Gapped Pattern Statistics

Philippe Duchon[1], Cyril Nicaud[2], and Carine Pivoteau[2]

1 **Univ. Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France**
CNRS, LaBRI, UMR 5800, F-33400 Talence, France
philippe.duchon@u-bordeaux.fr

2 **Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE Paris, UPEM, F-77454 Marne-la-Vallée, France**
cyril.nicaud@u-pem.fr,pivoteau@univ-mlv.fr

Abstract

We give a probabilistic analysis of parameters related to α-gapped repeats and palindromes in random words, under both uniform and memoryless distributions (where letters have different probabilities, but are drawn independently). More precisely, we study the expected number of maximal α-gapped patterns, as well as the expected length of the longest α-gapped pattern in a random word.

1998 ACM Subject Classification G.2.1 Combinatorics

Keywords and phrases combinatorics on words, α-gapped repeats, random words, memoryless sources, analytic combinatorics.

Digital Object Identifier 10.4230/LIPIcs.CPM.2017.40

1 Introduction

In this article, we are interested in the combinatorial aspects of the notion of α-gapped repeat and α-gapped palindromes [10, 7, 4]. An α-gapped repeat in a word is a factor of the form uvu, where u and v are words with $|uv| \leq \alpha|u|$. More precisely, such a pattern is essentially a repetition of u, but the second occurrence is not too far away from the first one. The definition for palindromes is similar, as we are looking for factors of the form $uv\overline{u}$ instead, where $\overline{u}$ is the reverse of u. The study of gapped patterns (see also [1, 12]) finds most of its motivation in bioinformatics. Recent works show that these patterns can be found in linear time [11, 17, 6], and there cannot be more than a linear number of them [2, 7]. Note that α-gapped repeats are also called fractional powers [16]: uvu is an α-gapped repeat if and only if it is a fractional power of uv with exponent at least $1 + \alpha^{-1}$.

When looking at patterns in words, there are usually two main categories of questions: providing efficient algorithms to find a specific set of patterns and studying the combinatorics of words with a focus on the appearance (or avoidance) of these patterns. These two points of view are of course directly related, as insights on the combinatorial properties often yield ideas for building new efficient algorithms.

In the sequel, we propose a probabilistic analysis of parameters related to α-gapped repeats and palindromes; more precisely, we answer the following questions:

- What is the expected number of α-gapped patterns in a random word?
- What is the expected length of the longest α-gapped pattern in a random word?

This only makes sense if one specifies what is meant by a random word, *i.e.*, what the distribution on words is. We first consider the uniform distribution, which often serves as an introductory example for the techniques we use and can provide, for instance, useful elements for average analysis of algorithms, while still being mathematically tractable. We

also consider memoryless sources, which give a more general, yet simple distribution where all letters are not constrained to have identical frequencies. In this model, each letter is drawn independently following a fixed, but possibly biased, distribution on the alphabet. In particular, we exhibit a noteworthy behavior on the longest α-gapped repeat: if each letter a_i has probability π_i, then a long random word of length n has about $\pi_i n$ occurrences of each letter a_i; however, in a long α-gapped repeat uvu, the frequencies of the letters in the u parts of the factor do not follow this typical distribution (see section 4.2).

Our work follows several other combinatorial and probabilistic results obtained for different kinds of patterns in words, such as the expected number of runs [14], the expected total run length [8], the expected number of distinct palindromic factors [15], *etc.* We use both techniques from analytic combinatorics, based on the definition of generating series for gapped patterns in words, as in [13], and classical discrete probabilities.

2 Preliminaries

For any two nonnegative integers i, j, let $[i, j]$ denote the integer interval $\{i, \dots, j\}$. By convention, $[i, j] = \emptyset$ if $j < i$. Let also $[i]$ denote the integer interval $[1, i]$.

In the sequel we consider words on a finite alphabet A, of cardinality $k \geq 2$. We assume the reader is familiar with the classical definitions on words [3], such as prefixes, suffixes, and factors. For $w \in A^*$ of length n and $i \in [n]$, let w_i (or $w[i]$) denote the i-th letter of w, with the convention that positions start at 1. The last letter of w is therefore $w_{|w|}$. Let also $w[i, j] = w_i \cdots w_j$ denote the factor of w that starts at position i and ends at position j, with $w[i, j] = \varepsilon$ if $i > j$ or if i or j is not in $[n]$. The factor of w of length ℓ that starts at position i is $w[i, i + \ell - 1]$. For a given length ℓ, a position i in w is *valid* if $i + \ell - 1 \leq n$.

A *gapped repeat* in a word w of length n is a triple (i, u, v), where $i \in [n]$ and u and v are nonempty words, such that the factor of w of length $|uvu|$ starting at position i is uvu. For a given real $\alpha \geq 1$, it is an α-gapped repeat if $|uv| \leq \alpha|u|$. A gapped repeat (i, u, v) of w is *maximal* if, when the positions exist, $w_{i-1} \neq w_{i+|uv|-1}$ and $w_{i+|uvu|} \neq w_{i+|u|}$, *i.e.*, the gapped repeat cannot be extended to the left or to the right.

Similar notions can be defined for palindromes. Under the same conditions for i, u and v, a triple (i, u, v) is an *α-gapped palindrome* if the factor of length $|uv\overline{u}|$ starting at position i in w is $uv\overline{u}$, where $\overline{u} = u_{|u|} \cdots u_1$ denote the reverse of u. It is an *α-gapped palindrome* if $|uv| \leq \alpha|u|$ and maximal if $w_{i-1} \neq w_{i+|uvu|}$ (when they exist) and either $|v| = 1$ or $v_1 \neq v_{|v|}$.

▶ **Example 1.** Consider $w = aab\underline{abbbab}ab$ and $\alpha = 2$. The triple $(4, ab, bb)$ is an α-gapped repeat, but it is not maximal since it can be extended to the left to form $(3, bab, b)$.

▶ Remark. In the sequel, we only consider α-gapped patterns (repeats or palindromes) for rational $\alpha \geq 1$. This really matters for Section 3 only, as the other results hold for any real $\alpha \geq 1$. It is also convenient to consider $\beta := \alpha - 1$ in most computations, as it changes the condition into $|u| \leq \beta|v|$, and we therefore use this notation from now on.

The *uniform distribution* on a finite set E is the probability π defined for all $e \in E$ by $\pi(e) = \frac{1}{|E|}$. By a slight abuse of notation, we will speak of the *uniform distribution on A^** to denote the sequence $(\pi_n)_{n\geq 0}$ of uniform distributions on A^n. For instance, if $A = \{a, b, c\}$, then each element of A^n has probability 3^{-n} under this distribution.

Another very classical distribution on A^n is the *memoryless distribution of probability* π, where π is a probability on the alphabet A. Under this distribution, the probability of a word $w = w_1 \cdots w_n \in A^n$ is $\mathbb{P}_n(w) = \pi(w_1) \cdots \pi(w_n)$. This distribution can also be seen as generating each letter of the word independently, following π.

It is convenient to fix a total order $a_1 < \ldots < a_k$ on A and to define $\pi_i = \pi(a_i)$, for all $i \in [k]$. We also see π as a vector $\vec{\pi} = (\pi_1, \ldots, \pi_k)$ of $[0,1]^k$. This notation will be used repeatedly in the sequel.

3 Number of gapped patterns

In this section, we compute the average number of maximal α-gapped patterns (repeats or palindromes) in random words of length n under a memoryless distribution. Our main tool is writing exact generating functions, which happen to be rational fractions; the asymptotic behavior is then obtained by using standard theorems of analytic combinatorics [5].

3.1 Framework

Let $A = \{a_1, \ldots, a_k\}$ be an alphabet and, for every $i \in [k]$, let z_i be a formal variable (associated with the letter a_i). To each word $w \in A^*$ we associate a monomial $c(w) = z_1^{|w|_1} \ldots z_k^{|w|_k}$, where $|w|_i$ is the number of occurrences of the letter a_i in w. In other words, the mapping c allows us to consider words as in the abelian world, where letters commute. Let $\vec{z} = (z_1, \ldots, z_k)$. If $\mathcal{X}$ is a set of words, its *formal power series* $X(\vec{z})$ is defined as the formal sum of the monomials associated with its words: $X(\vec{z}) = \sum_{w \in \mathcal{X}} c(w)$. As we shall see, this power series is a tool of choice to study the probabilistic properties of the set $\mathcal{X}$.

First, the *symbolic method* [5] can be used to build $X(\vec{z})$, directly from a nonambiguous regular description of $\mathcal{X}$: if $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are three sets of words whose respective series are $X(\vec{z})$, $Y(\vec{z})$ and $Z(\vec{z})$, then

- if $\mathcal{X}$ is the disjoint union of $\mathcal{Y}$ and $\mathcal{Z}$, then $X(\vec{z}) = Y(\vec{z}) + Z(\vec{z})$;
- if $\mathcal{X}$ is the nonambiguous concatenation of $\mathcal{Y}$ and $\mathcal{Z}$, then $X(\vec{z}) = Y(\vec{z})Z(\vec{z})$;
- if $\mathcal{X}$ is the nonambiguous Kleene star of $\mathcal{Y}$, then $X(\vec{z}) = \frac{1}{1-Y(\vec{z})}$.

Second, for a given probability $\vec{\pi} = (\pi_1, \ldots, \pi_k)$ on A, one can build the formal power series in a single variable $\overline{X}(z)$, by substituting $\pi_i\, z$ to each z_i. After the substitution, the contribution of each word of length n to the coefficient of z^n in $\overline{X}(z)$, in the memoryless model, is exactly its probability. By marking a certain set of patterns with a copy of the alphabet, one can effectively multiply the contribution of a word by its number of patterns, and hence compute the expected number of such patterns using this technique (another approach is to control the unambiguity of the description [13]). Once $\overline{X}(z)$ is known, analytic combinatorics can be used to estimate the quantities under study.

Let us illustrate this technique on a toy example. Assume that we want to compute the expected number of occurrences of the pattern aba in a random word of length n under the memoryless distribution on the alphabet $\{a, b\}$, with[1] $\pi_a = \frac{1}{3}$ and $\pi_b = \frac{2}{3}$. Observe that the word $w = bbababaaab$ contains two (overlapping) occurrences of the pattern. The *marking technique* consists in distinguishing these two occurrences by using another alphabet, say $\{\overline{a}, \overline{b}\}$ for the letters of the pattern. The associated regular language is $\mathcal{L} = (a+b)^*\overline{aba}(a+b)^*$. The two words $w = bb\overline{bab}abaaab$ and $w = bbab\overline{aba}aaab$ correspond to w, which therefore contributes twice, as the pattern occurs twice. Using the symbolic method directly yields that the generating series of $\mathcal{L}$ is

$$L(\vec{z}) = \frac{1}{1-(z_a+z_b)} \cdot z_a z_b z_a \cdot \frac{1}{1-(z_a+z_b)} = \frac{z_a^2 z_b}{(1-z_a-z_b)^2}.$$

[1] For readibility, we use π_a, π_b, z_a and z_b instead of π_1, π_2, z_1 and z_2.

Then, we compute $\overline{L}(z)$ by performing the substitutions $z_a \mapsto \pi_a z$ and $z_b \mapsto \pi_b z$:

$$\overline{L}(z) = \frac{\pi_a^2 \pi_b z^3}{(1 - \pi_a z - \pi_b z)^2} = \frac{\pi_a^2 \pi_b z^3}{(1-z)^2} = \frac{2z^3}{27\,(1-z)^2}.$$

The coefficient of z^n in $\overline{L}(z)$ is the expected number of occurrences of the pattern in a random word of length n. The expression above is amenable to the analytic technique presented below (see Section 3.3), yielding the (natural) estimate of $\frac{2n}{27}$ occurrences on average.

3.2 Generating series for the expected number of patterns

We now use this general framework to compute the expected number of maximal α-gapped patterns. To simplify the notations, for any positive integer i and any vector $\vec{z}$, let $N_i(\vec{z}) = z_1^i + \ldots + z_k^i$. In particular, $N_1(\vec{z}) = z_1 + \ldots + z_k$.

A gapped pattern is equivalent to a triple of words (u, v, u'), with a condition $u' = u$ (for gapped repeats) or $u' = \overline{u}$ (for gapped palindromes), and a length condition $1 \leq |v| \leq \beta|u|$, which we rewrite into the equivalent $|u| \geq |v|/\beta$ and $|v| \geq 1$. Because we are ultimately interested in *maximal* patterns, we need to keep track of the first and last letters of v; this, in turn, forces us to distinguish between the subcases $|v| = 1$ and $|v| \geq 2$.

• In the simpler case $|v| = 1$, a pattern is just given by a single letter $a \in A$, and an arbitrary word u of length at least $\lceil 1/\beta \rceil$. The generating series for words of length at least ℓ is $N_1(\vec{z})^\ell/(1 - N_1(\vec{z}))$. In our patterns, the letters of u are to be counted twice, once in u and once in u'. This is taken into account by just changing $N_1(\vec{z})$ into $N_2(\vec{z})$ into the formula. Hence, the generating series for α-gapped patterns with $v = a_i$ is $\frac{z_i N_2(\vec{z})^{\lceil 1/\beta \rceil}}{1 - N_2(\vec{z})}$.

We now want to add a prefix and a suffix (both possibly empty) to the patterns. To avoid ambiguity in the description, we duplicate the alphabet and consider that patterns are written using this newly introduced copy. We are therefore considering words with one *marked* pattern, clearly identified. We also want the marked patterns to be maximal; this adds a condition on the prefix (resp. suffix) when it is not empty. This condition is slightly different for gapped repeats and gapped palindromes; we deal with gapped repeats first. Then the condition is that both prefix and suffix can be empty, but if they are not, the last letter of the prefix and the first letter of the suffix must be different from a_i. The generating series for both the possible prefixes and suffixes are the same, and equal to $(1 - z_i)/(1 - N_1(\vec{z}))$. Summing over all possible i, the generating series for all words with a marked maximal α-gapped pattern having a gap of length exactly 1 is therefore

$$U_\alpha(\vec{z}) = \frac{(N_1(\vec{z}) - 2N_2(\vec{z}) + N_3(\vec{z}))N_2(\vec{z})^{\lceil 1/\beta \rceil}}{(1 - N_1(\vec{z}))^2(1 - N_2(\vec{z}))}.$$

For gapped palindromes, there is a condition on the prefix and suffix when they are both nonempty: the last letter of the prefix must be different from the first letter of the suffix. This leads to multiplying the generating series for all patterns by the generating series for this set of pairs of words, which is $\frac{1 - N_2(\vec{z})}{(1 - N_1(\vec{z}))^2}$. We thus get as the generating series for all words with a marked maximal α-gapped palindrome having a gap of length exactly 1,

$$\overline{U}_\alpha(\vec{z}) = \frac{(1 - N_2(\vec{z}))\, N_1(\vec{z})\, N_2(\vec{z})^{\lfloor 1/\beta \rfloor}}{(1 - N_1(\vec{z}))^2\, (1 - N_2(\vec{z}))}.$$

• We now turn to the case $|v| \geq 2$. For any two letters a_i and a_j, we consider the possible gapped patterns (see Figure 1) such that v starts with a_i and ends with a_j (for maximal

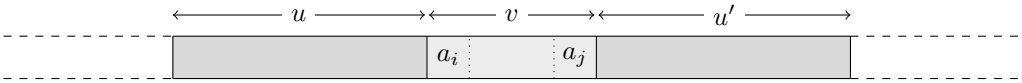

Figure 1 A gapped pattern uvu' with the first and last letters in v distinguished.

gapped palindromes, an additional condition is $i \neq j$). Let $\ell + 2$ be the length of such a word v; the α-gapped condition is thus $|u| \geq (\ell + 2)/\beta$. Writing $\beta = p/q$ with positive integers p and q, and writing the Euclidean division of ℓ by p as $\ell = tp + m$, the condition becomes $|u| \geq tq + (m + 2)/\beta$.

Thus, in the pattern uvu', u is obtained by concatenation of t arbitrary words of length q, plus one arbitrary word of length $\lceil (m + 2)/\beta \rceil$, plus an arbitrary (possibly empty) word; and v starts with a_i, concatenated with t arbitrary words of length p, plus one arbitrary word of length m, and ends with a_j. In the pattern composition, the composition of u has to be counted twice since u' also contributes and has the same composition. Summing over all possible values of t and m, we get the generating series for all α-gapped patterns such that v starts with a_i and ends with a_j:

$$G_{\alpha,i,j}(\vec{z}) = \frac{z_i z_j Q_\alpha(\vec{z})}{(1 - N_2(\vec{z}))(1 - N_1(\vec{z})^p N_2(\vec{z})^q)}, \text{ with } Q_\alpha(\vec{z}) = \sum_{m=0}^{p-1} N_1(\vec{z})^m N_2(\vec{z})^{\lceil (m+2)/\beta \rceil}.$$

Writing the generating functions for all words with a marked maximal gapped pattern again corresponds to adding a prefix and suffix, but leads to different generating functions for repeats and palindromes because the conditions on the suffix and prefix are slightly different.

For gapped repeats, both the prefix and the suffix can be empty, or an arbitrary word that does not end with a_j (for the prefix), or that does not start with a_i (for the suffix). This is done by multiplying the generating series $G_{\alpha,i,j}(\vec{z})$ by $(1 - z_i)(1 - z_j)/(1 - N_1(\vec{z}))^2$. Taking the sum over all possible j yields that the generating series of all words with a marked maximal α-gapped repeat and $|v| \geq 2$ is

$$V_\alpha(\vec{z}) = \frac{(N_1(\vec{z}) - N_2(\vec{z}))^2 \, Q_\alpha(\vec{z})}{(1 - N_1(\vec{z}))^2 (1 - N_2(\vec{z}))(1 - N_1(\vec{z})^p N_2(\vec{z})^q)}.$$

For gapped palindromes, maximality induces two conditions. First the last letter of v must be different from its first letter; this is taken into account by summing $G_{\alpha,i,j}(\vec{z})$ over all possible $i \neq j$. Second, the prefix and suffix must also satisfy the same conditions as for the case $|v| = 1$, which leads to multiply by $\frac{1 - N_2(\vec{z})}{(1 - N_1(\vec{z}))^2}$ as before. Hence, the generating series of all words with a marked maximal α-gapped palindrome and $|v| \geq 2$ is

$$\overline{V}_\alpha(\vec{z}) = \frac{(N_1(\vec{z})^2 - N_2(\vec{z})) \, Q_\alpha(\vec{z})}{(1 - N_1(\vec{z}))^2 (1 - N_1(\vec{z})^p N_2(\vec{z})^q)}.$$

We can now proceed with the substitution $z_i \to \pi_i z$, which changes $N_1(\vec{z})$ into z, $N_2(\vec{z})$ into $\lambda_2 z^2$ and $N_3(\vec{z})$ into $\lambda_3 z^3$, with $\lambda_j = \sum_i \pi_i^j$.

Let $\chi(w)$ (resp. $\xi(w)$) denote the number of maximal α-gapped repeats (resp. palindromes) in a word w. Let $R(z) = \sum_w \chi(w) \mathbb{P}(w) z^{|w|}$ and $P(z) = \sum_w \xi(w) \mathbb{P}(w) z^{|w|}$ be the generating series of the expectations of χ and ξ, that is, the coefficients of z^n of $R(z)$ and $P(z)$ are $\mathbb{E}_n[\chi]$ and $\mathbb{E}_n[\xi]$, respectively. These series $R(z)$ and $P(z)$ are obtained by the previous substitutions $z_i \to \pi_i z$ from the series $U_\alpha(\vec{z}) + V_\alpha(\vec{z})$ and $\overline{U}_\alpha(\vec{z}) + \overline{V}_\alpha(\vec{z})$, respectively. From the computations above, we obtain the following statement.

▶ **Theorem 2.** *For $\beta = \alpha - 1 = \frac{p}{q}$, the series $R(z)$ and $P(z)$ for the memoryless model of probability $\vec{\pi}$ are given by*

$$R(z) = \frac{(z - 2\lambda_2 z^2 + \lambda_3 z^3)\lambda_2^{\lceil 1/\beta \rceil} z^{2\lceil 1/\beta \rceil}}{(1-z)^2(1-\lambda_2 z^2)} + \frac{(z - \lambda_2 z^2)^2\, \overline{Q_\alpha}(z)}{(1-z)^2(1-\lambda_2 z^2)(1-\lambda_2^q\, z^{p+2q})},$$

$$P(z) = \frac{\lambda_2^{\lceil 1/\beta \rceil} z^{1+2\lceil 1/\beta \rceil}}{(1-z)^2} + \frac{(z^2 - \lambda_2 z^2)\, \overline{Q_\alpha}(z)}{(1-z)^2(1-\lambda_2^q\, z^{p+2q})},$$

with

$$\overline{Q_\alpha}(z) = \sum_{j=0}^{p-1} \lambda_2^{\lceil (j+2)/\beta \rceil} z^{j+2\lceil (j+2)/\beta \rceil},\ \lambda_2 = \sum_{i=1}^{k} \pi_i^2,\ \text{and } \lambda_3 = \sum_{i=1}^{k} \pi_i^3.$$

3.3 From generating series to asymptotics

Analytic combinatorics links asymptotic behavior of counting sequences to singularities of the corresponding generating functions, viewed as analytic functions of a complex variable. For rational generating series of one variable, as in Theorem 2, the situation is quite simple, and we use this simplified version of the Transfer Theorem [5] for rational functions:

▶ **Theorem 3** (Simplified Transfer Theorem [5])**.** *Assume $A(z) = F(z)(1-z)^{-\ell}$, where ℓ is a positive integer, $F(z)$ is a rational function with no pole in the closed disc of radius 1 and $F(1) \neq 0$. Then the n-th coefficient of $A(z)$ is asymptotically equivalent to $\frac{F(1)}{(\ell-1)!}\, n^{\ell-1}$.*

The series $R(z)$ and $P(z)$ of Theorem 2 both have a dominant pole of order 2 at $z = 1$. Applying Theorem 3 yields the following statement. Note that, though the generating series $R(z)$ and $P(z)$ are different, they lead to the same asymptotics for the coefficients; the difference is in lower order terms.

▶ **Theorem 4.** *Under the memoryless distribution of probability $\vec{\pi}$, and for any rational $\alpha = 1 + p/q$, the expected number of maximal α-gapped repeats (respectively, palindromes) in a random word of length n is asymptotically equivalent to $r_\alpha n$ (respectively, $p_\alpha n$) defined by*

$$r_\alpha = \frac{(1 - 2\lambda_2 + \lambda_3)\lambda_2^{\lceil q/p \rceil}}{1-\lambda_2} + \frac{(1-\lambda_2)}{1-\lambda_2^q} \sum_{j=2}^{p+1} \lambda_2^{\lceil jq/p \rceil} \text{ and } p_\alpha = \lambda_2^{\lceil q/p \rceil} + \frac{(1-\lambda_2)}{1-\lambda_2^q} \sum_{j=2}^{p+1} \lambda_2^{\lceil jq/p \rceil}.$$

In particular, when α is a positive integer, these reduce to

$$r_\alpha = (\alpha - 1)\lambda_2 + \frac{\lambda_2(\lambda_3 - \lambda_2^2)}{1-\lambda_2} \text{ and } p_\alpha = (\alpha-1)\lambda_2 + \lambda_2^2.$$

For the uniform distribution, we have $\lambda_2 = 1/k$ and $\lambda_3 = 1/k^2$, yielding the following result.

▶ **Corollary 5.** *For the uniform distribution on an alphabet of size $k \geq 2$, we have*

$$r_\alpha = \frac{k-1}{k}\left(k^{-\lceil q/p \rceil} + \frac{\sum_{j=2}^{p+1} k^{-\lceil jq/p \rceil}}{1-k^{-q}}\right) \text{ and } p_\alpha = r_\alpha + k^{-1-\lceil q/p \rceil}.$$

In particular, if α is a positive integer, then $r_\alpha = \frac{\alpha-1}{k}$ and $p_\alpha = \frac{\alpha-1}{k} + \frac{1}{k^2}$.

▶ Remark. As a function of $\alpha = 1 + \frac{p}{q}$, the value of r_α is not very regular. It is increasing, as expected, but there are some large variations when reaching a value with a small denominator (typically integers or half-integers). This also gives hints on the difficulty of giving a formula if α is not rational. Some examples are given in the table below, for $k = 4$.

α	5/4	3/2	7/4	2	9/4	5/2	11/4	3	13/4	7/2	15/4	4	17/4
r_α	0.002	0.05	0.061	0.25	0.252	0.3	0.311	0.5	0.502	0.55	0.561	0.75	0.752

4 Longest pattern

In this section we focus on the typical and expected length of the longest α-gapped patterns (repeat or palindromes) in a random word. Contrarily to the previous section, our analysis relies on discrete probabilities rather than on generating series.

Let L_n denote the random variable associated with the length of the longest α-gapped patterns in a random word of length n. We first focus on the uniform distribution, in order to introduce the main techniques of this section. For memoryless distributions, the computations are more involved, but the general idea remains the same.

If X_n is a random variable, we say that it is *concentrated around its mean* if there exists a sequence $(\nu_n)_{n\geq 1}$ that tends to 0 such that $\mathbb{P}(|X_n - \mathbb{E}[X_n]| > \nu_n \mathbb{E}[X_n]) \xrightarrow[n\to\infty]{} 0$.

In this whole section, whenever we say that some property holds *with asymptotic probability* 1, the property implicitly depends on some integer n, which denotes the length of the random words considered; and we mean that, as n goes to infinity, the probability tends to 1. The details in the text typically make it possible to derive a more explicit bound on the speed of convergence.

4.1 Uniform distribution

We establish the following theorem. Its proof is obtained by computing tight lower and upper bounds for the typical value of L_n, for the uniform distribution.

▶ **Theorem 6.** *For the uniform distribution on words of length n, on an alphabet of size k, the expected length of the longest α-gapped repeat (or palindrome) is asymptotically equivalent to $(\alpha+1)\log_k n$. Moreover, the random variable L_n is concentrated around its mean.*

Observe that a longest α-gapped repeat is necessarily maximal. Moreover, the proof is exactly the same for palindromes, so we focus on repeats only.

To establish the lower bound, we prove that with asymptotic probability 1 there is an α-gapped repeat of length t_0 in a random word of length n, where t_0 is a well chosen value, which is asymptotically equivalent to $(\alpha+1)\log_k n$. This property is proved to hold by just looking for α-gapped repeats lying at very specific positions: the word is split into roughly n/t_0 factors of length t_0, and we only compute the probability that at least one of these factors is an α-gapped repeat of a particular $|v|/|u|$ ratio. This fairly rough estimation is enough to establish a lower bound that is asymptotically tight.

For any $\ell \geq 1$, let $\mathcal{M}_\beta(\ell)$ denote the set of words of the form uvu, with $u \in A^\ell$ and $v \in A^{\lfloor \beta\ell \rfloor}$. The set $\mathcal{M}_\beta(\ell)$ therefore contains all the α-gapped repeats where u has size ℓ and v is of maximal length. Let $\ell_0 = \lfloor \log_k n - 2\log_k \log_k n \rfloor$. Every word of $\mathcal{M}_\beta(\ell_0)$ has length $t_0 = 2\ell_0 + \lfloor \beta\ell_0 \rfloor$, and t_0 is asymptotically equivalent to $(\alpha+1)\log_k n$, as required.

The probability for an element of $\mathcal{M}_\beta(\ell_0)$ to be a factor of a random word of length n is exactly the probability that, for some $i \in [n]$, the factor of length t_0 starting at position i belongs to $\mathcal{M}_\beta(\ell_0)$. Thus, it is at least the probability that the factor of length t_0 starting at position $1 + jt_0$ is in $\mathcal{M}_\beta(\ell_0)$ for some $j \geq 1$ such that $(j+1)t_0 \leq n$. For such a given j, the probability that the factor starting at position $1 + jt_0$ is in $\mathcal{M}_\beta(\ell_0)$ is $k^{-\ell_0}$, since $|\mathcal{M}_\beta(\ell_0)| = k^{\ell_0 + \lfloor \beta\ell_0 \rfloor}$ and each possible factor has probability $k^{-\ell_0 - 2\lfloor \beta\ell_0 \rfloor}$. Since the integer intervals $[1 + jt_0, (j+1)t_0]$ do not overlap, the factors they define are independent, and the probability that none of them is in $\mathcal{M}_\beta(\ell_0)$ is $(1 - k^{-\ell_0})^{\lfloor n/t_0 \rfloor}$. Therefore, with probability at least $1 - (1 - k^{-\ell_0})^{\lfloor n/t_0 \rfloor}$, a random word of length n contains an α-gapped repeat of length t_0.

Straightforward computations yield that $(1-k^{-\ell_0})^{\lfloor n/t_0 \rfloor} \leq \exp(-\log_k n)$, which tends to 0 as $n \to \infty$. Thus, with asymptotic probability 1, a random uniform word of length n contains an α-gapped repeat of length t_0, which is asymptotically equivalent to $(\alpha+1)\log_k n$.

We now proceed with the upper bound. Let $\mathcal{R}_\beta(t)$ denote the set of all words uvu such that $|uvu| = t$ and $|v| \leq \beta|u|$. The set $\mathcal{R}_\beta(t)$ contains all the possible α-gapped repeats of length t. Observe that, by summing over all the possible lengths ℓ for u, we have

$$|\mathcal{R}_\beta(t)| = \sum_{\ell=\lceil t/(2+\beta)\rceil}^{\lfloor (t-1)/2 \rfloor} k^{t-\ell} \leq k^{t-\lceil t/(2+\beta)\rceil} \sum_{j=0}^{\infty} k^{-j} \leq 2k^{(\beta+1)t/(\beta+2)}.$$

Let $t_1 = \lceil (\beta+2)\log_k n + 2(\beta+2)\log_k \log_k n \rceil + 1$. The probability that a random word contains a factor in $\mathcal{R}_\beta(t_1)$ at a given valid position is $|\mathcal{R}_\beta(t_1)|k^{-t_1} \leq 2k^{-t_1/(\beta+2)}$. Since the number of valid positions is no more than n, by the union bound, the probability that a uniform random word of length n contains an element of $\mathcal{R}_\beta(t_1)$ (as a factor in any position) is at most $2nk^{-t_1/(\beta+2)}$. These computations also hold if one substitutes $t_1 + i$ for t_1. This yields that the probability that a uniform random word of length n contains an element of $\mathcal{R}_\beta(t_1+i)$ is bounded from above by $\frac{2k^{-i/(\beta+2)}}{\log_k^2 n}$.

Using the union bound again, we sum these bounds for $i \geq 0$, and obtain that, with asymptotic probability 1, a uniform random word of length n contains no α-gapped repeat of length greater than or equal to t_1, which is asymptotically equivalent to $(\alpha+1)\log_k(n)$.

A bit more is required to estimate the expectation of L_n, but this can be easily done from the computations above: they yield that the contribution to the expectation of the values that are not between t_0 and t_1 is negligible, and $t_0 \sim t_1 \sim (\alpha+1)\log_k n$. The concentration around the mean can be proved by taking any sequence ν_n that tends to 0 and such that $\frac{\nu_n \log_k n}{\log_k \log_k n}$ tends to infinity.

4.2 Memoryless sources

In this section, we associate to each letter $a_i \in A = \{a_1, \ldots, a_k\}$ a probability $\pi_i = p(a_i)$ as described in Section 2. We assume all these probabilities to be positive (otherwise, reduce the alphabet size accordingly).

From the probability $\vec{\pi}$, we build another probability $\vec{\tau}$ proportional to the square of π: for every $i \in [k]$, $\tau_i = \pi_i^2/\lambda_2$, where $\lambda_2 = \sum_{i\in[k]} \pi_i^2$ is the *coincidence probability* of $\vec{\pi}$ (as in Section 3). The result for memoryless sources, which generalizes Theorem 6, is the following.

▶ **Theorem 7.** *For the memoryless source of probability $\vec{\pi}$, the expected length of the longest α-gapped repeat (or palindrome) in a random word of length n is asymptotically $\mathbb{E}[L_n] \sim (\alpha+1)\log_{1/\lambda_2} n$, where $\lambda_2 = \sum_{i\in[k]} \pi_i^2$. Moreover, L_n is concentrated around its mean.*

Though it follows the same main ideas as in the proof of Theorem 6, the proof of Theorem 7 is more technical. Due to lack of space, we only focus on the main steps in this extended abstract. We will focus on the most probable words, and the most probable longest α-gapped repeat. For this purpose, for a probability vector $\vec{s}$ on A and $\delta \geq 0$, we consider the set $\mathcal{W}_n(\vec{s}, \delta)$ of words whose letters roughly follow the distribution of $\vec{s}$, defined by

$$\mathcal{W}_n(\vec{s}, \delta) = \left\{ u \in A^n : \ \left| |u|_a - s(a)\, n \right| \leq \delta, \ \forall a \in A \right\}.$$

To establish the lower bound, we define the set $\mathcal{M}_\beta(\vec{\pi}, \ell)$ of α-gapped repeats uvu where the letters are distributed following $\vec{\pi}$ in v and following $\vec{\tau}$ in u. More formally:

$$\mathcal{M}_\beta(\vec{\pi}, \ell) = \left\{ uvu \in A^* : \ u \in \mathcal{W}_\ell(\vec{\tau}, \sqrt{\log n}) \text{ and } v \in \mathcal{W}_{\lfloor \beta\ell \rfloor}(\vec{\pi}, \sqrt{\log n}) \right\}.$$

The set $\mathcal{M}_\beta(\vec{\pi},\ell)$ will play the same role as the set $\mathcal{M}_\beta(\ell)$ of the previous section. They do not coincide if the distribution is uniform, but they still have the same order of size.

We now proceed with the lower bound. We define ℓ_0 and t_0 by

$$\ell_0 = \left\lfloor \frac{\log n}{\log(1/\lambda_2)} - \frac{(\log n)^{2/3}}{\log(1/\lambda_2)} \right\rfloor \text{ and } t_0 = 2\ell_0 + \lfloor \beta\ell_0 \rfloor,$$

then prove that long random words have a factor in $\mathcal{M}_\beta(\vec{\pi},\ell_0)$ with high probability.

For this purpose, we need to estimate the probability that a factor of length t_0 at a given position is in $\mathcal{M}_\beta(\vec{\pi},\ell_0)$. The computations are done as follows. Let $\vec{n} = (n_1,\dots,n_k)$ with $n_1+\ldots+n_k = \ell_0$ and let $\vec{m} = (m_1,\dots,m_k)$ with $m_1+\ldots+m_k = \lfloor\beta\ell_0\rfloor$. We are interested in the set of words $\mathcal{E}(\vec{n},\vec{m})$, with fixed compositions for u and v, defined by

$$\mathcal{E}(\vec{n},\vec{m}) = \{uvu : \ |u|_{a_i} = n_i \text{ and } |v|_{a_i} = m_i, \ \forall i \in [k]\}.$$

Observe that $\mathcal{M}_\beta(\ell_0)$ can be written as a union of $\mathcal{E}(\vec{n},\vec{m})$ for properly chosen ranges for $\vec{n}$ and $\vec{m}$. The probability that the factor of length t_0 at a given valid position lies in $\mathcal{E}(\vec{n},\vec{m})$ is

$$\mathbb{P}_{t_0}(\mathcal{E}(\vec{n},\vec{m})) = \binom{\ell_0}{n_1,\dots,n_k} \prod_{i\in[k]} \pi_i^{2n_i} \binom{\lfloor\beta\ell_0\rfloor}{m_1,\dots,m_k} \prod_{i\in[k]} \pi_i^{m_i}.$$

By estimating this quantity and summing for all $\vec{n}$ and $\vec{m}$ such that $\mathcal{E}(\vec{n},\vec{m}) \subseteq \mathcal{M}_\beta(\ell_0)$, we obtain that the probability of the factor of length t_0 at a given valid position not being in $\mathcal{M}_\beta(\vec{\pi},\ell_0)$ is $O(\frac{1}{\log^2 n})$. At this point, the proof continues exactly as in Section 4.1.

We now turn to the upper bound. As in Section 4.1 let $\mathcal{R}_\beta(t)$ be the set of all uvu such that $|uvu| = t$ and $1 \le |v| \le \beta|u|$. We want to compute an upper bound for the probability that the factor of length t at a given valid position lies in $\mathcal{R}_\beta(t)$.

We need to partition the set $\mathcal{R}_\beta(t)$ for our computations. Let $\vec{\ell} = (\ell_1,\dots,\ell_k)$ be a vector of non-negative integers such that $N_1(\vec{\ell}) = \ell_1+\cdots+\ell_k = \ell$. Let $\mathcal{R}_\beta(\vec{\ell},t)$ be the set of all words uvu such that $|uvu| = t$ and $|u|_{a_i} = \ell_i$ for every $i \in [k]$. Observe that $\mathcal{R}_\beta(t)$ can be written as the following disjoint union:

$$\mathcal{R}_\beta(t) = \bigcup_{\ell=\lceil t/(\beta+2)\rceil}^{\lfloor (t-1)/2\rfloor} \ \bigcup_{N_1(\vec{\ell})=\ell} \mathcal{R}(\vec{\ell},t).$$

Moreover, $\mathbb{P}_t(\mathcal{R}(\vec{\ell},t)) = \binom{\ell}{\ell_1,\dots,\ell_k} \prod_{i\in[k]} \pi_i^{2\ell_i}$. But $\sum_{N_1(\vec{\ell})=\ell} \binom{\ell}{\ell_1,\dots,\ell_k} \prod_{i\in[k]} \tau_i^{\ell_i} = 1$, as it is the sum of the probabilities of all the words of length ℓ for the memoryless distribution of probability vector $\vec{\tau}$. Hence, $\mathbb{P}_\ell(\cup_{N_1(\vec{\ell})=\ell} \mathcal{R}(\vec{\ell},t)) = \lambda_2^\ell$. Therefore, $\mathbb{P}_t(\mathcal{R}_\beta(t)) \le t\,\lambda_2^{t/(\beta+2)}$. In particular, for $t_1 = \lceil (\beta+2)\log_{1/\lambda_2} n + 3(\beta+2)\log_{1/\lambda_2}\log n\rceil$, we have

$$\mathbb{P}(\mathcal{R}_\beta(t_1+i)) \le \frac{2\,\lambda_2^{i/(\beta+2)}}{\log^2 n},$$

and the proof continues as in the uniform case.

▶ Remark. As a byproduct of the proof, we obtain the following interesting result on the probabilistic nature of the longest α-gapped repeat. Though a sufficiently large random word in the memoryless model contains roughly a proportion π_i of each letter a_i, the letters are distributed differently in the arms (the u's of uvu) of a typical longest α-gapped repeat: the proportion of each letter is roughly τ_i instead of π_i. This phenomenon is completely hidden in the uniform case, where $\tau_i = \pi_i = 1/k$ for every $i \in [k]$.

5 A remark on the number of distinct factors

In [15], Rubinchik and Shur estimated the expected number of distinct palindromes in a random word: these factors are counted only once, even if they have multiple occurrences. They established that for the uniform distribution, a random word contains $\Theta(\sqrt{n})$ distinct palindromes, and several refinements of this result.

In this short section we explain how their proof can be extended to estimate the expected number of distinct α-gapped repeats and palindromes, for the uniform distribution. There is no new idea, one just has to take care of the possibilities for the additional v part in the pattern. The result, however, is interesting on its own. It is stated as followed.

▶ **Theorem 8.** *For the uniform distribution over words of length n, the expected number of distinct α-gapped repeats (or palindromes) is in $\Theta(n^{\alpha/(\alpha+1)})$.*

We only consider repeats in our proof sketch; gapped palindromes are treated the same way. The lower bound is obtained using Guibas and Odlyzko's result on pattern avoidance [9]: the number of words of length n that avoid a given pattern w of length $m > 3$ is equal to $C_w\theta_w^n + O(1.7^n)$. In [15], the authors prove that the constants are maximal when $w = a^m$:

$$\theta_w \leq \theta_{a^m} = k\left(1 - \frac{k-1}{k^{m+1}} + O\left(\frac{m}{k^{2m+2}}\right)\right) \text{ and } C_w \leq C_{a^m} = 1 + O\left(\frac{m}{k^m}\right). \tag{1}$$

Let $\S_\beta(n)$ be the set of all uvu such that $|u| = \ell = \lfloor \frac{1}{\beta+2}\log_k n \rfloor$ and $|v| = \lfloor \beta\ell \rfloor$. Let $m = 2\ell + \lfloor \beta\ell \rfloor$. As a direct application of Equation (1), for a given $w \in \S_\beta(n)$, the probability that a random word of length n avoids w satisfies

$$\mathbb{P}_n(\text{avoiding } w) \leq \left(1 - \frac{k-1}{k^{m+1}} + O\left(\frac{m}{k^{2m+2}}\right)\right)^n \left(1 + O\left(\frac{m}{k^m}\right)\right),$$

which is at most C for some positive constant $C < 1$ and n sufficiently large (for our choice of ℓ, and thus of m). Hence, the probability for w to be a factor in a random word of length n is at least $1 - C$, and by linearity of the expectation, the expected number of distinct α-gapped repeats is greater than or equal to $(1 - C)|\S_\beta(n)| = (1 - C)k^{m-\ell} = \Omega(n^{\alpha/(\alpha+1)})$.

For the upper bound, let $\mathcal{R}_\beta(t)$ denote all the uvu such that $|uvu| = t$ and $|v| \leq \beta|u|$. Let also $t_0 = \lceil \log_k n \rceil$. We count differently the α-gapped repeats, depending on whether they are shorter or longer than t_0.

We count all the α-gapped repeats of length at most t_0 as contributing to the upper bound. By summing over all possible values for the length ℓ of the arms, we have

$$|\mathcal{R}_\beta(t)| = \sum_{\ell=\lceil t/(\beta+1)\rceil}^{\lfloor (t-1)/2 \rfloor} k^{t-\ell} \leq k^t \sum_{\ell=\lceil t/(\beta+1)\rceil}^{\infty} k^{t-\ell} \leq 2k^{(\beta+1)t/(\beta+2)}.$$

Hence,

$$\sum_{t=1}^{t_0} |\mathcal{R}_\beta(t)| \leq \sum_{t=1}^{t_0} 2k^{(\beta+1)t/(\beta+2)} \leq 4k^{(\beta+1)t_0/(\beta+2)} \leq 4n^{\alpha/(\alpha+1)}.$$

To obtain an upper bound for the expected number of patterns of length greater than t_0, we observe as in [15] that the probability for a given α-gapped repeat uvu to be a factor is at most the expected number of occurrences of uvu. As we shall see, this rough upper bound is enough to conclude. Let $\mathcal{R}_\beta(t, \ell)$ be the set of uvu such that $|uvu| = t$, $|u| = \ell$

and $|v| \leq \beta\ell$. The probability that there is pattern of $\mathcal{R}_\beta(t,\ell)$ at a given valid position in a random word is $k^{-\ell}$. Hence, the expected number of occurrences of such patterns is at most $nk^{-\ell}$, for given t and ℓ. By summing over all $t > t_0$ and all valid ℓ for each t we obtain the following upper bound:

$$\sum_{t=t_0}^{n} \sum_{\ell=\lceil t/(\beta+2)\rceil}^{\lfloor t/2\rfloor} n\,k^{-\ell} \leq \sum_{t=t_0}^{n} 2k^{-t/(\beta+2)} \leq 4k^{-t_0/(\beta+2)} \leq 4n^{\alpha/(\alpha+1)}.$$

Combining both results, for short and long α-gapped repeats, we get that the expected number of distinct such factors is bounded from above by $8n^{\alpha/(\alpha+1)}$, concluding the proof.

▶ Remark. Theorem 8 also holds for *maximal* patterns. The proof simply needs to be adapted for the lower bound, and there are sufficiently many of them to obtain the same result.

6 Conclusions

In this article we establish results about some statistics of random words related to the notion of α-gapped patterns, for both the uniform and memoryless distributions. We propose different techniques, generating series and discrete probabilities, to provide some tools for further analysis of statistics of interest. Amongst them, if would be natural to consider gapped patterns as a whole, *i.e.* if $uvu = u'v'u'$ then it is considered as one pattern instead of two different ones.

The biased distribution on letters in the arms of a typical α-gapped pattern, in the memoryless model, is something worth noticing (see Section 4.2). It may provide some leverage for speeding up algorithms, though the difference might be too thin to be exploited.

Finally, generalizing Theorem 8 to memoryless proves quite difficult. This is ongoing work, and the techniques involved are more advanced than what is presented in this article.

References

1. Gerth Stølting Brodal, Rune B. Lyngsø, Christian N. S. Pedersen, and Jens Stoye. Finding maximal pairs with bounded gap. In *Combinatorial Pattern Matching, 10th Annual Symposium, CPM 99, Warwick University, UK, July 22-24, 1999, Proceedings*, pages 134–149, 1999.
2. Maxime Crochemore, Roman Kolpakov, and Gregory Kucherov. Optimal bounds for computing α-gapped repeats. In *Language and Automata Theory and Applications - 10th International Conference, LATA 2016, Prague, Czech Republic, March 14-18, 2016, Proceedings*, pages 245–255, 2016.
3. Maxime Crochemore and Wojciech Rytter. *Text Algorithms*. Oxford University Press, 1994.
4. Marius Dumitran and Florin Manea. Longest gapped repeats and palindromes. In *Mathematical Foundations of Computer Science 2015 - 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part I*, pages 205–217, 2015.
5. Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2008.
6. Pawel Gawrychowski, Tomohiro I, Shunsuke Inenaga, Dominik Köppl, and Florin Manea. Efficiently finding all maximal α-gapped repeats. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016, February 17-20, 2016, Orléans, France*, pages 39:1–39:14, 2016.
7. Pawel Gawrychowski and Florin Manea. Longest α-gapped repeat and palindrome. In *Fundamentals of Computation Theory - 20th International Symposium, FCT 2015, Gdańsk, Poland, August 17-19, 2015, Proceedings*, pages 27–40, 2015.

8 Amy Glen and Jamie Simpson. The total run length of a word. *Theoretical Computer Science*, 501:41–48, 2013.

9 Leonidas J. Guibas and Andrew M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A*, 30(2):183–208, 1981.

10 Roman Kolpakov and Gregory Kucherov. Searching for gapped palindromes. *Theoretical Compututer Science*, 410(51):5365–5373, 2009.

11 Roman Kolpakov, Mikhail Podolskiy, Mikhail Posypkin, and Nickolay Khrapov. Searching of gapped repeats and subrepetitions in a word. In *Combinatorial Pattern Matching - 25th Annual Symposium, CPM 2014, Moscow, Russia, June 16-18, 2014. Proceedings*, pages 212–221, 2014.

12 Roman M. Kolpakov and Gregory Kucherov. Finding repeats with fixed gap. In *Seventh International Symposium on String Processing and Information Retrieval, SPIRE 2000, A Coruña, Spain, September 27-29, 2000*, pages 162–168, 2000.

13 Cyril Nicaud. Estimating statistics on words using ambiguous descriptions. In *27th Annual Symposium on Combinatorial Pattern Matching, CPM 2016, June 27-29, 2016, Tel Aviv, Israel*, pages 9:1–9:12, 2016.

14 Simon J. Puglisi and Jamie Simpson. The expected number of runs in a word. *Australasian Journal of Combinatorics*, 42:45–54, 2008.

15 Mikhail Rubinchik and Arseny M. Shur. The number of distinct subpalindromes in random words. *Fundam. Inform.*, 145(3):371–384, 2016.

16 Arseny M. Shur. Growth properties of power-free languages. *Computer Science Review*, 6(5-6):187–208, 2012.

17 Yuka Tanimura, Yuta Fujishige, Tomohiro I, Shunsuke Inenaga, Hideo Bannai, and Masayuki Takeda. A faster algorithm for computing maximal α-gapped repeats in a string. In *String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings*, pages 124–136, 2015.