Maximal Number of Subword Occurrences in a Word

Wenjie Fang
Univ Gustave Eiffel, CNRS, LIGM, F-77454 Marne-la-Vallée, France

Abstract

We consider the number of occurrences of subwords (non-consecutive sub-sequences) in a given word. We first define the notion of subword entropy of a given word that measures the maximal number of occurrences among all possible subwords. We then give upper and lower bounds of minimal subword entropy for words of fixed length in a fixed alphabet, and also showing that minimal subword entropy per letter has a limit value. A better upper bound of minimal subword entropy for a binary alphabet is then given by looking at certain families of periodic words. We also give some conjectures based on experimental observations.

2012 ACM Subject Classification Mathematics of computing → Enumeration; Mathematics of computing → Combinatorics on words

Keywords and phrases Subword occurrence, subword entropy, enumeration, periodic words

Digital Object Identifier 10.4230/LIPIcs.AofA.2024.3

Related Version *Full Version*: `https://arxiv.org/abs/2406.02971`

Supplementary Material
Software (Source Code): `https://github.com/fwjmath/maxocc-subword` [3]
archived at `swh:1:dir:fef689a6896632f63f67b460e989fc106d5899e0`

Acknowledgements I would like to thank Stéphane Vialette for bringing the question of maximal number of subword occurrences of a given word to our attention, and for giving an idea for Proposition 3.4.

1 Introduction

Enumeration problems concerning patterns have been rich sources of interesting combinatorics. The most famous examples are classes of permutations avoiding a given pattern. We refer readers to [9, 14] for an exposition of such results. In this article, we will consider enumeration about patterns in a word, which is in general easier than that for permutations.

There are two different widely used notions of patterns for words. The first notion is that of a *factor*. A word v occurs in another word w *as a factor* if there is a consecutive segment of w equal to v. The second notion is that of a *subword*. A word v occurs in another word w *as a subword* if we can obtain v by deleting letters in w. A factor of w is always a subword of w, but not *vice versa*. There are also other notions of patterns, such as the one in [2] that generalizes both factors and subwords, but we will not discuss them here.

Unlike for permutations, the enumeration of classes of words avoiding a (set of) given subwords or factors is already known in the sense that, for a given subword or factor, we can express their avoidance in regular expressions, leading automatically (no pun intended) to the generating function of such classes, which is always rational and can be effectively computed [5, Section V.5]. There is also some work on counting words with a fixed number of occurrences of a given pattern, for example [2, 11]. For the other end of the spectrum, the problem of maximal density of certain patterns in words is considered by Burstein, Hästö and Mansour in [1]. Readers are referred to the survey-book of Kitaev [9] for more of such results. In general, such results are non-trivial, due to the possible overlap of patterns.

35th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2024).
Editors: Cécile Mailler and Sebastian Wild; Article No. 3; pp. 3:1–3:12
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

We may also consider all patterns that occur in a given word. For the notion of factors, this idea leads to the notion of factor complexity of a word w, first defined by Morse and Hedlund in [13] and also called "subword complexity", which is a function f_w such that $f_w(k)$ is the number of distinct factors in w of length k. In [7], Gheorghiciuc and Ward studied the factor complexity of random words. We may also want to consider the number of occurrences of a given pattern in a word. The work of Flajolet, Szpankowski and Vallée [6] establishes a Gaussian limit law and large deviations for the number of occurrences of a given subword in a long random word, again by analyzing overlap of occurrences of subwords. The number of occurrences of a given pattern is of particular interest in algorithmics with applications in data mining, in which researchers propose algorithms finding patterns with large number of occurrences [8] and study complexity of such problems [15].

In this article, we take a further step on enumeration problems on patterns by considering the number of occurrences of all subwords in a word. More precisely, we can see a given word w that permits frequent occurrences of some subword w' as having some "large space" for such a subword, and we would like to measure the "extend" of such space, or from the opposite direction, the "disorder" generated by the possible different occurrences. To this end, we define a notion of *subword entropy*, which measures the maximal number of times that any subword can occur in a given word. We delay its precise definition to later sections. We then look at the minimal subword entropy of all words of a given length n in an alphabet of k letters, denoted by $\min S_{\mathrm{sw}}^{(k)}(n)$, as it is easy to find the ones with maximal subword entropy. Using the super-additivity of minimal subword entropy, we show that $\min S_{\mathrm{sw}}^{(k)}(n)/n$ has a finite limit L_k. We then concentrate on the binary case, showing some upper bounds of L_2 by looking at certain families of periodic words, inspired by experimental data. As a by-product, we also show that, given two words w and v, the generating function of the number of occurrences of v^r in w^m is rational.

The rest of this article is organized as follows. We first give necessary definitions in Section 2, then some basic results on subword occurrences and minimal subword entropy in Section 3, including the proof of the existence of the limit L_k of $\min S_{\mathrm{sw}}^{(k)}(n)/n$, and bounds of L_k. Then in Section 4, we focus on the case of binary alphabet, and shows a better upper bound of L_2 than the one given in Section 3. We end in Section 5 with a discussion on open problems partially inspired by experimental results obtained for the binary case.

2 Preliminaries

A *word* w of length n is a sequence $w = (w_1, \ldots, w_n)$ of elements in a finite set $\mathcal{A}$ called the *alphabet*. We denote by $|w|$ the length of w, and $|w|_a$ the number of letters a in w. For two words v, w, their concatenation is denoted by $v \cdot w$. We also denote by ϵ the empty word of length 0. A *run* in a word w is a maximal consecutive segment in w formed by only one letter in $\mathcal{A}$. Given a word w, if there is another word $w' = (w'_1, \ldots, w'_k)$ such that there is some set $P = \{p_1 < \cdots < p_k\}$ of integers from 1 to n satisfying $w_{p_j} = w'_j$ for all $1 \leq j \leq k$, then we say that w' is a *subword* of w, and we call the set P an *occurrence* of w' in w. We denote by $\mathrm{occ}(w, w')$ the number of occurrences of w' in w. For instance, for $w = 011001$ and $w' = 01$, there are 5 occurrences of w' in w, which are $\{1,2\}, \{1,3\}, \{1,6\}, \{4,6\}, \{5,6\}$. When w' is not a subword of w, we have $\mathrm{occ}(w, w') = 0$, and when $w' = \epsilon$, we have $\mathrm{occ}(w, \epsilon) = 1$.

It is easy to find words who have a subword with a large number of occurrences. For instance, with $w = a^n$ for some letter $a \in \mathcal{A}$, the subword $w' = a^{\lfloor n/2 \rfloor}$ appears $\binom{n}{\lfloor n/2 \rfloor} \sim \left(\frac{2}{\pi n}\right)^{1/2} 2^n$ times. It is more difficult to find words in which no subword occurs frequently. To quantify such intuition, we define the *maximal subword occurrences* $\mathrm{maxocc}(w)$ of a word

w to be the maximal value of $\mathrm{occ}(w, w')$, and subwords w' reaching this value are called *most frequent subwords* of w. We note that a word w may have several most frequent subwords. We then define the *subword entropy* $S_{\mathrm{sw}}(w)$ of w in an alphabet of size k by

$$S_{\mathrm{sw}}(w) := \log_2 \mathrm{maxocc}(w).$$

We note that this definition does not depend on the size of the alphabet, as subword occurrences are fundamentally about subsets of positions, and the size of the alphabet is implicit in the word w. Now, finding words in which no subword occurs frequently is to find words minimizing their subword entropy. We define the minimal subword entropy for words of length n in an alphabet of size k by

$$\mathrm{min}S_{\mathrm{sw}}^{(k)}(n) := \min_{w \in \mathcal{A}^n, |\mathcal{A}|=k} S_{\mathrm{sw}}^{(k)}(w).$$

3 Some basic results

We start with some simple properties of $\mathrm{occ}(w, u)$.

▶ **Lemma 3.1.** *For words w, w', u, u', we have $\mathrm{occ}(w \cdot w', u \cdot u') \geq \mathrm{occ}(w, u)\,\mathrm{occ}(w', u')$.*

Proof. Let P (resp. P') be an occurrence of u (resp. u') in w (resp. w'). The set $Q = P \cup \{p' + |w| \mid p' \in P'\}$ is an occurrence of $u \cdot u'$ in $w \cdot w'$, and the map $(P, P') \mapsto Q$ is clearly injective. ◀

▶ **Lemma 3.2.** *For a word w, it has a most frequent subword u with $w_1 = u_1$ and $w_{|w|} = u_{|w'|}$.*

Proof. Let v be a most frequent subword of w. If $v_1 \neq w_1$, then for all occurrences P of v in w, we have $1 \notin P$. Then, $\{1\} \cup P$ is an occurrence of $v' = w_1 \cdot v$, which is thus also a most frequent subword. Otherwise, we take $v' = v$. We repeat the same reasoning on v' for the last letter of w to obtain u. ◀

We now give simple upper and lower bounds for $\mathrm{maxocc}(w)$ for any word w.

▶ **Proposition 3.3.** *Given an alphabet $\mathcal{A}$ of size k and $n \geq 1$, for any word $w \in \mathcal{A}^n$, we have $\mathrm{maxocc}(w) \leq \binom{n}{\lceil n/2 \rceil}$, and it is realized exactly by $w = a^n$ for any letter $a \in \mathcal{A}$.*

Proof. For w' of length k, as occurrences of w' in w are subsets of $\{1, \ldots, n\}$, we have $\mathrm{occ}(w, w') \leq \binom{n}{k} \leq \binom{n}{\lceil n/2 \rceil}$. It is clear that only words composed by the same letter reach this bound. ◀

▶ **Proposition 3.4.** *Given an alphabet $\mathcal{A}$ of size k and $n \geq 1$, when $n \to \infty$, for any word $w \in \mathcal{A}^n$, we have $\ln \mathrm{maxocc}(w) \geq \max_{0 \leq \ell \leq n} \ln\left(\binom{n}{\ell} k^{-\ell}\right)$, with $\ell = \lfloor \frac{n}{k+1} \rfloor$ giving the asymptotically maximized value $n \ln(1 + k^{-1}) - \frac{1}{2}(\ln n) + O(1)$.*

Proof. Let u be a uniformly chosen word of length ℓ. We have

$$\mathbb{E}[\mathrm{occ}(w, u)] = \sum_{P \subseteq \{1, \ldots, n\}, |P| = \ell} \mathbb{P}[u \text{ occurs in } w \text{ at positions } P] = \binom{n}{\ell} k^{-\ell}.$$

The first equality is from linearity of expectation, and the second from the fact that u is uniformly chosen at random, and the probability does not depend on P. Hence, there is some u^* with $\mathrm{occ}(w, u^*) \geq \mathbb{E}[\mathrm{occ}(w, u)]$, implying the non-asymptotic part of our claim.

For the asymptotic part, take $\alpha = \ell/n$. Using Stirling's approximation, we have

$$\ln\left(\binom{n}{\ell}k^{-\ell}\right) = n\Big[-\alpha - \ln\alpha - (1-\alpha)\ln(1-\alpha) - \alpha\ln k\Big] - \frac{1}{2}\ln n + O(1).$$

The coefficient of n above is maximized for $\alpha = (k+1)^{-1}$, with value $\ln(1+k^{-1})$. We thus have our claim on the asymptotic growth. ◀

▶ **Corollary 3.5.** *There are constants c_1, c_2 such that, for all $n \in \mathbb{N}$ and $w \in \mathcal{A}^n$ with $|\mathcal{A}| = k \geq 2$, we have*

$$\log_2(1+k^{-1})n - \frac{1}{2}\log_2 n + c_1 \leq \min S_{\mathrm{sw}}^{(k)}(n) \leq S_{\mathrm{sw}}(w) \leq n - \frac{1}{2}\log_2 n + c_2.$$

Proof. The bounds on $S_{\mathrm{sw}}(w)$ result from combining Propositions 3.3 and 3.4 with $\ln\binom{n}{\lceil n/2\rceil} = n\ln 2 - \frac{1}{2}\ln n + O(1)$. The bounds for $\min S_{\mathrm{sw}}^{(k)}(n)$ then follows. ◀

We now show that there is a limit for $\min S_{\mathrm{sw}}^{(k)}(n)/n$. To this end, we need the well-known Fekete's lemma [4] for super-additive sequences.

▶ **Lemma 3.6.** *Suppose that a sequence $(g_n)_{n\geq 1}$ satisfies that, for all $n, m \geq 1$, we have $g_{n+m} \geq g_n + g_m$. Then, for $n \to +\infty$, the value of g_n/n either tends to $+\infty$ or converges to some limit L.*

We first show that the function $\min S_{\mathrm{sw}}^{(k)}(n)$ is super-additive.

▶ **Proposition 3.7.** *Given $k \geq 2$, for any $n, m \geq 1$, we have*

$$\min S_{\mathrm{sw}}^{(k)}(n+m) \geq \min S_{\mathrm{sw}}^{(k)}(n) + \min S_{\mathrm{sw}}^{(k)}(m).$$

Proof. Let w be a word of length $n+m$ achieving minimal subword entropy $\min S_{\mathrm{sw}}^{(k)}(n+m)$. We write $w = w' \cdot w''$, with $|w'| = n$ and $|w''| = m$. Let v' (resp. v'') be a most frequent subword of w' (resp. w''). We have

$$\operatorname{maxocc}(w) \geq \operatorname{occ}(w' \cdot w'', v' \cdot v'') \geq \operatorname{occ}(w', v')\operatorname{occ}(w'', v'') = \operatorname{maxocc}(w')\operatorname{maxocc}(w'').$$

The first inequality is from the definition of maxocc, the second from Lemma 3.1, and the equality comes from the definition of v' and v''. By the definition of w, we have

$$\min S_{\mathrm{sw}}^{(k)}(n+m) \geq \log_2 \operatorname{maxocc}(w') + \log_2 \operatorname{maxocc}(w'') \geq \min S_{\mathrm{sw}}^{(k)}(n) + \min S_{\mathrm{sw}}^{(k)}(m).$$

The second inequality is from the definition of $\min S_{\mathrm{sw}}^{(k)}$. ◀

▶ **Theorem 3.8.** *For any $k \geq 2$, the sequence $(\min S_{\mathrm{sw}}^{(k)}(n)/n)_{n\geq 1}$ converges to a certain limit $L_k < +\infty$.*

Proof. Proposition 3.7 shows that $\min S_{\mathrm{sw}}^{(k)}(n)$ is super-additive. We then apply Lemma 3.6, and as $\min S_{\mathrm{sw}}^{(k)}(n)/n$ is bounded above by some constant according to Corollary 3.5, we have the existence of the limit L_k which is finite. ◀

With the existence of the limit L_k, we can use known values of $\min S_{\mathrm{sw}}^{(k)}(n)$ to give lower bounds for L_k.

▶ **Proposition 3.9.** *Given $k \geq 2$, we have $L_k \geq \min S_{\mathrm{sw}}^{(k)}(n)/n$ for all n.*

Proof. By iterating Proposition 3.7, we have $\min S_{\mathrm{sw}}^{(k)}(rn) \geq r \min S_{\mathrm{sw}}^{(k)}(n)$ for all $r \geq 1$. Diving both sides by rn, it means that the limit L_k of $\min S_{\mathrm{sw}}^{(k)}(rn)/rn$ is also larger than $\min S_{\mathrm{sw}}^{(k)}(n)$. ◀

From Corollary 3.5, we know that

$$\log_2(1 + k^{-1}) \leq L_k \leq 1.$$

When $k \to \infty$, the lower bound is asymptotically $(\ln 2)^{-1}k^{-1}$, which tends to 0, while the upper bound stays constant. The next natural step is to try to give better bounds for L_k, and eventually compute the precise value of L_k. However, it seems to be a formidable task.

4 Better upper bound for binary alphabet

After the general basic results given in Section 3, we will focus hereinafter on the case of binary alphabet $\mathcal{A} = \{0, 1\}$. In this case, the bounds in Corollary 3.5 become $\log_2(3/2) \leq L_2 \leq 1$ for the limit L_2 in Theorem 3.8. The gap between the two bounds are significant, as $\log_2(3/2) \approx 0.585$. We now give a better upper bound of L_2 by constructing a family of periodic words with a small value of maximal subword occurrences.

▶ **Proposition 4.1.** *For $w = (0011)^m$, there is a most frequent subword w' of the form $(01)^r$.*

Proof. Take a most frequent subword u of w of length ℓ. Suppose that u has the form $u = s \cdot 00 \cdot t$. We take $u^{(1)} = s \cdot 010 \cdot t$ and $u^{(2)} = s \cdot 0 \cdot t$. Let $P = \{p_1, \ldots, p_\ell\}$ be an occurrence of u in w, and we suppose that the 00 occurs at p_i, p_{i+1}. Let $\mathcal{P}$ be the set of occurrences of u in w, which is divided into $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$, where $\mathcal{P}_1$ contains those P's with $p_i + 1 \neq p_{i+1}$, while $\mathcal{P}_2$ contains those with $p_i + 1 = p_{i+1}$. For any $P \in \mathcal{P}_1$, the two 0's occur in different runs, meaning that there is at least one run 11 in between. This leads to at least two choices for the extra 1 added in $u^{(1)}$. Therefore, $\mathrm{occ}(w, u^{(1)}) \geq 2|\mathcal{P}_1|$. For any $P \in \mathcal{P}_2$, the two 0's occur in the same run, meaning that replacing them by a single 0 leaves us two choices. We thus have $\mathrm{occ}(w, u^{(2)}) \geq 2|\mathcal{P}_2|$, meaning that $\mathrm{occ}(w, u^{(1)}) + \mathrm{occ}(w, u^{(2)}) \geq 2|\mathcal{P}| = 2\,\mathrm{maxocc}(w)$. We deduce that at least one of the $u^{(j)}$'s satisfies $\mathrm{occ}(w, u^{(j)}) = \mathrm{maxocc}(w)$. We observe that both $u^{(1)}$ and $u^{(2)}$ have one less pair of identical consecutive letters than u. We may then do the same for consecutive 1's. By iterating such a process, we get a most frequent subword without identical consecutive letters, thus alternating between 0 and 1. Then we conclude by Lemma 3.2. ◀

▶ Remark 4.2. We want to highlight the importance of Proposition 4.1 here. The main difficulty in the study of maximal subword occurrences is, in a sense, algorithmic. To the author's knowledge, we don't know whether there is a polynomial time algorithm to compute a most frequent subword of a given word, or to decide whether there is a subword that occurs at least a given number of times. However, in the case of words of the form $w = (0011)^m$, we manage to show some structure of their most frequent subwords, which then allows us to compute $\mathrm{maxocc}(w)$.

Let $a_{m,r} = \mathrm{maxocc}((0011)^m, (01)^r)$, and $f(x, y) = \sum_{m,r \geq 0} a_{m,r} x^m y^r$ be their generating function. We have the following counting result.

▶ **Proposition 4.3.** *We have*

$$f(x, y) = \frac{1 - x}{(1 - x)^2 - 4xy}, \quad a_{m,r} = 4^r \binom{m + r}{m - r}.$$

Proof. For an occurrence $P = \{p_1, \dots, p_{2r}\}$ of $(01)^r$ in $(0011)^m$, we have two cases.

- $p_{2r} < 4(m-1)$, meaning that P is also an occurrence of $(01)^r$ in $(0011)^{m-1}$;
- $p_{2r} \in \{4m-2, 4m-1\}$, meaning that the last letter 1 of $(01)^r$ occurs at the last segment of 0011. As the $(2r-1)$-st letter of $(01)^r$ is 0, we have $p_{2r-1} \in \{4m'+1, 4m'+2\}$ for some $0 \le m' \le m-1$. By removing both p_{2r-1} and p_{2r}, we obtain P', which is an occurrence of $(01)^{r-1}$ in $(0011)^{m'}$. To go back from P' to P given m', we have two choices for both p_{2r} and p_{2r-1}.

We thus have the recurrence for $m \ge 1$ that

$$a_{m,r} = a_{m-1,r} + \sum_{m'=0}^{m-1} 4a_{m',r-1}. \tag{1}$$

Subtracting Equation (1) for $a_{m,r}$ with that for $a_{m-1,r}$, we have

$$a_{m,r} - 2a_{m-1,r} + a_{m-2,r} - 4a_{m-1,r-1} = 0.$$

By the standard symbolic method, and with the initial conditions $a_{m,0} = 1$ and $a_{m,r} = 0$ for $r > m$, we obtain the claimed expression of $f(x,y)$. We can then compute $a_{m,r}$ by simply extracting the coefficient of y^r first, then that of x^m. ◀

▶ **Theorem 4.4.** *There is some constant c_3 such that, for all $n \in \mathbb{N}$, we have*

$$\min S_{\mathrm{sw}}^{(2)}(n) \le \frac{1}{2}\log_2(1+\sqrt{2})n - \frac{1}{2}\log_2 n + c_3.$$

Proof. For the case $n = 4m$, we have

$$\begin{aligned}\min S_{\mathrm{sw}}^{(2)}(4m) \le S_{\mathrm{sw}}((0011)^m) &= \max_{0\le r\le m} \log_2 \operatorname{occ}((0011)^m, (01)^r) \\ &= \frac{1}{\ln 2}\max_{0\le r\le m} \ln\left(4^r\binom{m+r}{m-r}\right).\end{aligned}$$

The first equality comes from Proposition 4.1, and the second from Proposition 4.3. We take $r = \alpha m$ for some fixed α with $0 < \alpha < 1$. Using Stirling's approximation, we have

$$\ln\left(4^r\binom{m+r}{m-r}\right) = s(\alpha)m - \frac{1}{2}\ln m + O(1),$$

where

$$s(\alpha) = \alpha\ln 4 + (1+\alpha)\ln(1+\alpha) - (1-\alpha)\ln(1-\alpha) - 2\alpha\ln(2\alpha).$$

The function $s(\alpha)$ is maximized at $\alpha = 2^{-1/2}$, with value $2\ln(1+\sqrt{2})$. We thus have, for some constant c_3, and in terms of $n = 4m$,

$$\min S_{\mathrm{sw}}^{(2)}(n) \le \frac{\ln(1+\sqrt{2})}{2\ln 2}n - \frac{1}{2\ln 2}\ln n + c_3 - \ln 4.$$

For the case $n = 4m + i$ with $1 \le i \le 3$, let u be a most frequent subword of $w = (0011)^m010$. For an occurrence P of u in w, we take $P' = P \cap \{n-2, n-1, n\}$. Then, $j = |P'|$ can be $0, 1, 2$ or 3. In each case, we define $u^{(j)}$ to be u with the last j letters removed, and there are at most 2 possibilities for P'. We also notice that $P \setminus P'$ is an occurrence of $u^{(j)}$ in $(0011)^m$. We thus have

$$\begin{aligned}\operatorname{maxocc}(w) = \operatorname{occ}(w,u) &= 2\operatorname{occ}((0011)^m, u^{(1)}) + \operatorname{occ}((0011)^m, u^{(2)}) + \operatorname{occ}((0011)^m, u^{(3)}) \\ &\le 4\operatorname{maxocc}((0011)^m).\end{aligned}$$

We conclude by

$$\min S_{\mathrm{sw}}^{(2)}(4m+i) \le S_{\mathrm{sw}}^{(2)}(w) \le \ln 4 + S_{\mathrm{sw}}^{(2)}((0011)^m) = \frac{\ln(1+\sqrt{2})}{2\ln 2} n - \frac{1}{2\ln 2}\ln n + c_3.$$

For the first inequality, we take w' to be the first $(4m+i)$ letters of w, and it is clear that $\mathrm{maxocc}(w') \le \mathrm{maxocc}(w)$, as each occurrence of some subword v' of w' is also one for w. ◀

The asymptotic upper bound of $\min S_{\mathrm{sw}}^{(2)}(n)/n$ given by Theorem 4.4 is $\frac{1}{2}\log_2(1+\sqrt{2}) \approx 0.636\ldots$, which is much better than that in Corollary 3.5. Furthermore, by regarding $(0011)^m$ as a word in a bigger alphabet, we have the following corollary, which also gives a better upper bound than that in Corollary 3.5.

▶ **Corollary 4.5.** *For all $k \ge 2$ and $n \in \mathbb{N}$, with the constant c_3 from Theorem 4.4, we have*

$$\min S_{\mathrm{sw}}^{(k)}(n) \le \frac{1}{2}\log_2(1+\sqrt{2})n - \frac{1}{2}\log_2 n + c_3.$$

There are also other families of words with which we have some knowledge on its most frequent subwords, with results similar to Proposition 4.1. Two of the families we have studied are $(01)^m$ and $(000111)^m$.

▶ **Proposition 4.6.** *For $w = (01)^m$, there is a most frequent subword w' of the form $(01)^r$. Furthermore, $\mathrm{maxocc}((01)^m, (01)^r) = \binom{m+r}{m-r}$, which is maximized asymptotically for $r = \lfloor n/\sqrt{5} \rfloor$, with the asymptotic maximal value $\exp\left(n \ln \frac{3+\sqrt{5}}{2} - \frac{\ln n}{2} + O(1)\right)$.*

▶ **Proposition 4.7.** *For $w = (000111)^m$, there is a most frequent subword w' of the form $(0011)^r$. Furthermore, let $f_{000111}(x,y) = \sum_{m,r\ge 0} \mathrm{occ}((000111)^m, (0011)^r) x^m y^r$, we have*

$$f_{000111}(x,y) = \frac{(1-x)^3}{(1-x)^4 - 9x(1+2x)^2 y}.$$

For $m \to \infty$, the value of $\mathrm{occ}((000111)^m, (0011)^r)$ is asymptotically maximized for $r = \alpha m$, with α an explicit value around $0.6597177\ldots$. The asymptotic maximal value is $\exp\left(\gamma m - \frac{\ln m}{2} + O(1)\right)$, where γ is an explicit value around $2.7182400\ldots$.

While the proof of Proposition 4.6 does not need heavy machinery, the proof of Proposition 4.7 needs the saddle-point estimates of large powers [5, Theorem VIII.8], during which a polynomial equation of degree 5 appears. Fortunately, the needed solution of the said equation has a radical expression, albeit complicated and making the explicit expressions of α and γ in Proposition 4.7 too long to fit here.

While interesting, the upper bounds of L_2 given by Propositions 4.6 and 4.7, which are approximately $0.6942\ldots$ and $0.6536\ldots$ respectively, are worse than the one from Theorem 4.4. It is natural to try to look at other families of periodic words. This is encouraged by the following theorem.

▶ **Theorem 4.8.** *For any words w, v in an alphabet $\mathcal{A}$ of size k, the generating function $f_{w,v}(x,y) = \sum_{m,r\ge 0} \mathrm{occ}(w^m, v^r) x^m y^r$ is rational in x, y.*

Proof. We define $a_{w,v}^{s,t}(m)$ with $1 \le s,t \le m$ to be the number of occurrences $P = \{p_1, \ldots, p_{|v|}\}$ of v in w^m such that $p_1 = s$ and $p_{|v|} = (m-1)|w| + t$. In other words, $a_{w,v}^{s,t}(m)$ counts the occurrences of v in w^m such that the first (resp. last) letter of v occurs in the first (resp. last) copy of w at position s (resp. t). Let $g_{w,v}^{s,t}(x) = \sum_{m\ge 1} a_{w,v}^{s,t}(m) x^{m-1}$. Note the extra -1 in the exponent of x in $g_{w,v}^{s,t}(x)$. We first show that $g_{w,v}^{s,t}(x)$ is rational.

For an occurrence P of v in w^m, some consecutive letters may occur in the same copy of w. We say that such letters form a cluster, and we denote by σ the integer composition of the number of letters in each cluster from left to right. We denote by $\ell(\sigma)$ the length of σ, which is also the number of clusters. We denote the clusters by $v_\sigma^{(1)}, \dots v_\sigma^{(\ell(\sigma))}$, and it is clear that they are obtained by cutting v into pieces whose lengths are the parts of σ. We then have

$$g^{s,t}_{w,v}(x) = a^{s,t}_{w,v}(1) \\ + \sum_{m\geq 2}\sum_{\sigma\vDash|v|} x^{m-1}\binom{m-2}{\ell(\sigma)-2}\left(\sum_{t'=s}^{|w|} a^{s,t'}_{w,v_\sigma^{(1)}}(1)\right)\left(\sum_{s'=1}^{t} a^{s',t}_{w,v_\sigma^{(\ell(\sigma))}}(1)\right)\prod_{i=2}^{\ell(\sigma)-1}\operatorname{occ}(w,v_\sigma^{(i)})).$$

Here, $\sigma \vDash |v|$ means that we go over all integer compositions of $|v|$. The first term is for $m = 1$. For the second term, we simply count all possibilities of how clusters of v appear in w^m with $m \geq 2$ while fixing the first and the last cluster. We observe that each $a^{s',t'}_{w,v_{\sigma^{(i)}}}(1)$ for any s', t', i is a constant, and the same holds for $\operatorname{occ}(w, v_\sigma^{(i)})$. By exchanging the two summations, and observing that $\sum_{m\geq 2}\binom{m-2}{d-2}x^{m-1} = x^{d-1}(1-x)^{-(d-1)}$, we see that $g^{s,t}_{w,v}(x)$ is rational in x with $(1-x)^{|v|-1}$ as denominator, as $\ell(\sigma) \leq |v|$ for $\sigma \vDash |v|$.

Now, for $1 \leq t \leq |w|$, we define $f^{(t)}_{w,v}(x,y) = \sum_{m\geq 1}\sum_{r\geq 1} b^{(t)}_{w,v}(m,r)x^{m-1}y^r$ with $b^{(t)}_{w,v}(m,r)$ counting the number of occurrences $P = \{p_1, \dots p_{|v|r}\}$ of v^r in w^m such that $p_{|v|r} = (m-1)|w|+t$. Again, we note the extra -1 in the exponent of x. We see that $b^t_{w,v}(m,r)$ is defined similarly as $a^{s,t}_{w,v}(m)$, except that we consider subwords of the form v^r, and we do not fix the position of the first letter of v^r in w^m. We thus have $b^{(t)}_{w,v}(m,1) = \sum_{s=1}^{|w|} a^{s,t}_{w,v}(m)$. Now, let P be an occurrence of v^r in w^m counted by $b^{(t)}_{w,v}(m,r)$. By considering the copies of w spanned by the last copy of v, we have

$$f^{(t)}_{w,v}(x,y) = y\sum_{s=1}^{|w|} g^{s,t}_{w,v}(x) + \frac{xy}{1-x}\left(\sum_{t'=1}^{|w|} f^{(t')}_{w,v}(x,y)\right)\left(\sum_{s=1}^{|w|} g^{s,t}_{w,v}(x)\right) \\ + y\sum_{t'=1}^{|w|-1}\left(f^{(t')}_{w,v}(x,y)\sum_{s=t'+1}^{|w|} g^{s,t}_{w,v}(x)\right).$$

Here, the first term is for $r = 1$, and the rest is for $r \geq 2$. There are two cases: either letters in the r-th and the $(r-1)$-st copies of v do not occur in the same copy of w in w^m, or they do. The first case is counted by the second term above, with the factor $(1-x)^{-1}$ for copies of w between the occurrences of the two last copies of v in w^m. The second case is accounted by the third term above, where we have the constraint that the last letter of the $(r-1)$-st copy of v occurs before the first letter of the r-th copy in the same copy of w.

Let $\mathbf{f} = {}^t(f^{(1)}_{w,v}, \dots, f^{(|w|)}_{w,v})$. The equation above can be seen as $\mathbf{Af} = \mathbf{b}$ for some matrix $\mathbf{A} = (A_{i,j})_{1\leq i,j\leq |w|}$ and some row vector $\mathbf{b}$, both with coefficients that are linear in y and rational in x, and with only powers of $(1-x)$ as denominators. We also observe that $A_{i,i}$ is of the form $1 + R(x)y$ with $R(x)$ rational in x, while $A_{i,j}$ for $i \neq j$ is of the form $R(x)y$. Hence, $\mathbf{A}$ is non-singular, and $f^{(i)}_{w,v}$ is rational in x, y for all $1 \leq i \leq |w|$. We conclude by observing that $f_{w,v}(x,y) = \frac{1}{1-x}(1 + x\sum_{t=1}^{|w|} f^{(t)}_{w,v}(x,y))$, with the 1 taking care of the case $m = 0$, then the factor $(1-x)^{-1}$ for the copies of w after the last cluster of v^r. ◀

Therefore, in principle, for any word w and v, we can first compute $f_{w,v}(x,y)$ effectively as in the proof of Theorem 4.8, then use analytic combinatorics in several variables [10, 12] to compute the asymptotically maximal value of $\operatorname{occ}(w^m, v^r)$ for fixed m. Although the computation of $f_{w,v}(x,y)$ would be tedious, it is still feasible in principle. The only problem is that, for w in general, we do not have results like Proposition 4.1 for the structure of most frequent subwords of w^m, meaning that $\mathrm{maxocc}(w^m)$ is not necessarily achieved for subwords of the form v^r.

5 Open questions

Generally, the "minimum of maximums" structure in the definition of $\mathrm{min}S_{\mathrm{sw}}^{(k)}(n)$ makes estimates difficult. Hence, not a lot is known about $\mathrm{min}S_{\mathrm{sw}}^{(k)}(n)$. Intuitively, we have the following conjecture that is surprisingly difficult to tackle.

▶ **Conjecture 5.1.** *For fixed $k \geq 2$, there is a value N such that the function $\mathrm{min}S_{\mathrm{sw}}^{(k)}(n)/n$ is increasing for $n \geq N$.*

We already know experimentally that Conjecture 5.1 is not universally true for all n (see Table 1). This point needs to be addressed in possible proofs.

We are also interested by better bounds of L_2. Inspired by Theorem 4.8, we looked at some families experimentally, and there are some with potential to give a slightly better upper bound of L_2 according to numerical evidence. However, we don't know how to prove the observed structure on most frequent subwords of such periodic words in general, which leads us to the following conjecture.

▶ **Conjecture 5.2.** *For a given word w, there is a word v such that, for all m large enough, there is a most frequent subword of w^m that takes the form $u \cdot v^r \cdot u'$, with u and u' of lengths bounded by $|v|$.*

If Conjecture 5.2 holds, then using arguments similar to those in Theorem 4.4, we can reduce the computation of $\mathrm{maxocc}(w^m)$ to that of maximizing $\mathrm{occ}(w^m, v^r)$ while losing only a multiplicative constant. We can then apply Theorem 4.8 and tools in analytic combinatorics in several variables to obtain a better upper bound for L_k. Again, Conjecture 5.2 seems natural, intuitive and supported by experimental evidence, but we don't see how to settle it.

Another intuitive idea on most frequent subwords of a given word w is that their length should be smaller than $|w|/2$. The reasoning is that longer subwords have more letters, thus more possible occurrences, but this effect only works up till length $|w|/2$. However, even such an intuitive idea, supported by Proposition 3.4, seems difficult to prove.

▶ **Conjecture 5.3.** *For a given word w of length at least 2, there is no most frequent subword of w with length at least $|w|/2$.*

We now present some concrete experimental results, based on which we have other conjectures. We denote by $\overline{w}$ the word obtained from w by switching 0 and 1, and $\overleftarrow{w}$ the reverse of w. By symmetry between the two letters, we have the following simple observation.

▶ **Lemma 5.4.** *For any $w \in \{0,1\}^n$ with $n \geq 0$, we have* $\mathrm{maxocc}(w) = \mathrm{maxocc}(\overline{w}) = \mathrm{maxocc}(\overleftarrow{w})$.

We now give the words achieving minimal subword entropy of length up to 35 in Table 1, up to the symmetries in Lemma 5.4. These results are computed using a program written in C on one core on a local computation server, and it took around 11 days for $n = 35$. The source code can be found at `https://github.com/fwjmath/maxocc-subword`. We do not include the most frequent subwords we find, because there may be several of them for a word, taking up too much space in the table. There are several observations we can draw from Table 1, but few is without exception.

Table 1 Binary words achieving minimal subword entropy of length from 1 to 35. In each equivalent class defined by the symmetries in Lemma 5.4, only one representative is given. Numerical values are rounded to three digits after the decimal point when needed.

n	Words w with lowest $S^{(2)}_{\mathrm{sw}}(w)$	maxocc(w)	$S^{(2)}_{\mathrm{sw}}(w)$	$S^{(2)}_{\mathrm{sw}}(w)/n$	#runs
1	0	1	0	0	1
2	01	1	0	0	2
3	001	2	1	0.333	2
4	0110	2	1	0.25	3
5	01110	3	1.585	0.317	3
6	011001	5	2.322	0.387	4
7	0110001	6	2.585	0.369	4
8	01110001	9	3.170	0.396	4
9	011000110	16	4	0.444	5
10	0110001110	22	4.459	0.446	5
11	01110001110	33	5.044	0.459	5
12	011000111001	52	5.700	0.475	6
13	0111001001110	72	6.170	0.475	7
14	01100010111001	108	6.755	0.482	8
15	011000101110001	162	7.340	0.489	8
16	0111000101110001	252	7.977	0.499	8
17	01100011111000110	390	8.607	0.506	7
18	011100100101110001	588	9.200	0.511	10
19	0110001011101000110 0110001110110001110	900	9.814	0.517	11
20	01110001011011000110	1320	10.366	0.518	11
21	011100011011010001110	2049	11.000	0.524	11
22	0110001110101000111001	2958	11.530	0.524	12
23	01110001011011010001110	4473	12.127	0.527	13
24	011000111010101000111001	6979	12.769	0.532	14
25	0111000101101101000111001	10602	13.372	0.535	14
26	01110001011011001000111001	15962	13.962	0.537	14
27	011100010101110101000111001	24150	14.560	0.539	16
28	0110001111010010010111000110 0111000101110101000101110001	36450	15.154	0.541	15 16
29	01100011101010001010111000110	53671	15.712	0.542	17
30	011000111001100010101111000110	83862	16.356	0.545	15
31	0110001110101000101011110001110	127998	16.966	0.547	17
32	01100011101010001010111010001110	189131	17.529	0.548	19
33	011000111101010001011011010001110	288900	18.140	0.550	19
34	0110001110101000101011101001001110	442386	18.755	0.552	21
35	01110001011011001000110111001001110	681966	19.379	0.554	19

- The words of length n achieving $\min S_{\mathrm{sw}}^{(2)}(n)$ are palindromic, *i.e.*, $w = \overleftarrow{w}$, or anti-palindromic, *i.e.*, $\overline{w} = \overleftarrow{w}$, for many values of n, such as 1, 2, 4, 5, 6, 8, 9, 11, 12, 13, 14, 16, 17, 22, 23, 24, 29. Moreover, for $n = 19$ (resp. $n = 28$), one of the two words is palindromic (resp. anti-palindromic), the other not.
- The value of $\min S_{\mathrm{sw}}^{(2)}(n)/n$ increases with n in general, but with the exceptions of $n = 3, 4$, $n = 6, 7$ and $n = 12, 13$ (although the rounded numbers are the same). We believe that the exceptions are due to the effect of small size, and should not reproduce for larger n.
- The number of runs for words of length n achieving $\min S_{\mathrm{sw}}^{(2)}(n)$ is increasing with n, with the exception of $n = 17, 28, 30, 35$. Moreover, for $n = 28$ only one word among the two that has less runs than the word for $n = 27$.
- The maximal run length for words of length n achieving $\min S_{\mathrm{sw}}^{(2)}(n)$ is at most 3, with the exception of $n = 17, 28, 30, 31, 33$. Moreover, for $n = 28$, one of the two words has maximal run length 3, and the other 4.
- There is only one word of length n up to symmetries in Lemma 5.4 that achieves $\min S_{\mathrm{sw}}^{(2)}(n)$, with the exception of $n = 19, 28$, where there are two such words.

However, we should note that we only have very limited data, as we were only able to perform exhaustive search for small values of n. A naïve method requires looking at $\Theta(4^n)$ word-subword pairs. Although some optimizations are possible, such as using Lemma 5.4 to reduce the number of words to examine, the time taken remains exponential, against which we cannot push too far. An evidence is that, although asymptotically $\min S_{\mathrm{sw}}^{(2)}(n)/n$ should be bounded from below by $\log_2(3/2) \approx 0.585$ by Corollary 3.5, all the values of $\min S_{\mathrm{sw}}^{(2)}(n)/n$ in Table 1 are smaller than this asymptotic bound, meaning that the values of n tested here are not large enough. Nevertheless, we can still formulate reasonable conjectures based on these observations.

▶ **Conjecture 5.5.** *For $k \geq 2$, let w be a word of length $n \geq 1$ achieving the minimal subword entropy $\min S_{\mathrm{sw}}^{(k)}(n)$. Then, except for a finite number of n, the longest run in w has length 3. Furthermore, the average run length converges when $n \to +\infty$.*

Given Proposition 3.9, we may be tempted to use experimental results to give better lower bound for L_2. However, all such bounds are worse than the one in Corollary 3.5, which is around 0.585 for $k = 2$. Judging from their gap, it seems impractical or even impossible to obtain a better bound in this way. In fact, with examples obtained from searches using various heuristics, it seems that $\mathrm{maxocc}(w)$ for w of length n achieving the lowest subword entropy has an exponential growth in n with a growth constant close to but slightly larger than 1.5, which is the value given by the lower bound. We thus have the following conjecture.

▶ **Conjecture 5.6.** *We have $L_2 > \log_2(3/2)$.*

The value in Conjecture 5.6 comes from the lower bound in Corollary 3.5, which is in fact the expectation of the number of occurrences of a random subword of length $n/3$. Hence, Conjecture 5.6 implies that, for all large values of n, there are binary words of length n in which each subword of length $n/3$ occurs much more often than others. The question remains on how to find such subwords, which probably have relatively high self-correlations.

References

1 A. Burstein, P. Hästö, and T. Mansour. Packing Patterns into Words. *Eletron. J. Combin.*, 9(2), 2003. doi:10.37236/1692.

2 A. Burstein and T. Mansour. Counting occurrences of some subword patterns. *Discrete Mathematics & Theoretical Computer Science*, Vol. 6 no. 1, January 2003. doi:10.46298/dmtcs.320.

3 Wenjie Fang. fwjmath/maxocc-subword. Software, swhId: swh:1:dir:fef689a6896632f63f67b460e989fc106d5899e0 (visited on 2024-07-05). URL: https://github.com/fwjmath/maxocc-subword.

4 M. Fekete. Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Math. Z.*, 17(1):228–249, 1923. doi:10.1007/bf01504345.

5 Ph. Flajolet and R. Sedgewick. *Analytic combinatorics.* Cambridge University Press, Cambridge, 2009. doi:10.1017/CBO9780511801655.

6 Ph. Flajolet, W. Szpankowski, and B. Vallée. Hidden word statistics. *Journal of the ACM*, 53(1):147–183, 2006. doi:10.1145/1120582.1120586.

7 I. Gheorghiciuc and M. D. Ward. On Correlation Polynomials and Subword Complexity. *Discrete Mathematics & Theoretical Computer Science*, DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 2007), 2007. doi:10.46298/dmtcs.3553.

8 K. Iwanuma, R. Ishihara, Y. Takano, and H. Nabeshima. Extracting Frequent Subsequences from a Single Long Data Sequence: A Novel Anti-Monotonic Measure and a Simple On-Line Algorithm. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*. IEEE, 2005. doi:10.1109/icdm.2005.60.

9 S. Kitaev. *Patterns in Permutations and Words.* Springer Berlin Heidelberg, 2011. doi:10.1007/978-3-642-17333-2.

10 S. Melczer. *Algorithmic and Symbolic Combinatorics: An Invitation to Analytic Combinatorics in Several Variables.* Springer International Publishing, 2021. doi:10.1007/978-3-030-67080-1.

11 K. Menon and A. Singh. Subsequence frequency in binary words. *Discrete Mathematics*, 347(5):113928, May 2024. doi:10.1016/j.disc.2024.113928.

12 M. Mishna. *Analytic combinatorics: a multidimensional approach.* Discrete Mathematics and its Applications (Boca Raton). CRC Press, 2020.

13 M. Morse and G. A. Hedlund. Symbolic dynamics. *Amer. J. Math.*, 60(4):815, October 1938. doi:10.2307/2371264.

14 V. Vatter. Permutation classes. In *Handbook of Enumerative Combinatorics.* CRC Press, 2015.

15 G. Yang. The complexity of mining maximal frequent itemsets and maximal frequent patterns. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD04. ACM, August 2004. doi:10.1145/1014052.1014091.