mathfrak{O}) \ p(\mathfrak{O})}{\mathbb{P}_n(\mathcal{G}_n \cap \mathfrak{L}_n)} \ \mathbb{U}_{\mathfrak{O}}.$$

*For $\mathfrak{O} \in \mathcal{C}_n$, given that $\mathtt{w} \in \mathfrak{O}$, the ranks of the blocks $(X_i)_{1 \leq i \leq K_n(\mathfrak{O})}$ are uniformly distributed.*

*Proof.* The weight $p(\mathtt{w})$ depends only on the number of letters $a_1$, $a_2$, ... that $\mathtt{w}$ contains, not on the order of the letters in $\mathtt{w}$, so that $p(.)$ is constant on each $\mathfrak{O} \in \mathcal{C}_n$: thus, under $\widehat{\mathbb{G}}_n$, the conditional distribution of $\mathtt{w}$ given that $\mathtt{w} \in \mathfrak{O}$ is $\mathbb{U}_{\mathfrak{O}}$. As a consequence of Proposition 6.4, $\mathcal{C}_n$ is a partition of $\mathcal{G}_n \cap \mathfrak{L}_n$, so the relation in Corollary 6.5 is just the desintegration of $\widehat{\mathbb{G}}_n$ according to its conditional distributions given $\mathcal{C}_n$. $\square$

As in [MZA07, Theorem 6.5], asymptotic independence between $\mathfrak{C}_n$ and $V_{2,n}$ holds under $\widehat{\mathbb{G}}_n$: for a $\mathfrak{C}_n$-measurable $\mathbb{R}$-valued statistic $T_n$ with probability distribution $\chi_n$,

$$\mathcal{W}_2\left((T_n, V_{2,n}), \chi_n \otimes \mathbb{U}_1\right) = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right). \tag{10}$$

In order to prove Theorem 1.4, let $\mu_n$ (resp. $\tilde{\mu}_n$) denote the image of $\mathbb{L}_n$ (resp. of $\widehat{\mathbb{G}}_n$) by $r_n$. Set

$$\mathfrak{L}_n^1 = (\mathcal{G}_n \cap \mathfrak{L}_n) \cap a_1\mathfrak{L}_{n-1} = \mathcal{G}_n \cap a_1\mathfrak{L}_{n-1}, \quad \text{and} \quad \mathfrak{L}_n^2 = (\mathcal{G}_n \cap \mathfrak{L}_n)\backslash\mathfrak{L}_n^1.$$

We remark that:

**i.** if $\mathtt{w} \in \mathfrak{L}_n^1$, $r_n(\mathtt{w}) = 1 - \frac{1}{n}$ holds true[3] ;
**ii.** if $\mathtt{w} \in \mathfrak{L}_n^2$, $r_n(\mathtt{w}) = 1 - V_{2,n}(\mathtt{w})$ ;
**iii.** when $\mathtt{w} \in \mathfrak{L}_n\backslash(\mathcal{G}_n \cap \mathfrak{L}_n)$, the crude bound $0 \le r_n(\mathtt{w}) \le 1$ will prove to be more than sufficient for our purposes.

First, the conditional law $\tilde{\nu}$, given $A$, of a bounded r.v. $X$, defined on a probabilistic space $\Omega$, is Wasserstein-close to its unconditional law $\nu$, if $A$ is close to $\Omega$. More precisely

$$\mathcal{W}_2\left(\nu, \tilde{\nu}\right) \le 2\,\mathbb{P}\left(\Omega\backslash A\right)^{1/2} \left\|X\right\|_\infty. \tag{11}$$

As a consequence, point **iii.**, together with Proposition 4.1, entails that

$$\mathcal{W}_2(\mu_n, \tilde{\mu}_n) = \mathcal{O}\left(n^{-1/2+\varepsilon}\ \log n\right).$$

Thus we shall now work on $\mathcal{G}_n \cap \mathfrak{L}_n$, under $\widehat{\mathbb{G}}_n$, for $\mu_n$ has the same asymptotic behavior as $\tilde{\mu}_n$.

On $\mathcal{G}_n \cap \mathfrak{L}_n$, we have, according to points **i.** and **ii.**,

$$r_n = f_n\left(V_{2,n}, \mathbf{1}_{\mathcal{L}_n^2}\right) = (1 - V_{2,n})\mathbf{1}_{\mathcal{L}_n^2} + \left(1 - \frac{1}{n}\right)(1 - \mathbf{1}_{\mathfrak{L}_n^2}).$$

The $\mathfrak{C}_n$-measurability of $\mathfrak{L}_n^2$ (see [MZA07, Section 7] for more details) and relation (10) entails asymptotic independence between $\mathbf{1}_{\mathfrak{L}_n^2}$ and $V_{2,n}$, and more precisely it entails that

$$\mathcal{W}_2\left((\mathbf{1}_{\mathfrak{L}_n^2}, V_{2,n}), \chi_n \otimes \mathbb{U}_1\right) = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right). \tag{12}$$

[3] Actually, $r_n(\mathtt{w}) = 1 - \frac{1}{n}$ holds true if $\mathtt{w} \in a_k\mathfrak{L}_{n-1}(a_k, a_{k+1}, \ldots, a_n)$, but, since $\mathtt{w} \in \mathcal{G}_n$, $\mathtt{w}$ contains at least one occurrence of the letter $a_1$, which precludes $\mathtt{w} \in a_k\mathfrak{L}_{n-1}(a_k, a_{k+1}, \ldots, a_n)$ for $k \ge 2$.

in which $\chi_n$ denotes the probability distribution of $\mathbf{1}_{\mathfrak{L}_n^2}$. Thus, there exists a probability space, and, defined on this probability space, a couple $(T_n, U)$ with distribution $\chi_n \otimes \mathbb{U}_1$, and a copy of $(\mathbf{1}_{\mathfrak{L}_n^2}, V_{2,n})$ whose $\mathbb{L}^2$ distance satisfies

$$\left\|\mathbf{1}_{\mathfrak{L}_n^2} - T_n\right\|_2^2 + \|M_n - U\|_2^2 = \mathcal{O}\left(\frac{\log n}{n}\right).$$

Set

$$\tilde{r}_n = (1-U)T_n + \left(1 - \frac{1}{n}\right)(1 - T_n).$$

The inequality

$$\left|f_n(d,w) - f_n(d',w')\right|^2 \leq 2\left(\left|d-d'\right|^2 + \left|w-w'\right|^2\right),$$

that holds for $(w, w', d, d') \in [0,1]^4$, entails that

$$\mathcal{W}_2(\tilde{\mu}_n, \tilde{r}_n) = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right).$$

Finally, using an optimal coupling $\left(T_n, \widehat{T}_n\right)$ in which $\widehat{T}_n$ is a Bernoulli random variable with expectation $1 - p_1$, independent of $U$, set

$$\widehat{r}_n = (1-U)\widehat{T}_n + \left(1 - \frac{1}{n}\right)(1 - \widehat{T}_n).$$

As above, we obtain easily

$$\begin{aligned}\mathcal{W}_2(\tilde{r}_n, \widehat{r}_n) &\leq \mathcal{W}_2\left(T_n, \widehat{T}_n\right)\\ &\leq \sqrt{\left|\widehat{\mathbb{G}}_n(\mathfrak{L}_n^2) - (1-p_1)\right|}.\end{aligned}$$

Also

$$(1-U)\widehat{T}_n + (1-\widehat{T}_n) = \widehat{r}_n + \frac{1}{n}(1-\widehat{T}_n)$$

has distribution $\mu$. Thus

$$\mathcal{W}_2(\widehat{r}_n, \mu) \leq \frac{1}{n}.$$

Now

$$\mathbb{P}_n(\mathtt{a}_1\mathfrak{L}_{n-1}) - \mathbb{P}_n(\mathfrak{L}_n \backslash \mathcal{G}_n) \leq \mathbb{P}_n(\mathfrak{L}_n^1) \leq \mathbb{P}_n(\mathtt{a}_1\mathfrak{L}_{n-1}).$$

So by Proposition 4.1 and the fact that $\mathbb{P}_n(\mathfrak{L}_n) = \frac{1}{n}(1 - O(\beta^{n/2}))$, we obtain

$$\left|\widehat{\mathbb{G}}_n(\mathfrak{L}_n^1) - p_1\right| = \mathcal{O}\left((\log n)^2\, n^{2\varepsilon - 1}\right) \tag{13}$$

and

$$\mathcal{W}_2(\tilde{r}_n, \hat{r}_n) = \mathcal{O}\left(n^{-1/2+\varepsilon}\log n\right).$$

With (12), this yields

$$\mathcal{W}_2(\mu_n, \mu) = \mathcal{O}\left(n^{-1/2+\varepsilon}\log n\right).$$

Since $0 \leq r_n \leq 1$, convergence of moments follows. □

## References


ABT93. Richard Arratia, A. D. Barbour, and Simon Tavaré, *On random polynomials over finite fields*, Math. Proc. Cambridge Philos. Soc. **114** (1993), no. 2, 347–368. MR MR1230136 (95a:60011)

ABT99. _____, *On Poisson-Dirichlet limits for random decomposable combinatorial structures*, Combin. Probab. Comput. **8** (1999), no. 3, 193–208. MR MR1702562 (2001b:60029)

ABT03. _____, *Logarithmic Combinatorial Structures: a probability approch*, European Mathematical Society Zurich, 2003.

BCN05. Frédérique Bassino, Julien Clément, and Cyril Nicaud, *The standard factorization of Lyndon words: an average point of view*, Discrete Math. **290** (2005), no. 1, 1–25. MR MR2116634 (2005j:68084)

BD92. Dave Bayer and Persi Diaconis, *Trailing the dovetail shuffle to its lair*, Ann. Appl. Probab. **2** (1992), no. 2, 294–313. MR MR1161056 (93d:60014)

Bil99. P. Billingsley, *Probability and measure*, John Wiley & Sons, New York, 1999.

BP85. Jean Berstel and Dominique Perrin, *Theory of codes*, Pure and Applied Mathematics, vol. 117, Academic Press, Inc., Orlando, FL, 1985. MR 797069

BP07. _____, *The origins of combinatorics on words*, European J. Combin. **28** (2007), no. 3, 996–1022.

CG96. J.H. Conway and R.K. Guy, *The book of numbers*, Springer-Verlag, 1996.

DMP95. P. Diaconis, M.J. McGrath, and J. Pitman, *Riffle shuffles, cycles, and descents*, Combinatorica **15, no. 1** (1995), 11–29.

GR93. Ira M. Gessel and Christophe Reutenauer, *Counting permutations with given cycle structure and descent set*, J. Combin. Theory Ser. A **64** (1993), no. 2, 189–215. MR MR1245159 (95g:05006)

GS12. Joseph Yossi Gil and David Allen Scott, *A bijective string sorting transform*, CoRR **abs/1201.3077** (2012).

Han93. Jennie C. Hansen, *Factorization in* $\mathbf{F}_q[x]$ *and Brownian motion*, Combin. Probab. Comput. **2** (1993), no. 3, 285–299. MR MR1264035 (95f:11056)

Han94. _____, *Order statistics for decomposable combinatorial structures*, Random Structures Algorithms **5** (1994), no. 4, 517–533. MR MR1293077 (96f:60010)

Kal97. O. Kallenberg, *Foundations of Modern Probability*, Springer series in Statistics Probability and its applications, 1997.

Kin75. J. F. C. Kingman, *Random discrete distributions*, Journal of the Royal Statistical Society. Series B (Methodological) **37** (1975), no. 1, 1–22.

Lot97. M. Lothaire, *Combinatorics on words*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997.

Lot02. M. Lothaire, *Algebraic Combinatorics on Words*, vol. 90 of Encyclopedia of mathematics and its applications, Cambridge University Press, 2002.

Lyn54. R. Lyndon, *On Burnside problem I*, Trans. American Math. Soc. **77** (1954), 202–215.

McC65. J. W. T. McCloskey, *A model for the distribution of individuals by species in an environment*, ProQuest LLC, Ann Arbor, MI, 1965, Thesis–MSU.

MZA07. R. Marchand and E. Zohoorian Azad, *Limit law of the length of the standard right factor of a Lyndon word*, Combin. Probab. Comput. **16** (2007), no. 3, 417–434. MR MR2312436 (2008e:68120)

Oka58. M. Okamoto, *Some inequalities related to the partial sum of binomial probabilities*, Ann. Inst. Statist. Math. **10** (1958), 29–35.

PPY92. Mihael Perman, Jim Pitman, and Marc Yor, *Size-biased sampling of Poisson point processes and excursions*, Probab. Theory Related Fields **92** (1992), no. 1, 21–39. MR MR1156448 (93d:60088)

Rac91. S.T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, U.K., 1991.

Reu93. C. Reutenauer, *Free lie algebras*, Oxford Science Publications, 1993.

SW09. G.R. Shorack and J.A. Wellner, *Empirical processes with applications to statistics*, Society for Industrial Mathematics, 2009.

## 7 Runs statistics: proofs

### 7.1 Asymptotically almost sure properties in $\mathcal{A}^n$ vs $\mathcal{P}_n$: proof of Lemma 3.1

This proof rephrases in probabilistic terms some results of [Reu93, Section 7.1], to which the reader is referred for definitions. Let us define two sequences of subsets of $\mathcal{A}^n$,

$$\mathcal{A}_{n,k} = \left\{ \mathtt{w} \in \mathcal{A}^n \mid \exists \mathtt{v} \in \mathcal{A}^k \text{ such that } \mathtt{w} = \mathtt{v}^{n/k} \right\},$$
$$\mathcal{P}_{n,k} = \mathcal{A}_{n,k} \setminus \left( \bigcup_{1 \le i < k} \mathcal{A}_{n,i} \right),$$

with probabilities $\nu_k = \mathbb{P}_n(\mathcal{A}_{n,k})$ and $\xi_k = \mathbb{P}_n(\mathcal{P}_{n,k})$, respectively. Clearly

$$\mathcal{A}_{n,n} = \mathcal{A}^n, \qquad \mathcal{P}_{n,n} = \mathcal{P}_n.$$

Also, if $k|n$, $(\mathcal{P}_{n,i})_{i|k}$ is a partition of $\mathcal{A}_{n,k}$ (else, both $\mathcal{A}_{n,k}$ and $\mathcal{P}_{n,k}$ are empty). Thus

$$\nu_k = \sum_{d|k} \xi_d,$$

and, by the Möbius inversion formula,

$$\xi_k = \sum_{d|k} \mu(d) \nu_{k/d}, \tag{14}$$

in which $\mu(d)$ denotes the Möbius function. On the other hand, when $k|n$,

$$\begin{aligned}
\nu_k &= \sum_{\mathtt{w} \in \mathcal{A}_{n,k}} p(\mathtt{w}) \\
&= \sum_{\mathtt{v} \in \mathcal{A}^k} p(\mathtt{v})^{n/k} \\
&= \sum_{\sum_i r_i = k} \binom{k}{r_1, r_2, \dots} \left(p_1^{r_1} p_2^{r_2} \dots\right)^{n/k} \\
&= \|p\|_{n/k}^n .
\end{aligned}$$

Specializing (14) to $k = n$, we obtain

$$\mathbb{P}_n(\mathcal{P}_n) = \sum_{d|n} \mu(d) \|p\|_d^n . \tag{15}$$

Let the set of divisors of $n$ be $\{1 < d_1 < d_2 < \dots < d_\ell = n\}$. Then, by (15),

$$\begin{aligned}
\left| \mathbb{P}_n(\mathcal{P}_n) - 1 + \|p\|_{d_1}^n \right| &\le (\ell - 1) \|p\|_{d_2}^n \\
&\le (n-2) \|p\|_{d_2}^n ,
\end{aligned}$$

if $n$ is not prime. Else $\mathbb{P}_n(\mathcal{P}_n) = 1 - \|p\|_{d_1}^n$. In any case, $\left|\mathbb{P}_n(\mathcal{P}_n) - 1 + \|p\|_{d_1}^n\right|$ is a $o\left(\|p\|_{d_1}^n\right)$, and, since $d_1 \geq 2$,

$$\mathbb{P}_n(\mathcal{P}_n^c) = \mathcal{O}\left(\|p\|_2^n\right). \tag{16}$$

Lemma 3.1 is a direct consequence of

$$\mathbb{L}_n(A) = \frac{\mathbb{P}_n(\pi^{-1}(A))}{\mathbb{P}_n(\mathcal{P}_n)},$$

and of (16).

### 7.2 Alternative representations for $\mathbb{P}_n$

The asymptotic behaviour of the factorizations of $n$-letters general random words is predicted by the lengths and positions of runs of the letter $\mathtt{a}_1$, provided that these lengths and positions satisfy a set of properties that hold true, but for a vanishing probability as $n$ grows, for what we call *good words* (see Definition 7). Thus, a random word $\mathtt{w} \in \mathcal{A}^n$ is a good word, or not, depending on $\varphi_n(\mathtt{w})$. The proof that the probability of bad words vanishes relies on two descriptions, given in this section, of the probability distribution $\mathbb{B}_n$ of $\varphi_n(\mathtt{w})$ under $\mathbb{P}_n$.

Under $\mathbb{P}_n$, $\varphi_n$ is a $\mathcal{B}^n$-valued random variable, a random word, a sequence of $n$ independant symbols, each of them being a $\mathtt{0}$ with probability $p_1$, a $\mathtt{1}$ with probability $1 - p_1$ : $\mathbb{B}_n$ denote the probability distribution of $\varphi_n$, i.e. the push-forward of $\mathbb{P}_n$ by $\varphi_n$. For the next proofs, however, $\mathbb{B}_n$ shall be seen as the push-forward of two probability measures on the set $\mathcal{B}^{\mathbb{N}}$ of infinite words, by the truncation operation $\psi_n$ defined, for $\omega \in \mathcal{B}^{\mathbb{N}}$, by:

$$\omega = \omega_1\omega_2\omega_3\ldots \quad \longrightarrow \quad \psi_n(\omega) = \omega_{[1,n]}.$$

First, $\mathbb{B}_n$ is the probability distribution of $\psi_n$ under the product measure

$$\mathbb{B} = \left(p_1\delta_{\mathtt{0}} + (1-p_1)\delta_{\mathtt{1}}\right)^{\otimes\mathbb{N}}.$$

Next, let $\eta = (\eta_n)_{n\geq 1}$ (resp. $\theta = (\theta_n)_{n\geq 1}$) be a sequence of independent geometric random variables with expectation $(1-p_1)^{-1}$ (resp. with expectation ${p_1}^{-1}$), defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\xi$ be a Bernoulli random variable with parameter $1 - p_1$, and assume that $\xi$, $\eta$ and $\theta$ are independent. For $m \geq 1$, set:

$$S_m = \sum_{k=1}^{m}(\eta_k + \theta_k),$$

and consider the infinite random word

$$\Upsilon = \begin{cases} \mathtt{0}^{\eta_1}\mathtt{1}^{\theta_1}\mathtt{0}^{\eta_2}\mathtt{1}^{\theta_2}\ldots & \text{if } \xi = 0, \\ \mathtt{1}^{\theta_1}\mathtt{0}^{\eta_1}\mathtt{1}^{\theta_2}\mathtt{0}^{\eta_2}\ldots & \text{if } \xi = 1, \end{cases}$$

that is, $\Upsilon$ is defined by the sequences $\eta$ and $\theta$ of its runs' lengths. Then

**Proposition 7.1.** *The probability distribution of $\Upsilon$ is $\mathbb{B}$. As a consequence, $\Upsilon_{[1,n]}$, $\psi_n$ and $\varphi_n$ have the same distribution $\mathbb{B}_n$.*

*Proof.* We already know that, for all $n$, $\psi_n$ and $\varphi_n$ have the same distribution $\mathbb{B}_n$. We need to prove that $\mathbb{B}_n$ is also the distribution of $\Upsilon_{[1,n]}$ : for any $\ell \geq 1$, and any finite word $\mathtt{w} \in \mathcal{B}^\ell$, for instance of the form $\mathtt{0v10}$, i.e. having an even number of runs, say $2m$, followed by $\mathtt{0}$, the first run being thus a run of $\mathtt{0}$s, we can write

$$\mathtt{w} = \mathtt{0}^{k_1}\mathtt{1}^{\ell_1}\mathtt{0}^{k_2}\mathtt{1}^{\ell_2}\dots\mathtt{0}^{k_m}\mathtt{1}^{\ell_m}\mathtt{0},$$
$$\ell - 1 = s_m = \sum_{i=1}^{m}(k_i + \ell_i),$$

and we have

$$\begin{aligned}\mathbb{B}_\ell(\{\mathtt{w}\}) &= p_1^{1+k_1+\dots+k_m}(1-p_1)^{\ell_1+\dots+\ell_m}\\ &= p_1\prod_{i=1}^{m}p_1^{k_i-1}(1-p_1)\prod_{i=1}^{m}(1-p_1)^{\ell_i-1}p_1\\ &= \mathbb{P}(\xi=0)\prod_{i=1}^{m}\mathbb{P}(\eta_i = k_i, \theta_i = \ell_i)\\ &= \mathbb{P}\left(\Upsilon_{[1,\ell]} = \mathtt{w}\right).\end{aligned}$$

For $\xi = 1$, and for $\mathtt{w}$ of the form $\mathtt{1v01}$, or even when $\mathtt{w}$ does not end with the beginning of a new run, the computation is similar, in the last case using $\mathbb{P}(\theta_i > k_i) = (1-p_1)^{k_i}$, for instance. This also entails that the probability distribution of $\Upsilon$ is $\mathbb{B}$. □

## 7.3 Number of runs: proof of Lemma 3.2

**Proposition 7.2.** $\mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} < m\right) \leq \mathbb{P}(S_m > n)$.

*Proof.* If $N_n^{(\mathtt{0})} \circ \psi_n(\omega) < m$, the $m$th run of $\mathtt{0}$s of $\psi(\omega)$ begins after its $n$th letter. According to Proposition 7.1, this last event has the same probability for $\omega$ or for the infinite random word $\Upsilon$, but the $m$th run of $\mathtt{0}$s of $\Upsilon$ begins either at the position $1+S_{m-1}(< S_m)$, or at the position $\theta_m+1+S_{m-1}(\leq S_m)$, according to the value of $\xi$. □

Thus, by Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} < m\right) &\leq \mathbb{P}(S_m > n)\\ &\leq \frac{\mathrm{Var}(S_m)}{(n - m\mathbb{E}[\eta_k + \theta_k])^2}.\end{aligned}$$

Since $\mathbb{E}\left[\eta_k + \theta_k\right] = \sigma^{-2}$, with the choice $m = an + b$, $b \in \mathbb{R}$, $a < \sigma^2$, we obtain

$$\mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} < an + b\right) = \mathcal{O}\left(n^{-1}\right). \tag{17}$$

For a primitive word $\mathtt{w}$, $N_n^{(\mathtt{a}_1)}(\mathtt{w}) \leq N_n^{(\mathtt{a}_1)}(\pi(\mathtt{w})) + 1$, thus

$$\mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} \circ \pi < an + b\right) \leq \mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} < an + b + 1\right).$$

With that in view, Lemma 3.1 extends (17) to $\mathbb{L}_n$. Note that, with some additional work, one obtains easily, for any $\varepsilon > 0$,

$$\mathbb{P}_n\left(N_n^{(\mathtt{a}_1)} < \sigma^2(1-\varepsilon)n\right) = \mathcal{O}\left(e^{-\eta n}\right),$$

for a suitable $\eta > 0$. However, the weaker Lemma 3.2 suits our aims here.

### 7.4 Number of long runs : proof of Lemma 3.3

Given that the probability of a run of $\mathtt{a}_1$ longer than $(1-\varepsilon)\log_{1/p_1} n$ is approximately $n^\varepsilon/n$, an (admittedly flawed) argument suggests that in an $n$-letters long random word, there are many (that is, $\Theta\left(n^\varepsilon\right)$) runs longer than $(1-\varepsilon)\log_{1/p_1} n$. For a more precise and concise argument, let $\hat{N}_n = N_n^{(\mathtt{0})}\left(\Upsilon_{[1,n]}\right)$ (resp. $\hat{H}_n$) denote the number of runs of $\mathtt{0}$'s in the word $\Upsilon_{[1,n]}$ (resp. the number of runs with length at least $(1-\varepsilon)\log_{1/p_1} n$). Each run of the letter $\mathtt{a}_1$ in the word $\mathtt{w}$ is matched with a run of $\mathtt{0}$ of the same length in the word $\varphi(\mathtt{w})$, thus, according to Proposition 7.1, $(N_n^{(\mathtt{a}_1)}, H_n)$ (defined on $(\mathcal{A}^n, \mathbb{P}_n)$) and $(\hat{N}_n, \hat{H}_n)$ (defined on $(\Omega, \mathcal{F}, \mathbb{P})$) have the same probability distribution.

We let, for $i \geq 1$,

$$B_i = \mathbf{1}_{\{\eta_i \geq (1-\varepsilon)\log_{1/p_1} n\}}, \quad \hat{S}_m = \sum_{i=1}^{m} B_i.$$

The sequence of lengths of runs of $\mathtt{0}$ in $\Upsilon_{[1,n]}$ differs from $(\eta_i)_{1 \leq i \leq \hat{N}_n}$, possibly, only at the last term, due to the truncation of $\Upsilon$. As a consequence,

$$\hat{H}_n \geq \hat{S}_{\hat{N}_n - 1}. \tag{18}$$

Note also that, under $\mathbb{P}$, $(B_i)_{i \geq 1}$ is a Bernoulli process, and that its parameter $p(n,\varepsilon)$ satisfies $n^{\varepsilon-1} \leq p(n,\varepsilon) \leq n^{\varepsilon-1}/p_1$. Thus relation (18), with Lemma 3.2, entails that, under $\mathbb{P}$, $\hat{H}_n$ is, roughly speaking, stochastically larger than the binomial distribution with parameters $an + b$ and $p(n,\varepsilon)$, provided that $a < \sigma^2$. More precisely,

$$\begin{aligned}\mathbb{P}\left(\hat{H}_n < \alpha n^\varepsilon\right) &\leq \mathbb{P}\left(\hat{N}_n < an + 2\right) + \mathbb{P}\left(\hat{N}_n \geq an + 2 \text{ and } \hat{S}_{\hat{N}_n - 1} < \alpha n^\varepsilon\right) \\ &\leq \mathbb{P}\left(\hat{N}_n < an + 2\right) + \mathbb{P}\left(\hat{S}_{\lceil an \rceil} < \alpha n^\varepsilon\right).\end{aligned}$$

Lemma 3.2 takes care of the first term on the right hand side. For the second term, by Okamoto's inequality [Oka58, Th. 2(ii)], a binomial random variable $S_{n,p}$ with parameters $n$ and $p < 1/2$ satisfies :

$$\mathbb{P}\left(S_{n,p} - pn \leq -cn\right) < \exp\left(-nc^2/(2pq)\right).$$

As a consequence

$$\begin{aligned}\mathbb{P}\left(\hat{S}_{\lceil an\rceil} < \alpha n^{\varepsilon}\right) &\leq \mathbb{P}\left(\hat{S}_{\lceil an\rceil} - \lceil an\rceil p(n,\varepsilon) < \alpha n^{\varepsilon} - \lceil an\rceil p(n,\varepsilon)\right)\\ &\leq \mathbb{P}\left(\hat{S}_{\lceil an\rceil} - \lceil an\rceil p(n,\varepsilon) < (\alpha - a)n^{\varepsilon}\right),\end{aligned}$$

and, for $\alpha < a$ and $\lceil an\rceil \leq 2an$, Okamoto's inequality entails that

$$\mathbb{P}\left(\hat{S}_{\lceil an\rceil} < \alpha n^{\varepsilon}\right) \leq \exp\left(-\frac{(a-\alpha)^2 p_1}{4a} n^{\varepsilon}\right).$$

The first statement of Lemma 3.3 follows. For the proof of the second statement, we note that if $\mathtt{w}$ is a primitive word,

$$H_n \circ \pi(\mathtt{w}) \geq H_n(\mathtt{w}) - 1, \tag{19}$$

with equality when $\mathtt{w}$ begins and ends with long runs. Together with Lemma 3.1, it entails that

$$\begin{aligned}\mathbb{L}_n\left(H_n < \alpha n^{\varepsilon}\right) &\leq \mathbb{P}_n\left(\{\mathtt{w} \in \mathcal{P}_n, H_n \circ \pi(\mathtt{w}) < \alpha n^{\varepsilon}\}\right) + \mathcal{O}\left(\beta^{n/2}\right)\\ &\leq \mathbb{P}_n\left(H_n < \alpha n^{\varepsilon} + 1\right) + \mathcal{O}\left(\beta^{n/2}\right).\end{aligned}$$

and the Lemma follows since, as above, $\mathbb{P}\left(\hat{S}_{\lceil an\rceil} < 1 + \alpha n^{\varepsilon}\right) = \mathcal{O}\left(n^{-1}\right)$.

### 7.5 Large values of the longest runs : proof of Lemma 3.4

Recall that $M_n^{(0)}(\varphi(\mathtt{w}))$ (resp. $M_n^{(1)}(\varphi(\mathtt{w}))$) denote the length of the largest runs of the letter $\mathtt{a}_1$ (resp. non-$\mathtt{a}_1$ letters) of some word $\mathtt{w} \in \mathcal{A}^n$, see Definition 5, and set

$$A_{1,n} = \left\{M_n^{(1)} \circ \varphi_n \geq 2\log_{1/(1-p_1)} n\right\}, \quad A_{0,n} = \left\{M_n^{(0)} \circ \varphi_n \geq 2\log_{1/p_1} n\right\}.$$

In order to prove that $\mathbb{P}_n(A_{i,n})$ or $\mathbb{L}_n(A_{i,n})$ are $\mathcal{O}\left(n^{-1}\right)$, we use again Proposition 7.1 then Lemma 3.1 : for any $\mathtt{i} \in \{\mathtt{0},\mathtt{1}\}$, let $\hat{M}_n^{(\mathtt{i})}$ denote the length of the largest run of $\mathtt{i}$'s of the word $\Upsilon_{[1,n]}$, so that, by Proposition 7.1, $(\hat{M}_n^{(0)}, \hat{M}_n^{(1)})$ has the same probability distribution as $(M_n^{(0)} \circ \varphi_n, M_n^{(1)} \circ \varphi_n)$. As a consequence, for $y > 0$, we have:

$$\begin{aligned}\mathbb{P}_n(M_n^{(0)} \circ \varphi_n \leq y) &= \mathbb{P}(\hat{M}_n^{(0)} \leq y)\\ &\geq \mathbb{P}(\forall i \in \{1,\ldots,n\},\ \eta_i \leq y)\\ &\geq \left(1 - p_1^{\lfloor y\rfloor}\right)^n,\end{aligned}$$

the first inequality due to $\hat{N}_n \leq n$. Choosing $y = \left\lceil 2\log_{1/p_1} n \right\rceil - 1$, we obtain that

$$\mathbb{P}_n\left(A_{0,n}\right) = \mathcal{O}\left(n^{-1}\right).$$

Note that for a primitive word $\mathtt{w}$, $M_n^{(0)} \circ \varphi_n \circ \pi(\mathtt{w})$ differs from $M_n^{(0)} \circ \varphi_n(\mathtt{w})$ only if the word $\mathtt{w}$ begins *and* ends with the letter $\mathtt{a}_1$. More precisely, we have, according to Definition 5,

$$\begin{aligned} M_n^{(0)} \circ \varphi_n \circ \pi(\mathtt{w}) &= \max\{M_n^{(0)} \circ \varphi_n(\mathtt{w}), (W_1(\mathtt{w}) + W_{N_n} \circ \varphi_n(\mathtt{w}))\, \mathbb{1}_{\mathtt{w}_1=\mathtt{a}_1=\mathtt{w}_n}\} \\ &\leq \max\left\{M_n^{(0)}(\mathtt{w}), (W_1(\mathtt{w})\mathbb{1}_{\mathtt{w}_1=\mathtt{a}_1} + W_{N_n}(\mathtt{w})\mathbb{1}_{\mathtt{w}_n=\mathtt{a}_1})\right\}, \end{aligned}$$

Since $\mathbb{P}_n$ is invariant under words' reversal, $(W_1, \mathtt{w}_1)$ and $(W_{N_n}, \mathtt{w}_n)$ have the same probability distribution. Thus, from Lemma 3.1, we deduce that

$$\mathbb{L}_n\left(A_{0,n}\right) \leq 2\,\mathbb{P}_n\left(W_1\mathbb{1}_{\mathtt{w}_1=\mathtt{a}_1} \geq \log_{1/p_1} n\right) + \mathbb{P}_n\left(A_{0,n}\right) + \mathcal{O}\left(\|p\|_2^n\right).$$

which leads to the desired bound for $\mathbb{L}_n\left(A_{0,n}\right)$, since, for $1 \leq k \leq n$,

$$\mathbb{P}_n\left(W_1\mathbb{1}_{\mathtt{w}_1=\mathtt{a}_1} \geq k\right) = p_1^k.$$

Similar arguments hold for $\mathbb{P}_n\left(A_{1,n}\right)$ and $\mathbb{L}_n\left(A_{1,n}\right)$.

### 7.6 $\Lambda_n$ is large : proof of Proposition 5.1

Proposition 5.1 asserts that with a probability close to 1, $\Lambda_n$ is at least of order $\log n$. By invariance of $\mathbb{G}_n$ under uniform random permutations of the blocks $Y_i$, the sequence of ranks of the long blocks is, conditionally given that $H_n = k$, a random uniform permutation of $\mathfrak{S}_k$. Thus the conditional distribution of the number $\Lambda_n$ of Lyndon factors obtained this way, given that $H_n = k$, has the same law as the number of records (or of cycles) of a uniform random permutation in $\mathfrak{S}_k$ (see [ABT03, Ch. 1] or [Lot02, Ch. 11]), with generating function

$$\frac{1}{k!}\, x(x+1)(x+2)\ldots(x+k-1) = \frac{1}{k!} \sum_{0\leq j\leq k} \begin{bmatrix} k \\ j \end{bmatrix} x^j,$$

in which $\left[\begin{smallmatrix} k \\ j \end{smallmatrix}\right]$ is a Stirling number of the first kind. We can thus describe the conditional law of $\Lambda_n$ as follows : consider a sequence $B = (B_i)_{i\geq 1}$ of independent Bernoulli random variables with respective parameters $1/i$, $B$ and $H_n$ being independent. Set

$$\tilde{S}_n = \sum_{1\leq i\leq n} B_i$$

and

$$\tilde{\Lambda}_n = \tilde{S}_{H_n} = \sum_i B_i\ \mathbb{1}_{1\leq i\leq H_n}.$$

Then $\Lambda_n$ and $\tilde{\Lambda}_n$ have the same distribution, and we shall use the notation $\Lambda_n$ for both of them.

The $m$-th harmonic number has the asymptotic expansion

$$\sum_{i=1}^{m} 1/i \;=\; \mathfrak{H}_m \;=\; \ln m + \gamma + \frac{1}{2m} - \frac{1}{12m^2} + \dots,$$

in which $\gamma$ is the Euler-Mascheroni constant (see [CG96]). We have

$$\mathbb{E}(\tilde{S}_n) = \mathfrak{H}_n \qquad \text{and} \qquad \mathrm{Var}(\tilde{S}_n) = \mathfrak{H}_n - \sum_{i=1}^{n} \frac{1}{i^2}.$$

By Lemma 3.3 :

$$\mathbb{P}_n\left(\Lambda_n < \varepsilon \log n/3\right) \leq \mathbb{P}_n\left(\Lambda_n < \varepsilon \log n/3 \mid H_n \geq \alpha n^{\varepsilon}\right) + \mathcal{O}\left(n^{-1}\right).$$

In addition

$$\begin{aligned}
\mathbb{P}_n\left(\Lambda_n < \varepsilon \log n/3 \mid H_n \geq \alpha n^{\varepsilon}\right) &\leq \mathbb{P}_n\left(\tilde{S}_{\alpha n^{\varepsilon}} < \varepsilon \log n/3\right) \\
&\leq \mathbb{P}_n\left\{\left|\tilde{S}_{\alpha n^{\varepsilon}} - \mathbb{E}(\tilde{S}_{\alpha n^{\varepsilon}})\right| \geq \varepsilon \log n/2\right\} \\
&= \mathcal{O}\left(\frac{1}{\log n}\right),
\end{aligned}$$

in which the second inequality holds true provided that

$$\mathbb{E}(\tilde{S}_{\alpha n^{\varepsilon}}) - \varepsilon \log n/3 \geq \varepsilon \log n/2,$$

i.e. for $n$ large enough, and the last equality follows from the Bienaymé-Chebyshev inequality and from the asymptotic behavior of $\mathfrak{H}_n$.