On the Expected Number of Distinct Gapped Palindromic Factors

Philippe Duchon[1] and Cyril Nicaud[2]

[1] Univ. Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France
CNRS, LaBRI, UMR 5800, F-33400 Talence, France
[2] Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE Paris, UPEM,
F-77454 Marne-la-Vallée, France

Abstract. An α-gapped palindromic factor of a word is a factor of the form $uv\overline{u}$, where $\overline{u}$ is the reversal of u and where $|uv| \leq \alpha|u|$ for some fixed $\alpha \geq 1$. We give an asymptotic estimate of the expected number of distinct palindromic factors in a random word for a memoryless source, where each letter is generated independently from the other, according to some fixed probability distribution on the alphabet.

1 Introduction

An α-gapped palindrome is a word of the form $uv\overline{u}$, where $\overline{u}$ is the reversal[3] of u and where $|uv| \leq \alpha|u|$, for some fixed $\alpha \geq 1$. Initially motivated by applications to bioinformatics, several articles in the literature focus on studying the α-gapped palindromic factors that occur in a given word [7, 5]. Different directions were taken in these studies, and it is now known that there are at most a linear number of distinct α-gapped palindromic factors in a word [1, 5], and that they can be computed in linear time [8, 13, 6].

In this paper, we are interested in the probabilistic properties related to this notion: if w is a random word of length n, what can be said about its α-gapped palindromic factors? To answer this kind of question, the probabilistic model must be specified. In the sequel, we will consider words generated using a *memoryless source*: each letter is chosen independently from each other, following a fixed probability distribution on the alphabet. Together with C. Pivoteau, we gave several results in [4]: the expected number of α-gapped palindromic factors and the expected length of the longest such factor. These were obtained using classical techniques from analytic combinatorics, together with elementary discrete probabilities. We also adapted, almost readily, a result by M. Rubinchik and A. Shur [11] on the expected number of *distinct* palindromic factors when the distribution is the *uniform distribution*, to *distinct* α-gapped palindromic factors. But this technique fails to work when the distribution is not uniform.

We aim at completing the works [11, 4] by studying the number of distinct α-gapped palindromic factors in a random word generated by a memoryless source. Beside the combinatorial and probabilistic motivations, note that knowledge on

[3] The reversal of $u = u_1 \cdots u_n$ is $\overline{u} = u_n \cdots u_1$.

the typical number of distinct factors can be useful in the design of data structures, as it can give hints on the likely memory size needed for a typical input. For instance, the number of vertices of the graph EERTREE introduced in [12] is the number of distinct palindromic factors, which is in $\Theta(\sqrt{n})$ in expectation for the uniform distribution [11].

The classical techniques we used in [4] are not well suited to handle distinct factors. This is why we propose in [3] a probabilistic process that focuses on the notion of distinctness, in order to develop useful methodologies. The process we studied is the following: generate N random words of length L, independently, and remove duplicates. What does the resulting random set S looks like? We gave a precise characterization of the typical composition of letters of a word of S. More than the result itself, the techniques we used, based on classical analysis of functions with several variables, can be used to try to tackle other questions involving distinctness and non-uniform models. It also hints that uniform distributions are really singular, hiding some complicated situations that appear for non-uniform distribution only.

In this article, we use techniques that are similar to those introduced in [3] in order to estimate the expected number of distinct α-gapped palindromic factors in a random word, generated by a memoryless source. Our problem is more difficult than [3] for two main reasons: factors of a random word are not independent and α-gapped palindromic factors have various lengths. As we will see, these technical difficulties can be overcome, and we give in the sequel two main results: an estimate of the expected number of distinct α-gapped palindromic factors and a description of the factors that are more likely to occur, in term of their lengths and of their composition of letters.

2 Definitions and notations

If k is a positive integer, let $[k] = \{1, \dots, k\}$. For $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$, let $\|\mathbf{x}\| = \sqrt{\sum_{i\in[k]} x_i^2}$ denote the Euclidean norm. A vector $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$ is a *probability vector* if every $x_i \in [0, 1]$ and $\sum_{i=1}^k x_i = 1$.

Words. Let $A = \{a_1, \dots, a_k\}$ be an alphabet with $k \geq 2$ letters. Throughout the article, the alphabet A, and therefore k, are fixed. We denote the empty word by ε. If $u = u_1 \cdots u_n$ is a word of length n on A, then its *reversal* is the word $\overline{u} = u_n \cdots u_1$. A *palindrome* is a word of the form $u\lambda\overline{u}$ where $\lambda \in A \cup \{\varepsilon\}$, that is, where λ is either a letter or empty. Let $\alpha \geq 1$ be a real number. An *α-gapped* palindrome is a word of the form $uv\overline{u}$ where $|uv| \leq \alpha|u|$.

For any word $w \in A^*$, the *composition vector* (or *Parikh vector*) of w is the vector $(|w|_1, \dots, |w|_k)$, where $|w|_i$ is the number of occurrences of a_i in w. If w is not empty, its *frequency vector* is the probability vector $(\frac{|w|_1}{|w|}, \dots, \frac{|w|_k}{|w|})$. We let $\mathcal{W}_m(\mathbf{x})$ denote the set of words of length m with frequency vector $\mathbf{x}$ (which is empty if $m\mathbf{x}$ does not have nonnegative integer coordinates).

Probabilities. Throughout the article, we assume some probability vector $\mathbf{p} = (p_1, \dots, p_k)$ to be fixed, with $p_i \neq 0$ for every $i \in [k]$, and we consider statistics

in the memoryless model where each letter a_i has probability p_i. We also assume that $\mathbf{p}$ is not the uniform distribution, *i.e.* there exists $i \in [k]$ such that $p_i \neq \frac{1}{k}$. Let $p_{\max} = \max_{i\in[k]} p_i < 1$ denote the maximal value of $\mathbf{p}$.

We will use the (natural-based) entropy function [9] on k positive variables, which is defined by $H(\mathbf{x}) = H(x_1, \ldots, x_k) = -\sum_{i=1}^{k} x_i \log x_i$. We also borrow the function Φ from [3], defined for any non-negative t by $\Phi(t) = \sum_{i=1}^{k} p_i^t$.

3 Expected number of distinct α-gapped factors

Recall that the $\tilde{\Theta}$ notation means "asymptotically of the same growth, up to some polylogarithmic multiplicative factors". More precisely, a positive sequence $(u_n)_{n\geq 0}$ is $\tilde{\Theta}(n^d)$ if there exists $\delta > 0$ such that $n^d(\log n)^{-\delta} \leq u_n \leq n^d(\log n)^{\delta}$, for n sufficiently large. It is in the same vein, though a bit more precise, as saying that for all $\varepsilon > 0$, u_n is $\mathcal{O}(n^{d+\varepsilon})$ and u_n is $\Omega(n^{d-\varepsilon})$. Our main result is the following.

Theorem 1. *Let $\alpha \geq 1$. The expected number of distinct α-gapped palindromic factors is $\tilde{\Theta}(n^{c^*})$, where $c^* \in (0, 1)$ is the unique positive solution of the equation*

$$\Phi(2c)\Phi(c)^{\alpha-1} = 1.$$

Observe that, even if we assume that $\mathbf{p}$ is not the uniform distribution in this paper, the result is compatible with the estimations of [11, 3]: we have $\Phi(t) = k^{1-t}$ for the uniform distribution, so the equation rewrites $k^{1-2c+(\alpha-1)(1-c)=1} = 1$, so that $c^* = \frac{\alpha}{\alpha+1}$. The value of c^* for non-uniform $\mathbf{p}$ are depicted in Fig. 1.

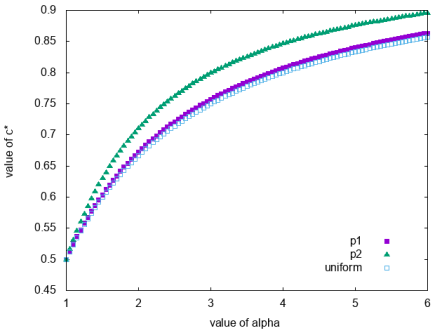

Fig. 1. The values of c^* as α ranges from 1 to 6, for three different probability vectors on an alphabet with three letters ($k = 3$): $\overline{\mathbf{x}}$ the uniform distribution plotted with empty squares, $p_1 = (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ plotted with squares, and $p_2 = (\frac{4}{5}, \frac{1}{10}, \frac{1}{10})$ plotted with triangles. For these three examples, the more the probability approaches the uniform distribution, the smaller the expected number of distinct factors.

The proof, presented in the rest of this section, consists in finding the frequency vector that contributes the most to the number of distinct α-gapped palindromic factors. This information will be crucial, as we will see that everything is concentrated on words with approximatively this frequency vector. We first work solely on the upper bound, then provide a matching lower bound.

3.1 Upper bound for the probability of a given α-gapped pattern

In this section we compute an upper bound for the probability that a given α-gapped palindromic factor $uv\overline{u}$ appears at position j in a random word w of length n, with $i \in [n+1-|uvu|]$. We introduce several real variables to express this probability in a convenient way: Let ℓ be positive real such that $|u| = \ell \log n$. Let r be the non-negative real such that $|v| = r|u|$. The length of $uv\overline{u}$ is therefore $(2+r)\ell \log n$, and $r \leq \alpha - 1$ because of the α-gapped condition. Let also $\boldsymbol{x} = (x_1, \ldots, x_k)$ be the frequency vector of u and let $\boldsymbol{y} = (y_1, \ldots, y_k)$ be the frequency vector of v.

The probability that $uv\overline{u}$ appears as a factor at position j of a random word of length n is $p_n(\boldsymbol{x}, \boldsymbol{y}, \ell, r) := \mathbb{P}_n(uv\overline{u} \text{ factor at position } j)$ defined by

$$p_n(\boldsymbol{x}, \boldsymbol{y}, \ell, r) = \prod_{i=1}^{k} p_j^{2x_i \ell \log n + y_i r \ell \log n} = n^{\ell \sum_{i \in [k]} (2x_i + r y_i) \log p_i}. \tag{1}$$

In particular, this probability does not depend on the position j.

By linearity of the expectation, the expected number of occurrences of $uv\overline{u}$ in a random word of length n is simply $(n+1-|uv\overline{u}|)p_n(\mathbf{x}, \mathbf{y}, \ell, r)$. Since the probability that a word appears as a factor is bounded from above by its expected number of occurrences, the probability $q_n(\mathbf{x}, \mathbf{y}, \ell, r)$ that $uv\overline{u}$ is factor of a random word of length n satisfies

$$q_n(\mathbf{x}, \mathbf{y}, \ell, r) \leq n p_n(\boldsymbol{x}, \boldsymbol{y}, \ell, r) = n^{1 + \ell \sum_{i \in [k]} (2x_i + r y_i) \log p_i}. \tag{2}$$

This upper bound can be greater than one, if the exponent is positive. Thus, we will use the following upper bound for the probability $q_n(\mathbf{x}, \mathbf{y}, \ell, r)$:

$$q_n(\mathbf{x}, \mathbf{y}, \ell, r) \leq \min(1, n p_n(\boldsymbol{x}, \boldsymbol{y}, \ell, r)) \leq n^{\min(0, 1 + \ell \sum_{i \in [k]} (2x_i + r y_i) \log p_i)}. \tag{3}$$

This idea that the probability of appearance is bounded by the minimum between 1 and the expected number of occurrences is central in [11, 3].

3.2 Upper bound for given frequency vectors

The result of this section is the following lemma.

Lemma 1. *Let $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{y} = (y_1, \ldots, y_k)$ be two frequency vectors, such that $(\ell \log n)\mathbf{x}$ and $(r\ell \log n)\mathbf{y}$ are integer-valued. The expected number of distinct α-gapped factors of the form $uv\overline{u}$ in a random word of length n, such that $|u| = \ell \log n$ and $|v| = r\ell \log n$, and such that u has frequency vector $\mathbf{x}$ and v has frequency vector $\mathbf{y}$, is bounded from above by $\lambda n^{\ell \min(J_r(\mathbf{x},\mathbf{y}), K_{\ell,r}(\mathbf{x},\mathbf{y}))}$, for some positive constant λ, where $J_r(\mathbf{x}, \mathbf{y})$ and $K_{\ell,r}(\mathbf{x}, \mathbf{y})$ are defined by*

$$J_r(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + rH(\mathbf{y}) \text{ and } K_{\ell,r}(\mathbf{x}, \mathbf{y}) = J_r(\mathbf{x}, \mathbf{y}) + \frac{1}{\ell} + \sum_{i=1}^{k} (2x_i + r y_i) \log p_i.$$

Proof. Let $\mu_n(\mathbf{x}, \mathbf{y}, \ell, r)$ be the expected number of distinct α-gapped factors described in the statement of the lemma. By linearity of the expectation, we have

$$\mu_n(\mathbf{x}, \mathbf{y}, \ell, r) = \sum_{\substack{u \in \mathcal{W}_{\ell \log n}(\mathbf{x}) \\ v \in \mathcal{W}_{r\ell \log n}(\mathbf{y})}} q_n(\mathbf{x}, \mathbf{y}, \ell, r) = |\mathcal{W}_{\ell \log n}(\mathbf{x})| \cdot |\mathcal{W}_{r\ell \log n}(\mathbf{y})| \cdot q_n(\mathbf{x}, \mathbf{y}, \ell, r).$$

So we just have to bound from above the cardinalities of $\mathcal{W}_{\ell \log n}(\mathbf{x})$ and of $\mathcal{W}_{r\ell \log n}(\mathbf{y})$ to conclude using Eq. (3). This is done using Lemma 1 of [3], which states that $|\mathcal{W}_{\ell \log n}(\mathbf{x})| \leq C n^{\ell H(\mathbf{x})}$ for some positive constant C. □

Observe that there can be no possibilities for u or v if the values of $\mathbf{x}$, $\mathbf{y}$, ℓ, n and r are such that one of the quantities $x_i \ell \log n$ or $y_i \ell \log n$ is not an integer. This is not a problem, as our statement is an upper bound.

3.3 Optimizing $J_r(\mathbf{x}, \mathbf{y})$ and $K_{\ell,r}(\mathbf{x}, \mathbf{y})$, for fixed ℓ and r

In this section we start to work on the upper bound provided by Lemma 1 by studying separately the two functions $J_r(\mathbf{x}, \mathbf{y})$ and $K_{\ell,r}(\mathbf{x}, \mathbf{y})$. The following lemma is a consequence of the properties of the entropy function.

Lemma 2. *For given $r \geq 0$, the function $J_r(\mathbf{x}, \mathbf{y})$ is maximized on the set of probability vectors for $\mathbf{x}$ and $\mathbf{y}$ when $\mathbf{x} = \mathbf{y} = \overline{\mathbf{x}}$, where $\overline{\mathbf{x}} = (\frac{1}{k}, \ldots, \frac{1}{k})$ is the uniform probability vector. For these values we have $J_r(\overline{\mathbf{x}}, \overline{\mathbf{x}}) = (1 + r) \log k$.*

The analysis of $K_{\ell,r}$ is not really complicated either. For any non-negative c, we define the probability vector $\mathbf{x}(c)$ by

$$\mathbf{x}(c) = \left(\frac{p_1^c}{\Phi(c)}, \ldots, \frac{p_i^c}{\Phi(c)}, \ldots, \frac{p_k^c}{\Phi(c)} \right).$$

In particular $\mathbf{x}(0)$ is the uniform probability vector $\overline{\mathbf{x}}$ and $\mathbf{x}(1) = \mathbf{p}$ is the distribution of the source. Recall also Gibbs' inequality [9], stating that if $\boldsymbol{s} = (s_1, \ldots, s_k)$ and $\boldsymbol{t} = (t_1, \ldots, t_k)$ are two probability vectors, then we have[4] $-\sum_{i=1}^k s_i \log s_i \leq -\sum_{i=1}^k s_i \log t_i$, with equality if and only if $\boldsymbol{s} = \boldsymbol{t}$.

Lemma 3. *For given $r \geq 0$ and $\ell > 0$, the function $K_{\ell,r}(\mathbf{x}, \mathbf{y})$ is maximized on the set of probabilities vector for $\mathbf{x}$ and $\mathbf{y}$ when $\mathbf{x} = \mathbf{x}(2)$ and $\mathbf{y} = \mathbf{p}$. For these values we have $K_{\ell,r}(\mathbf{x}(2), \mathbf{p}) = \frac{1}{\ell} + \log \Phi(2)$.*

Proof. We can rewrite $K_{\ell,r}(\mathbf{x}, \mathbf{y})$ the following way:

$$K_{\ell,r}(\mathbf{x}, \mathbf{y}) = \left(H(\mathbf{x}) + \sum_{i=1}^k x_i \log p_i^2 \right) + r \left(H(y) + \sum_{i=1}^k y_i \log p_i \right) + \frac{1}{\ell}.$$

[4] The case where some coordinates are zero is covered by setting $x \log x = 0$ for $x = 0$.

The second parenthesis is non-positive by direct application of Gibbs' inequality, and maximal for $\mathbf{y} = \mathbf{p}$, in which case it is equal to zero. For the first parenthesis, Gibbs' inequality also applies as

$$H(\mathbf{x}) + \sum_{i=1}^{k} x_i \log p_i^2 = -\sum_{i=1}^{k} x_i \log x_i + \sum_{i=1}^{k} x_i \log \frac{p_i^2}{\Phi(2)} + \log \Phi(2).$$

It is therefore maximal for $\mathbf{x} = \mathbf{x}(2)$, where it is equal to $\log \Phi(2)$. □

3.4 Optimizing $G_{\ell,r}(\mathbf{x}, \mathbf{y})$, for fixed ℓ and r

Let $G_{\ell,r}(\mathbf{x}, \mathbf{y}) = \min(J_r(\mathbf{x}, \mathbf{y}), K_{\ell,r}(\mathbf{x}, \mathbf{y}))$. This is the function to maximize in order to locate the maximum of the upper bound given in Lemma 1. We distinguish three cases (recall that J_ℓ is maximal at $(\overline{\mathbf{x}}, \overline{\mathbf{x}})$ and $K_{\ell,r}$ at $(\mathbf{x}(2), \mathbf{p})$):

(a) If $J_r(\overline{\mathbf{x}}, \overline{\mathbf{x}}) \leq K_{\ell,r}(\overline{\mathbf{x}}, \overline{\mathbf{x}})$: then $G_{\ell,r}(\overline{\mathbf{x}}, \overline{\mathbf{x}}) = J_r(\overline{\mathbf{x}}, \overline{\mathbf{x}})$, and thus, $G_{\ell,r}$ is maximal at $(\overline{\mathbf{x}}, \overline{\mathbf{x}})$, since $G_{\ell,r}(\mathbf{x}, \mathbf{y}) \leq J_r(\mathbf{x}, \mathbf{y})$.

(b) If $K_{\ell,r}(\mathbf{x}(2), \mathbf{p}) \leq J_r(\mathbf{x}(2), \mathbf{p})$: for the same reasons, $G_{\ell,r}$ is maximal at the point $(\mathbf{x}(2), \mathbf{p})$.

(c) If $J_r(\overline{\mathbf{x}}, \overline{\mathbf{x}}) > K_{\ell,r}(\overline{\mathbf{x}}, \overline{\mathbf{x}})$ and $K_{\ell,r}(\mathbf{x}(2), \mathbf{p}) > J_r(\mathbf{x}(2), \mathbf{p})$: since both J_r and $K_{\ell,r}$ are strictly concave (on the set of pairs of probabilities vectors), then $G_{\ell,r}$ has no local maximum on the set defined by $\{J_r(\mathbf{x}, \mathbf{y}) < K_{\ell,r}(\mathbf{x}, \mathbf{y})\}$, nor on the set $\{J_r(\mathbf{x}, \mathbf{y}) > K_{\ell,r}(\mathbf{x}, \mathbf{y})\}$. The global maximum of $G_{\ell,r}$ therefore lies on the set defined by the condition $J_r(\mathbf{x}, \mathbf{y}) = K_{\ell,r}(\mathbf{x}, \mathbf{y})$.

For convenience, we introduce the set $\mathcal{E}_{\ell,r}$ defined by $\mathcal{E}_{\ell,r} = \{(\mathbf{x}, \mathbf{y}) : J_r(\mathbf{x}, \mathbf{y}) = K_{\ell,r}(\mathbf{x}, \mathbf{y})\}$.

It is not difficult to identify the ranges for the different cases depending on the values of ℓ and r. The properties of the first two cases are summarized in the lemma below.

Lemma 4. *The first two cases are characterized as follows:*

- *We are in Case (a) if and only if* $(2 + r) \log k < \frac{1}{\ell}$. *In this case,* $G_{\ell,r}$ *has its maximum at* $(\overline{\mathbf{x}}, \overline{\mathbf{x}})$, *with* $G_{\ell,r}(\overline{\mathbf{x}}, \overline{\mathbf{x}}) = (2 + r) \log k$.
- *We are in Case (b) if and only if* $H(\mathbf{x}(2)) + rH(p) + \log \Phi(2) > \frac{1}{\ell}$. *In this case,* $G_{\ell,r}$ *has its maximum at* $(\mathbf{x}(2), \mathbf{p})$, *with* $G_{\ell,r}(\mathbf{x}(2), \mathbf{p}) = \frac{1}{\ell} + \log \Phi(2)$.

Observe that each condition in Lemma 4 defines an open subset of the (ℓ, r) domain, each delimited by a hyperbole segment. These two subsets cannot intersect, since in each domain $G_{\ell,r}$ is maximized at a different point (recall that we assumed $\mathbf{p} \neq \overline{\mathbf{x}}$). Thus, the closed subset of the (ℓ, r) domain defining Case (c) cannot be empty. Our result for Case (c), which constitutes the main technical contribution of this article, is the following.

Lemma 5. *Let* $\ell > 0$ *and* $r \geq 0$ *be two real numbers such that* $(2 + r) \log k \geq \frac{1}{\ell}$ *and* $H(\mathbf{x}(2)) + rH(p) + \log \Phi(2) \leq \frac{1}{\ell}$. *Then* $G_{\ell,r}(\mathbf{x}, \mathbf{y})$ *reaches its unique maximum*

for $\mathbf{x} = \mathbf{x}(2c)$ *and* $\mathbf{y} = \mathbf{x}(c)$*, where* $c \in (0,1)$ *is the unique positive solution of the equation*

$$1 + \ell \frac{\Phi'(2t)}{\Phi(2t)} + \ell r \frac{\Phi'(t)}{\Phi(t)} = 0. \tag{4}$$

Proof. By Lemma 4, the hypothesis of the lemma implies that we are in Case (c). Hence, $G_{\ell,r}$ reaches its maximum, for probability vectors of $\mathcal{E}_{\ell,r}$; this is equivalent to $\frac{1}{\ell} + \sum_i (2x_i + r y_i) \log p_i = 0$, which is a linear condition on the $2k$ variables defining $(\mathbf{x}, \mathbf{y})$.

On the considered domain we have $G_{\ell,r}(\mathbf{x},\mathbf{y}) = J_r(\mathbf{x},\mathbf{y}) = K_{\ell,r}(\mathbf{x},\mathbf{y})$, so we take $J_r(\mathbf{x},\mathbf{y})$ which is easier to study. Its gradient, as a function of $2k$ non-negative variables (and thus, not only for probability vectors) is the following:

$$\frac{\partial J_r}{\partial x_i} = -1 - \log x_i, \text{ and } \frac{\partial J_r}{\partial y_i} = -r - r \log y_i, \ \forall i \in [k].$$

Since we are looking for a pair of probability vectors $(\mathbf{x}, \mathbf{y})$ that lies in the set $\mathcal{E}_{\ell,r}$ the following linear constraints must be satisfied: $\sum_{i\in[k]} x_i = 1$, $\sum_{i\in[k]} y_i = 1$ and $\frac{1}{\ell} + \sum_{i\in[k]} (2x_i + r y_i) \log p_i = 0$. Let $\boldsymbol{n}_1$, $\boldsymbol{n}_2$ and $\boldsymbol{n}$ be the three vectors normal to the constraints defined by

$$\boldsymbol{n}_1 = (\underbrace{1,\dots,1}_{k \text{ times}}, \underbrace{0,\dots,0}_{k \text{ times}}); \quad \boldsymbol{n}_2 = (\underbrace{0,\dots,0}_{k \text{ times}}, \underbrace{1,\dots,1}_{k \text{ times}});$$
$$\boldsymbol{n} = (2\log p_1, \dots, 2\log p_k, r\log p_1, \dots, r \log p_k).$$

To locate our optimal value, we have to find where the gradient lies in the vector space spanned by $\boldsymbol{n}_1$, $\boldsymbol{n}_2$ and $\boldsymbol{n}$. We are thus looking for vectors $\mathbf{x}$ and $\mathbf{y}$ for which there exist three constants c_1, c_2, c_3 such that, for all $i \in [k]$,

$$-1 - \log x_i = c_1 + 2c_3 \log p_i, \text{ and } -r - r\log y_i = c_2 + c_3 r \log p_i.$$

By differences (keeping the equations involving x_1 and y_1), this leads to

$$\log\left(\frac{x_i}{x_1}\right) = -2c_3 \log\left(\frac{p_i}{p_1}\right) \text{ and } \log\left(\frac{y_i}{y_1}\right) = -c_3 \log\left(\frac{p_i}{p_1}\right).$$

Introducing parameter $c = -c_3$, we get that the x_i's must be proportional to p_i^{2c} (with a normalizing constant $\Phi(2c)$) and the y_i's must be proportional to p_i^c (with a normalizing constant $\Phi(c)$) for the *same* constant c. That is, we must take $(\mathbf{x},\mathbf{y}) = (\mathbf{x}(2c), \mathbf{x}(c))$ for some constant c; the parameters c_1 and c_2 are then easily recovered as, from the equation involving x_1:

$$c_1 = -1 - \log\left(\frac{p_1^{2c}}{\Phi(2c)}\right) + 2c \log p_i = -1 + \log \Phi(2c),$$

and, from the equation involving y_1, we have $c_2 = -r - r\log p_i + cr \log p_i$. Thus the only remaining equation is the one asking for $(\mathbf{x},\mathbf{y})$ to be in the set $\mathcal{E}_{\ell,r}$. Setting $\mathbf{x} = \mathbf{x}(2c)$ and $\mathbf{y} = \mathbf{x}(c)$ in the equation, we get

$$-\frac{1}{\ell} = \frac{2}{\Phi(2c)} \sum_{i=1}^{k} p_i^{2c} \log p_i + \frac{r}{\Phi(c)} \sum_{i=1}^{k} p_i^c \log p_i, \tag{5}$$

or equivalently,

$$-\frac{1}{\ell} = 2\frac{\Phi'(2c)}{\Phi(2c)} + r\frac{\Phi'(c)}{\Phi(c)}. \tag{6}$$

Now consider the function $\frac{\Phi'}{\Phi}$ appearing on the right-hand side of (6). Its derivative is $\frac{\Phi''\Phi - \Phi'^2}{\Phi^2}$, which is strictly positive by an application of Cauchy-Schwarz inequality (we use here that $\mathbf{p} \neq \overline{\mathbf{x}}$). Thus, the right-hand side of (4) is an increasing, continuous function of c when r and ℓ are considered as fixed parameters. Evaluating this function for $c = 0$ yields a value less than $-1/\ell$, from the condition that we are not in Case (a), and for $c = 1$, a value more than $-1/\ell$, from the condition that we are not in Case (b). This proves that the equation has a unique solution c, and that this solution lies strictly between 0 and 1. □

3.5 Optimizing the exponent on ℓ and r

At this point, we know where the maximum is when ℓ and r are fixed, for all three ranges. We still have to find the values of ℓ and r that maximize the exponent $\ell G_{\ell,r}(\mathbf{x}, \mathbf{y})$, in each case, for the maximal $\mathbf{x}$ and $\mathbf{y}$. Recall that $r \in [0, \alpha - 1]$ in our settings. The next two lemmas are directly derived from Lemma 4.

Lemma 6. *For Case (a), the maximum of $\ell G_{\ell,r}(\mathbf{x}, \mathbf{y})$ is reached for $\ell = \ell_0$, $r = \alpha - 1$ and $\mathbf{x} = \mathbf{y} = \overline{\mathbf{x}}$, where $\ell_0 = \frac{1}{(2+r)\log k}$; we have $\ell_0 G_{\ell_0,\alpha-1}(\overline{\mathbf{x}}, \overline{\mathbf{x}}) = \frac{\alpha}{\alpha+1}$.*

Lemma 7. *For Case (b), the maximum of $\ell G_{\ell,r}(\mathbf{x}, \mathbf{y})$ is reached for $\ell = \ell_1$, $r = \alpha - 1$, $\mathbf{x} = \mathbf{x}(2)$ and $\mathbf{y} = \mathbf{p}$, where $\ell_1 = \frac{1}{H(\mathbf{x}(2)) + rH(p) - \log \Phi(2)}$; we have $\ell_1 G_{\ell_1,\alpha-1}(\overline{\mathbf{x}}, \overline{\mathbf{x}}) = \frac{H(\mathbf{x}(2)) + (\alpha-1)H(p)}{H(\mathbf{x}(2)) + (\alpha-1)H(p) - \log \Phi(2)}$.*

Observe that ℓ_0 and ℓ_1 are not formally in the ranges for Case (a) and Case (b) as they have been defined. We allow this abuse of notation, as the exponent function can be extended by continuity at ℓ_0 and at ℓ_1.

We now focus on Case (c). By Lemma 5 we know that the maximum of $G_{\ell,r}$ is reached at some point $(\mathbf{x}(2c), \mathbf{x}(c))$ of $\mathcal{E}_{\ell,r}$, for some $c \in (0, 1)$ that is implicitly defined. Our optimization is on the variables ℓ and r, and c is viewed as a function of these two variables. We introduce the notation $h(c) = H(\mathbf{x}(c))$, for which an elementary calculation yields $h(c) = \log \Phi(c) - c\frac{\Phi'(c)}{\Phi(c)}$. As $G_{\ell,r} = J_r$ on $\mathcal{E}_{\ell,r}$, our problem reduces to maximizing the exponent $E(r, \ell, c) = \ell h(2c) + \ell r h(c)$ subject to the constraint $1 + 2\ell\frac{\Phi'(2c)}{\Phi(2c)} + \ell r\frac{\Phi'(c)}{\Phi(c)} = 0$. The solution is given in the following lemma.

Lemma 8. *For Case (c), the maximum of $\ell G_{\ell,r}(\mathbf{x}, \mathbf{y})$ is reached for $\ell = \ell_{c^*}$, $r = \alpha - 1$, $\mathbf{x} = \mathbf{x}(2c^*)$ and $\mathbf{y} = \mathbf{x}(c^*)$, where c^* is the solution of the equation*

$$\Phi(2c)\Phi(c)^{\alpha-1} = 1. \tag{7}$$

At this point, the value of the exponent is $\ell_{c^*} G_{\ell_{c^*},\alpha-1}(\mathbf{x}(2c^*), \mathbf{x}(c^*)) = c^*$.

Proof. Remark that, using the simplification of $h(c)$, we have $E(\ell, r, c) = c + \ell\left(\log \Phi(2c) + r \log \Phi(c)\right)$. Lets us consider c as an implicit function of (ℓ, r), so that E becomes a function of only two variables ℓ and r, for which we compute partial derivatives:

$$\begin{aligned}
\frac{\partial E}{\partial r}(\ell, r) &= \frac{\partial c}{\partial r}(\ell, r) + 2\ell \frac{\partial c}{\partial r}(\ell, r)\frac{\Phi'(2c)}{\Phi(2c)} + \ell \log \Phi(c) + \ell r \frac{\partial c}{\partial r}(\ell, r)\frac{\Phi'(c)}{\Phi(c)} \\
&= \ell \log \Phi(c) + \frac{\partial c}{\partial r}(\ell, r)\left(1 + 2\ell\frac{\Phi'(2c)}{\Phi(2c)} + \ell r \frac{\Phi'(c)}{\Phi(c)}\right) = \ell \log \Phi(c); \\
\frac{\partial E}{\partial \ell}(\ell, r) &= \frac{\partial c}{\partial \ell}(\ell, r) + \log \Phi(2c)) + r \ln(\Phi(c)) + \ell\left(2\frac{c}{\ell}\frac{\Phi'(2c)}{\Phi(2c)} + \frac{\partial c}{\partial \ell}(\ell, r) r \frac{\Phi'(c)}{\Phi(c)}\right) \\
&= \log \Phi(2c) + r \log \Phi(c).
\end{aligned}$$

The expression for $\frac{\partial E}{\partial r}$ shows it to be positive ($c < 1$ implies $\Phi(c) > 1$), so the maximum is obtained when r is as large as possible, namely, $r = \alpha - 1$. The expression for $\frac{\partial E}{\partial \ell}$ provides a candidate for a maximum, where it reaches zero. This happens for c solution of the equation

$$\log \Phi(2c) + (\alpha - 1) \log \Phi(c) = 0. \tag{8}$$

We compute the second derivative in ℓ to verify that it is a local maximum:

$$\begin{aligned}
\frac{\partial^2 E}{\partial \ell^2}(\ell, r) &= 2\frac{\partial c}{\partial \ell}(\ell, r)\frac{\Phi'(2c)}{\Phi(2c)} + r\frac{\partial c}{\partial \ell}(\ell, r)\frac{\Phi'(c)}{\Phi(c)} \\
&= \frac{\partial c}{\partial \ell}(\ell, r)\underbrace{\left(2\frac{\Phi'(2c)}{\Phi(2c)} + r\frac{\Phi'(c)}{\Phi(c)}\right)}_{=-\frac{1}{\ell}} = -\frac{1}{\ell}\frac{\partial c}{\partial \ell}(\ell, r).
\end{aligned}$$

A closer look at Equation (4), the implicit equation for c, yields that c is an increasing function of ℓ. Hence $\frac{\partial^2 E}{\partial \ell^2}(\ell, r)$ is negative, and thus it has a unique maximum for c solution of Equation (8). This equation is equivalent to Equation (7), concluding the proof. □

Lemma 6, Lemma 7 and Lemma 8 describe the various maximum exponents obtained in the three cases; we now combine the three into a single result. It is obtained by remarking that in all three cases, the maximum exponent is reached on the line $r = \alpha - 1$, and that it is a continuous function of ℓ.

Proposition 1. *The maximum exponent over all choices of ℓ, r, $\mathbf{x}$ and $\mathbf{y}$ is obtained for Case (c), namely, $r = \alpha - 1$, $\ell = \ell_{c^*}$ where $\Phi(2c^*)\Phi(c^*)^{\alpha-1} = 1$, $\mathbf{x} = \mathbf{x}(2c^*)$ and $\mathbf{y} = \mathbf{x}(c^*)$; this maximum exponent is equal to c^*.*

3.6 Proof of Theorem 1

Up to now, we have optimized over continuous domains for ℓ, r, $\mathbf{x}$ and $\mathbf{y}$. We now deal properly with the fact that they must be rational numbers and vectors that correspond to actual composition vectors for words. We let ℓ^*, c^*,

$\mathbf{x}^* = (x_1^*, \ldots, x_k^*)$ and $\mathbf{y}^* = (y_1^*, \ldots, y_k^*)$ denote the real-valued optimal solution describe in the previous section. Let $L = \lfloor \ell^* \log n \rfloor$, $R = \lfloor (\alpha - 1)L \rfloor$, and define the vectors $\mathbf{x}$ and $\mathbf{y}$ as follows: for any $i \in [k-1]$, $x_i = \frac{1}{L}\lfloor x_i^* L \rfloor$ and $y_i = \frac{1}{R}\lfloor y_i^* R \rfloor$; finally, let $x_k = 1 - \sum_{i<k} x_i$ and let $y_k = 1 - \sum_{i<k} y_i$. Defined this way, $L\mathbf{x}$ and $R\mathbf{y}$ are vectors of integers summing to L and R respectively, so we can look at the expected number of distinct α-gapped palindromic factors $uv\overline{u}$ where u has composition exactly $L\mathbf{x}$, and v has composition exactly $L\mathbf{y}$; we write this expected number as n^c (equivalently, c is the logarithm to base n of the expected number of factors).

Since $||\mathbf{x} - \mathbf{x}^*|| = O(1/\log n)$ and $||\mathbf{y} - \mathbf{y}^*|| = O(1/\log n)$, we have that $c = c^* - O(1/\log n)$ (this would be $1/\log^2 n$ if both partial derivatives of E vanished at $(\ell^*, \alpha - 1)$, but $\partial E/\partial r$ does not vanish). Thus, we have $n^c = n^{c^* - O(\log n)} = e^{c^* \log n - O(1)} = \Theta(n^{c^*})$.

Now let $\ell^+ = \frac{3}{|\log p_{\max}|}$. For any choice of ℓ and r with $\ell^+ \leq \ell \leq n/\log n$ and $0 \leq R \leq \alpha \ell \log n$, the probability of having an α-gapped palindromic factor $uv\overline{u}$ at some position j, with $|u| = \ell \log n$ and $|v| = R$, is at most $p_{\max}^{\ell \log n} \leq n^{-3}$: for any choice of u and v, the probability that the next $\ell \log n$ letters are exactly those of u is upper bounded by $p_{\max}^{\ell \log n}$. Since there are fewer than n^3 choices for the triple $(|u|, |v|, j)$, the expected number of such "long" factors is less than 1.

As a consequence, the dominant contribution to the expected number of gapped palindromic factors comes from those with $|u| \leq \ell^+ \log n$. Each possible composition vector for u and v contributes less than n^{c^*}, and there are at most $(\ell^+ \log n)^{2k}(\alpha - 1)^k = \tilde{O}(1)$ such composition vectors; thus the $\tilde{\Theta}(n^{c^*})$ bound carries over for the expected total number of distinct α-gapped palindromic factors of all possible lengths. □

4 Typical composition vectors of palindromic factors

In this section, we show that with asymptotic probability 1, *most* gapped palindromic factors present in a large random word will be as described in the upper bound computations of the previous section. For this, we must first prove that our previous result on the *expected* number of gapped palindromic factors hold with good enough probability for the random variable that counts these factors.

Theorem 2. *There exist two constants $a < 0$ and $b > 0$ such that, with asymptotic probability 1 (when n tends to infinity), the number $\Gamma_{\alpha,n}$ of distinct α-gapped palindromic factors in a random word of length n lies between $n^{c^*} \log^a(n)$ and $n^{c^*} \log^b(n)$.*

Proof. The upper bound is a direct consequence of applying Markov's inequality to the bound on the expectation of Theorem 1: if we simply multiply by $\log^\varepsilon n$ the upper bound in the $\tilde{\Theta}(n^{c^*})$ in the theorem, the probability that $\Gamma_{\alpha,n}$ is higher than this new bound is $O((\log n)^{-\varepsilon}) = o(1)$.

For the lower bound, we now prove that with high enough probability, $\Gamma_{\alpha,n}$ is at least $\mathbb{E}(\Gamma_{\alpha,n})/\log n^d$ for some $d > 0$. We will do this by proving such a lower bound for the factors appearing in a subset of the possible positions in the word:

Let $m = n/(2+\alpha)\ell^* \log n$. Our word of length n is obtained[5] by concatenating m independent words $W_1, \ldots, W_m$, each of length $(2+\alpha)\ell^* \log n$.

Now, for each possible α-gapped palindrome w of length $(2+\alpha)\ell^* \log n$, and for each integer $1 \leq i \leq m$, define the Bernoulli random variable $X_{i,w}$ as 1 if $W_i = w$, and 0 otherwise; then define X_w as $\max_i X_{i,w}$, *i.e.*, $X_w = 1$ if and only if w appears in a factor in one of the m positions in the whole random word. Finally, set $X = \sum_w X_w$: X is the total number of distinct α-gapped palindromic factors that appear in at least one of the m positions. Thus, $X \leq G_{\alpha,n}$, but $\mathbb{E}(X) = \tilde{\Theta}(n^{c^*})$ (we lose a factor of at most $\log n$ because we only consider $\Theta(n/\log n)$ positions instead of n, but this is absorbed by the $\tilde{\Theta}$ notation).

The collection of random variables $(X_{i,w})$ is negatively associated in the sense of [2], so that (by [2], Prop. 7), the classical Chernoff-Hoeffding bounds apply to X. This is enough to prove (using, for example, [10], Thm 4.2) that the probability of X being less than half its expectation is exponentially small. □

Our final result ensures that, with probability close to 1, almost all α-gapped palindromic factors that appear in the random word have composition vectors close to the typical vectors described earlier.

Theorem 3. *We again let $\Gamma_{\alpha,n}$ denote the total number of distinct α-gapped palindromic factors of a random word of length n; and, for any $\varepsilon > 0$, we let $\Gamma_{\alpha,n,\varepsilon}$ denote the total number of these factors whose frequency vectors lie within distance ε of the optimal vectors $\mathbf{x}^*$ and $\mathbf{y}^*$. Then, for any $\delta > 0$, with asymptotic probability 1, we have $\Gamma_{\alpha,n,\varepsilon} \geq (1-\delta)\Gamma_{\alpha,n}$.*

Proof. We already know that, with asymptotic probability 1, $\Gamma_{\alpha,n,\varepsilon} \geq n^{c^*}/\log n^a$ for some $a > 0$. From the proof of Lemma 8, we know that any frequency vectors at distance at least ε from $(\mathbf{x}, \mathbf{y})$ come with an exponent at most $c^* - \beta\varepsilon^2$ (for some $\beta > 0$ which only depends on the second derivative of E at the critical point), so that the total expected number of distinct α-gapped factors with such frequency vectors is $\tilde{\Theta}(n^{c^*-\beta\varepsilon^2})$. Markov's inequality turns this into a high probability bound at the cost of logarithmic factors absorbed into the $\tilde{\Theta}$ notation, and with high probability, $\Gamma_{\alpha,n,\varepsilon}$ is within a factor $1 - \tilde{O}(n^{-\beta\varepsilon^2})$ of $\Gamma_{\alpha,n}$.

5 Conclusion

In this article we show that the expected number of distinct α-gapped palindromic factors in a random word of length n is $\tilde{\Theta}(n^{c^*})$, where c^* is implicitly defined as the solution of some equation depending on the probability $\mathbf{p}$ of the source and of α. Moreover, for any positive ε, the frequency vectors of u and v of such factors $uv\overline{u}$ are likely to be at distance at most ε of $\mathbf{x}^*$ and $\mathbf{y}^*$.

To conclude, we want to emphasize that the techniques we used follow those we introduced in [3]. These methods therefore prove useful to study the number

[5] We disregard rounding errors in lengths; properly dealing with them by means of integer parts would only yield clumsier notations without changing the asymptotic results.

of distinct elements (palindromes, subwords, ...) in different settings, for random words generated by a memoryless distribution.

References

1. M. Crochemore, R. Kolpakov, and G. Kucherov. Optimal bounds for computing \alpha α -gapped repeats. In A. Dediu, J. Janousek, C. Martín-Vide, and B. Truthe, editors, *Language and Automata Theory and Applications - 10th International Conference, LATA 2016, Prague, Czech Republic, March 14-18, 2016, Proceedings*, volume 9618 of *Lecture Notes in Computer Science*, pages 245–255. Springer, 2016.
2. D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.
3. P. Duchon and C. Nicaud. On the biased partial word collector problem. In M. A. Bender, M. Farach-Colton, and M. A. Mosteiro, editors, *LATIN 2018: Theoretical Informatics - 13th Latin American Symposium, Buenos Aires, Argentina, April 16-19, 2018, Proceedings*, volume 10807 of *Lecture Notes in Computer Science*, pages 413–426. Springer, 2018.
4. P. Duchon, C. Nicaud, and C. Pivoteau. Gapped pattern statistics. In J. Kärkkäinen, J. Radoszewski, and W. Rytter, editors, *28th Annual Symposium on Combinatorial Pattern Matching, CPM 2017, July 4-6, 2017, Warsaw, Poland*, volume 78 of *LIPIcs*, pages 21:1–21:12. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017.
5. M. Dumitran, P. Gawrychowski, and F. Manea. Longest gapped repeats and palindromes. *Discrete Mathematics & Theoretical Computer Science*, 19(4), 2017.
6. P. Gawrychowski, T. I, S. Inenaga, D. Köppl, and F. Manea. Tighter bounds and optimal algorithms for all maximal α-gapped repeats and palindromes - finding all maximal α-gapped repeats and palindromes in optimal worst case time on integer alphabets. *Theory Comput. Syst.*, 62(1):162–191, 2018.
7. R. Kolpakov and G. Kucherov. Searching for gapped palindromes. *Theoretical Computer Science*, 410(51):5365–5373, 2009.
8. R. Kolpakov, M. Podolskiy, M. Posypkin, and N. Khrapov. Searching of gapped repeats and subrepetitions in a word. *Journal of Discrete Algorithms*, 46-47:1–15, 2017.
9. D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.
10. R. Motwani and P. Raghavan. Randomized algorithms. *ACM Computing Surveys (CSUR)*, 28(1):33–37, 1996.
11. M. Rubinchik and A. M. Shur. The number of distinct subpalindromes in random words. *Fundam. Inform.*, 145(3):371–384, 2016.
12. M. Rubinchik and A. M. Shur. EERTREE: an efficient data structure for processing palindromes in strings. *Eur. J. Comb.*, 68:249–265, 2018.
13. Y. Tanimura, Y. Fujishige, T. I, S. Inenaga, H. Bannai, and M. Takeda. A faster algorithm for computing maximal \alpha -gapped repeats in a string. In C. S. Iliopoulos, S. J. Puglisi, and E. Yilmaz, editors, *String Processing and Information Retrieval - 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings*, volume 9309 of *Lecture Notes in Computer Science*, pages 124–136. Springer, 2015.