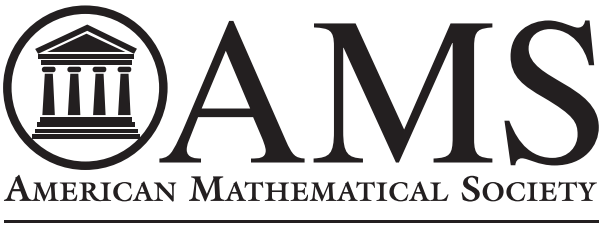
A characterization of words of linear complexity

Julien Cassaigne, Anna Frid, Svetlana Puzynina, Luca Q. Zamboni



Julien Cassaigne, Anna E. Frid, Svetlana Puzynina, Luca Q. Zamboni

A characterization of words of linear complexity

Proceedings of the American Mathematical Society

DOI: 10.1090/proc/14440

Accepted Manuscript

A CHARACTERIZATION OF WORDS OF LINEAR COMPLEXITY

JULIEN CASSAIGNE, ANNA E. FRID, SVETLANA PUZYNINA, AND LUCA Q. ZAMBONI

Abstract. Given an infinite word $x = x_0x_1x_2\cdots \in \mathbb{A}^{\mathbb{N}}$ over some finite alphabet $\mathbb{A}$, the *factor complexity* $p_x(n)$ counts the number of distinct factors of x of each given length n, i.e., the number of distinct blocks $x_ix_{i+1}\cdots x_{i+n-1} \in \mathbb{A}^n$ occurring in x. The factor complexity provides a useful measure of the extent of randomness of x : Periodic words have bounded factor complexity while digit expansions of normal numbers have maximal complexity. In this paper we obtain a new characterization of infinite words x of sub-linear complexity, namely we show that $p_x(n) = O(n)$ if and only if there exists a set $S \subseteq \mathbb{A}^*$ of bounded complexity (meaning $\limsup p_S(n) < +\infty$) such that each factor w of x is a concatenation of two elements of S, i.e., $w = uv$ with $u, v \in S$. In the process we introduce the notions of *marker words* and *marker sets* which are both new and may be of independent interest. Marker sets defined by right special factors constitute the key tool needed to split each factor of an infinite word of linear complexity into two pieces.

1. Introduction

Let $\mathbb{A}$ be a finite non-empty set. For each infinite word $x = x_0x_1x_2\cdots \in \mathbb{A}^{\mathbb{N}}$, the complexity or *factor complexity* $p_x(n)$ counts the number of distinct blocks $x_ix_{i+1}\cdots x_{i+n-1} \in \mathbb{A}^n$ of length n occurring in x. In other words, the complexity of x is taken to be the complexity of the language of its factors $\mathrm{Fac}(x) = \{x_ix_{i+1}\cdots x_j \,|\, 0 \le i \le j\} \cup \{\varepsilon\}$. First introduced by Hedlund and Morse in their seminal 1938 paper [15] under the name of *block growth*,[1] the factor complexity provides a useful measure of the extent of randomness of x and more generally of the subshift it generates. Periodic words have bounded factor complexity while digit expansions of normal numbers have maximal complexity. A celebrated theorem of Morse and Hedlund in [15] states that every aperiodic (meaning not ultimately periodic) word contains at least $n+1$ distinct factors of each length n. Sturmian words are those aperiodic words of minimal factor complexity: $p_x(n) = n+1$ for each $n \ge 1$. In [16] Hedlund and Morse showed that each Sturmian word may be realized geometrically by an irrational rotation on the circle. More precisely, every Sturmian word is obtained by coding the symbolic orbit of a point x on the circle (of circumference one) under a rotation by an irrational angle α where the circle is partitioned into two complementary intervals, one of length α and the other of length $1-\alpha$. And conversely each such coding gives rise to a Sturmian word.

Date: March 27, 2018.

2000 *Mathematics Subject Classification.* Primary 68R15; Secondary 37B10.

Key words and phrases. factor complexity, Sturmian words, linear complexity.

Svetlana Puzynina is partially supported by Russian Foundation of Basic Research (grant 18-31-00118).

[1] In [9], Ehrenfeucht, Lee, and Rozenberg adopted the term *subword complexity*.

Sturmian words admit various other types of characterizations of geometric and combinatorial nature (see for instance [13]) and constitute one important class of infinite words of linear complexity. In addition to Sturmian words, there are many other examples of families of infinite words of linear complexity including codings of rotations [19], interval exchange transformations [2], Arnoux-Rauzy sequences [3], episturmian sequences [10], primitive substitution sequences [18] and automatic sequences [1]. Results on the complexity of words are generally one of two kinds: Either they provide conditions or formulae for the complexity of a given family of words, for instance Pansiot's work in [17] on the classification of the factor complexities of purely morphic words. Or they give conditions on words, or rules for generating them, subject to specified constraints on their complexity. An example of a deep and difficult problem of this kind is the so-called S-adic conjecture on words of linear complexity (see for instance [11] and the references therein).

The set $\mathbb{A}^*$ consisting of all finite words over the alphabet $\mathbb{A}$ is naturally a free monoid under the operation of concatenation, with the empty word ε playing the role of the identity. Given an infinite word $x \in \mathbb{A}^{\mathbb{N}}$ one may ask whether the set of factors $\mathrm{Fac}(x)$ is contained in a finite product of the form S^k where S is a subset of $\mathbb{A}^*$ of strictly lower complexity.

Example 1.1. Consider the *Thue-Morse infinite word*

$$x = 01101001100101101001 0 \cdots$$

where for each $n \geq 0$, the nth term x_n is defined as the sum modulo 2 of the digits in the binary expansion of n. The origins of this word date back to the beginning of the last century with the works of A. Thue [20, 21] in which he proves amongst other things that x is *overlap-free*, i.e., contains no word of the form uuu' where u' is a non-empty prefix of u. It is well known that x is also a fixed point of the substitution $\varphi : 0 \to 01, 1 \to 10$. The factor complexity of the Thue-Morse word, first precisely computed by Brlek [4] and independently by de Luca and Varricchio [8], oscillates between two linear functions: $3(n-1) \leq p_x(n) \leq \frac{10}{3}(n-1)$, with each bound attained an infinite number of times.

For each $m \geq 0$, let $t_m = \varphi^m(0)$ and $\bar{t}_m = \varphi^m(1)$. Then both t_m and $\bar{t}_m$ are factors of x of length 2^m. Let $S \subseteq \{0,1\}^*$ be the set consisting of all prefixes and suffixes (including ε) of t_m and $\bar{t}_m$ for each $m \geq 0$. Since $t_{m+1} = \varphi^{m+1}(0) = \varphi^m(01) = t_m\bar{t}_m$ and similarly $\bar{t}_{m+1} = \bar{t}_m t_m$, it follows that S contains at most 4 words of each length n. We claim that $\mathrm{Fac}(x) \subseteq S^2$. To see this, let $u \in \mathrm{Fac}(x)$. Since S contains ε, 0 and 1, we may suppose $|u| \geq 2$. Consider the least $m \geq 0$ such that u is a factor of t_{m+1} or a factor of $\bar{t}_{m+1}$. If u is a factor of t_{m+1}, by minimality of m we can write $u = vw$ where v is a non-empty suffix of t_m and w a non-empty prefix of $\bar{t}_m$. Whence $u \in S^2$. A similar argument applies in case u is a factor of $\bar{t}_{m+1}$. Thus while $\mathrm{Fac}(x)$ is of linear complexity, it is contained in a product S^2 where S is of complexity bounded by 4.

The above example is an illustration of the following more general result proved herein:

Theorem 1.2. *An infinite word $x \in \mathbb{A}^{\mathbb{N}}$ is of sub-linear complexity (i.e., $p_x(n) = O(n)$) if and only if $\mathrm{Fac}(x) \subseteq S^2$ for some $S \subseteq \mathbb{A}^*$ of bounded complexity (i.e., $\limsup p_S(n) < +\infty$).*

Theorem 1.2 is actually a consequence of a somewhat more general result (see Theorem 3.4) combined with an earlier result of the first author which gives a uniform bound on the number of right special factors of each length n of an infinite word word of linear complexity. In order to construct a set S of bounded complexity satisfying the condition in Theorem 1.2, we introduce the notions of *marker words* and *marker sets* which are both new and may be of independent interest. Marker sets defined by right special factors constitute the key tool needed to split each factor of an infinite word of linear complexity into two pieces.

Theorem 1.2 does not extend to arbitrary languages of sub-linear complexity. We give an example of a (non-factorial) language L of complexity $p_L(n) = O(\log n)$ which is not contained in any finite product of the form S^k where $S \subseteq \mathbb{A}^*$ is of bounded complexity.

2. Preliminaries

In this section we recall some basic definitions and notations concerning finite and infinite words which are relevant to the subsequent sections. For more details the reader is referred to one of the standard texts in combinatorics on words such as the Lothaire books [12, 13, 14]. For the sake of clarity and self-containment, we develop in detail some notions which are less mainstream in the area of combinatorics on words but which will be relevant in the proofs of the main results. They include the notions of internal and extremal occurrences of factors in both finite and infinite words which are defined in terms of virtual occurrences.

Let $\mathbb{A}$ be a finite non-empty set (the *alphabet*). Let $\mathbb{A}^*$ denote the set of all finite words $u = u_0u_1 \cdots u_{n-1}$ with $u_i \in \mathbb{A}$. We call n the *length* of u and denote it $|u|$. The empty word is denoted ε, and by convention $|\varepsilon| = 0$. We put $\mathbb{A}^+ = \mathbb{A}^* \setminus \{\varepsilon\}$. A subset $L \subseteq \mathbb{A}^*$ is called a *language*.

Let $\mathbb{A}^{\mathbb{N}}$ denote the set of all right infinite words $x = x_0x_1x_2 \cdots$ with $x_i \in \mathbb{A}$. Given $x = x_0x_1x_2 \cdots \in \mathbb{A}^* \cup \mathbb{A}^{\mathbb{N}}$ let $\mathrm{Fac}(x) = \{x_i \cdots x_{i+n} \,|\, i, n \geq 0\} \cup \{\varepsilon\}$ denote the set of factors of x. We will frequently use the notation $x[i, j]$ for $x_i \cdots x_j$.

A language L is said to be *factorial* if $\mathrm{Fac}(u) \subseteq L$ for each $u \in L$.

Given a language $L \subseteq \mathbb{A}^*$, we define its complexity $p_L : \mathbb{N} \to \mathbb{N}$ by

$$p_L(n) = \mathrm{Card}(L \cap \mathbb{A}^n).$$

The complexity of the set of factors of a word x is its *factor complexity*

$$p_x(n) = \mathrm{Card}(\mathrm{Fac}(x) \cap \mathbb{A}^n).$$

We say that $x \in \mathbb{A}^{\mathbb{N}}$ (resp., $L \subseteq \mathbb{A}^*$) is of *bounded complexity* if there exists a positive integer C such that $p_x(n) \leq C$ (resp., $p_L(n) \leq C$) for all $n \in \mathbb{N}$. An infinite word x is called *ultimately periodic*, or *ultimately* $|v|$*-periodic*, if $x = uvvv \cdots = uv^\omega$ for some words $u \in \mathbb{A}^*$ and $v \in \mathbb{A}^+$. An infinite word is said to be *aperiodic* if it is not ultimately periodic. A factor u of x is called *right* (resp., *left*) *special* if $ua, ub \in \mathrm{Fac}(x)$ (resp., $au, bu \in \mathrm{Fac}(x)$) for distinct letters $a, b \in \mathbb{A}$. Every aperiodic word contains a right and a left special factor of each length. An infinite word x is said to be *recurrent* if each prefix of x occurs infinitely often in x.

Analogously we can consider bi-infinite words indexed by $\mathbb{Z}$. The definitions above extend in the obvious ways. In particular, a bi-infinite word x is said to be eventually right (left) periodic if x admits a a suffix of the form $vvv \cdots$ (respectively, a prefix of the form $\cdots vvv$) for some $v \in \mathbb{A}^+$. Otherwise x is said to be right (or left) aperiodic.

Definition 2.1. Let $u = u_0u_1\cdots u_{n-1}$ and v belong to $\mathbb{A}^+$. Fix an integer i, $0 \le i \le n$. We say that there is a *virtual occurrence* of v in u beginning (ending, respectively) at position i if the shorter of v and $u_i\cdots u_{n-1}$ ($u_0\cdots u_{i-1}$, respectively) is a prefix (suffix) of the other. That is, $v\mathbb{A}^*\cap u[i,n-1]\mathbb{A}^* \neq \emptyset$ ($\mathbb{A}^*v\cap\mathbb{A}^*u[0,i-1] \neq \emptyset$, respectively).

Definition 2.2. For $u = u_0u_1\cdots u_{n-1}$ and $0 \le i \le n$, we say that u has a *virtual square* centered at position i if there exists a word $v \in \mathbb{A}^+$ (the *witness*) and a virtual occurrence of v in u both beginning and ending at position i.

For example, the word $u = 00101101$ has a virtual square of length 2 centered at position $i = 3$ (witnessed by $v = 01$) as well as a virtual square of length 3 centered at position $i = 7$ (witnessed by $v = 110$.)

The above definitions extend in the obvious way to define a virtual occurrence of a word $v \in \mathbb{A}^+$ beginning or ending at a position $i \ge 0$ in an infinite word $x = x_0x_1\cdots$. In this way we can talk about virtual squares occurring in an infinite word. For instance, the word $x = 0100101001001010010\cdots$ has virtual squares of length 2 and 3 centered at position 1, and of lengths 3 and 5 at position 2.

Definition 2.3. Let $v = v_0v_1\cdots v_{n-1} \in \mathbb{A}^+$. Define the (least) *period* of v, denoted $\pi(v)$, to be the least positive integer m such that $v_i = v_{i+m}$ for all $0 \le i \le n-1-m$.

For instance, for $v = 00110$ we have $\pi(v) = 4$ while for $v = 00101101$ we have $\pi(v) = 8 = |v|$. Clearly in general $\pi(v) \le |v|$.

Let $x \in \mathbb{A}^+ \cup \mathbb{A}^{\mathbb{N}}$ be a finite or infinite word, and let $v \in \mathbb{A}^+$ be a word of length n occurring in x at a position $i \ge 0$, meaning $v = x[i, i+n-1]$. We say that the occurrence of v at position i is *internal* if x has virtual squares of length $\pi(v)$ centered at positions i and $i+n$. An occurrence of v in x which is not internal is called *extremal*. More precisely, an extremal occurrence is called *initial* if x does not have a virtual square of length $\pi(v)$ centered at position i, and *final* if x does not have a virtual square of length $\pi(v)$ centered at position $i+n$. For instance, if $x = 01001010100\cdots$, then the occurrence of $v = 010$ at position 0 is not initial: in fact, despite the name, the prefix occurrence of a factor is never initial since there is always a virtual square centered at position 0. Instead the prefix occurrence of v is final (even if it is immediately followed by another occurrence of v) since x does not have a virtual square of length 2 centered at position 3. On the other hand, the occurrence of v at position 3 is initial since x does not have a virtual square of length $2 = \pi(v)$ centered at position 3. In contrast, the occurrence of v in position 5 is internal. Note that an occurrence of a word v in x can be both initial and final. We also note that if x is an aperiodic infinite word, then each factor v of x admits a final occurrence in x.

3. Marker sets and words of sub-linear complexity

In this section, we prove the main result of this paper. For $x \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^{\mathbb{Z}}$, and $n \ge 0$, let $\mathcal{R}_x(n)$ denote the set of right special factors of x of length n and $\mathcal{R}_x = \bigcup_{n\ge 0}\mathcal{R}_x(n)$.

Definition 3.1. Let D be a positive integer. A subset $M \subseteq \mathbb{A}^*$ is called a *D-marker set* for x if for each $n \ge 1$ and each factor u of x of length $|u| \ge Dn$ we have $\mathrm{Fac}(u) \cap M \cap \mathbb{A}^n \neq \emptyset$. The elements of M are called D-markers.

Lemma 3.2. *Let C be a positive integer. Then for each aperiodic word $x \in \mathbb{A}^{\mathbb{N}}$ or right aperiodic $x \in \mathbb{A}^{\mathbb{Z}}$ with $p_x(n) \leq Cn$ for each $n \geq 1$, the set $\mathcal{R}_x$ is a $(C+1)$-marker set for x.*

Proof. Fix a positive integer n, and let u be any factor of x of length $(C+1)n$. We show that u contains some element of $\mathcal{R}_x(n)$. Since $p_x(n) \leq Cn$, and there are $Cn+1$ positions for factors of length n in u, by the pigeon-hole principle there exists a factor v of x of length n which occurs in u at least twice. Thus u contains as a factor a word w of length $|w| > n$ which begins and ends in v. Hence there exists a prefix w' of w of length $|w'| \geq n$ which is a right special factor of x. Otherwise, every occurrence of v in x is an occurrence of w, whence x is ultimately (right) periodic, a contradiction. It follows that the suffix w'' of w' of length n belongs to $\mathcal{R}_x(n)$. □

What is important to know about right special words is that in an infinite word of linear complexity, their number of each length is bounded ([5], see also [7]).

The proof of the following proposition describes a more general method for constructing marker sets whose complexity is related to the complexity of the underlying word:

Proposition 3.3. *For each aperiodic word $x \in \mathbb{A}^{\mathbb{N}}$ or right aperiodic $x \in \mathbb{A}^{\mathbb{Z}}$ there exists a 3-marker set M for x with*

$$p_M(n) \leq \frac{p_x(4n)}{n}$$

for each $n \geq 1$.

Proof. For each $n \geq 1$, we build recursively (relative to the index i) sets $M_n(i)$ consisting of factors of x of length n, and $W_n(i)$ consisting factors of x of length $3n$. In each case $\mathrm{Card}(M_n(i)) = \mathrm{Card}(W_n(i)) \leq i$. The process terminates when each factor of x of length $3n$ contains a factor from $M_n(i)$. Starting with $M_n(0)$ and $W_n(0)$ both empty, let w_1 be the factor of x of length $3n$ beginning in position n, and let m_1 be the middle block of w_1 of length n, i.e., $w_1 = x[n, 4n-1]$ and $m_1 = w_1[n, 2n-1] = x[2n, 3n-1]$. Then set $W_n(1) = \{w_1\}$ and $M_n(1) = \{m_1\}$.

For the inductive step, fix $i \geq 1$ and suppose we have constructed sets $M_n(i)$ and $W_n(i)$ as required. Consider the factors of x of length $3n$. If each of them contains a factor from $M_n(i)$, then we are done and we set $M_n = M_n(i)$, $W_n = W_n(i)$. Otherwise, pick a factor w_{i+1} of x of length $3n$ not containing any element of $M_n(i)$ and set $W_n(i+1) = W_n(i) \cup \{w_{i+1}\}$ and $M_n(i+1) = M_n(i) \cup \{m_{i+1}\}$ where m_{i+1} is the middle block of w_{i+1} of length n. Note that if x is a one-sided infinite word, then $w_{i+1} = x[m, m+3n-1]$ where $m \geq n$. Since all w_i are distinct and there are a finite number of factors of x of length $3n$, this process terminates at some point $i \geq 1$. Finally, we set $M = \cup_{n \geq 1} M_n$. By construction, M is a 3-marker set. It remains to prove the upper bound on the complexity of M.

For each element w_i of W_n, we consider a final occurrence $w_i = x[k_i, k_i + 3n - 1]$ of w_i in x. Since x is aperiodic (or bi-infinite and right aperiodic), each factor of x admits at least one final occurrence in x. Now for each $j = 0, \ldots, n-1$ consider its *covering factor* $c(i,j) = x[k_i + j - n, k_i + 3n + j - 1]$. Then the length of $c(i,j)$ is $4n$ and $w_i = c(i,j)[n-j, 4n-j-1]$. Note that even if x is one-sided infinite, each $c(i,j)$ is well defined since each w_i occurs in x at a position n or greater.

Now let us prove that if $c(i,j) = c(i',j')$, then $i = i'$ and $j = j'$. Indeed, suppose that $c(i,j) = c(i',j')$ but $i' < i$. Then $w_i = c(i,j)[n-j, 4n-j-1]$. Analogously, $w_{i'} = c(i,j)[n-j', 4n-j'-1]$ and thus $m_{i'} = c(i,j)[2n-j', 3n-j'-1]$. But since $j, j' \in \{0, \dots, n-1\}$, we have $2n - j' \geq n - j$ and $3n - j' \leq 4n - j$. So, $m_{i'}$ is a factor of w_i, a contradiction to our definition of w_i. We have proved that $i = i'$.

Next suppose that $j' < j$. Then $w_i = c(i,j)[n-j, 4n-j-1] = c(i,j)[n-j', 4n-j'-1]$. Consider the word $s = c(i,j)[n-j, 4n-j'-1]$. It is $(j-j')$-periodic, and in particular, its prefix w_i is $(j-j')$-periodic. So, $\pi(w_i) \leq j - j' \leq n$. The prefix occurrence of w_i to s overlaps with the suffix occurrence of w_i to s by $3n-(j-j') \geq 2n > \pi(w_i)$ symbols, and thus s is also $\pi(w_i)$-periodic. In particular, s has a virtual square of length $\pi(w_i)$ at the end of the prefix occurrence of w_i, that is, at the position $3n$. But s is a factor of $c(i,j) = x[k_i + j - n, k_i + 3n + j - 1]$, namely, $s = c(i,j)[n-j, 4n-j'-1] = x[k_i, k_i + 3n + j - j' - 1]$. So, x has an occurrence of w_i (of length $3n$) at position k_i, followed by a virtual square of length $\pi(w_i)$ at position $k_i + 3n$. It means exactly that this occurrence of w_i is not final, a contradiction.

So, $c(i,j) \neq c(i',j')$ for $i \neq i'$ or $j \neq j'$. Thus, the total number of covering factors $c(i,j)$ is given by

$$\mathrm{Card}\left(\{c(i,j) \mid 1 \leq i \leq \mathrm{Card}(W_n),\, 0 \leq j < n\}\right) = n\mathrm{Card}(W_n) = n\mathrm{Card}(M_n).$$

On the other hand, each covering factor $c(i,j)$ is a factor of x of length $4n$ whence their number is bounded above by $p_x(4n)$. Thus

$$p_M(n) = \mathrm{Card}(M_n) \leq \frac{p_x(4n)}{n}$$

as required. □

Theorem 3.4. *Assume either $y \in \mathbb{A}^{\mathbb{N}}$ and y is recurrent, or $y \in \mathbb{A}^{\mathbb{Z}}$. Let D be a positive integer and assume that M is a D-marker set for y. Then there exist $S, T \subseteq \mathbb{A}^*$ such that $\mathrm{Fac}(y) \subseteq ST$ and for each $n \geq 2D$ we have*

$$p_S(n), p_T(n) \leq \sum_{k \in I_n \cap \mathbb{N}} p_M(2^k)\left(1 + \frac{4p_y(3n)}{2^k}\right) \tag{3.1}$$

where $I_n = (\log_2\left(\frac{n}{2D}\right), \log_2(2n)]$.

Proof. Let us fix a D-marker set M for y. For each $k \geq 1$, let $M_k = \{\mathfrak{m} \in M \mid |\mathfrak{m}| = 2^k\}$. The elements of M_k are called markers of order k.

Consider a factor v of y with $|v| \geq 2D$. We shall define a rule for decomposing v as a product $v = s(v)t(v)$. The sets S and T will then be defined as the collection of all $s(v)$ and all $t(v)$ corresponding to all factors v of y of length $|v| \geq 2D$. Let $k \geq 1$ be the largest positive integer such that $\mathrm{Fac}(v) \cap M_k \neq \emptyset$, and fix $\mathfrak{m} \in \mathrm{Fac}(v) \cap M_k$. Thus $\mathfrak{m}$ is a marker word contained in v of length $|\mathfrak{m}| = 2^k$. First suppose some occurrence of $\mathfrak{m}$ in v is extremal. In this case, we arbitrarily pick one such occurrence, say at position j, and cut v precisely in the middle of this extremal occurrence of $\mathfrak{m}$ so that $s(v) = v[0, j+2^{k-1}-1]$ and $t(v) = v[j+2^{k-1}, |v|-1]$. In case all occurrences of $\mathfrak{m}$ in v are internal, then again arbitrarily pick one such internal occurrence, say at position j, and cut v precisely in the middle of this internal occurrence of $\mathfrak{m}$ so that $s(v) = v[0, j + 2^{k-1} - 1]$ and $t(v) = v[j + 2^{k-1}, |v| - 1]$ (see

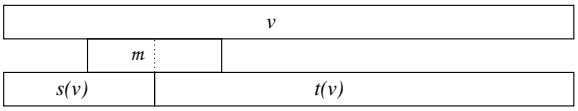


FIGURE 1. Building elements of S and T from a word v and an occurrence of a marker.

Fig. 1). Note that our cutting rule gives preference to extremal occurrences of the marker word.

Now set

$$S = (\mathrm{Fac}(y) \cap \mathbb{A}^{<2D}) \cup \{s(v) \,|\, v \in \mathrm{Fac}(y) \cap \mathbb{A}^{\geq 2D}\},$$
$$T = \{\varepsilon\} \cup \{t(v) \,|\, v \in \mathrm{Fac}(y) \cap \mathbb{A}^{\geq 2D}\},$$

where $\mathbb{A}^{<n} = \bigcup_{k=0}^{n-1} \mathbb{A}^k$ and $\mathbb{A}^{\geq n} = \mathbb{A}^* \setminus \mathbb{A}^{<n}$.

It follows immediately from the definitions that $\mathrm{Fac}(y) \subseteq ST$. It remains to show that complexities of S and T satisfy (3.1). We prove this only for T as the proof for S works in very much the same way.

Fix $n \geq 2D$, and let us estimate $p_T(n)$. Recall that each $u \in T \cap \mathbb{A}^n$ is obtained by cutting some factor v of y in the middle of an occurrence of some marker $\mathfrak{m}$ of maximal order k occurring in v and $u = t(v)$ is the resulting suffix of v. Then since $t(v)$ begins with the suffix of $\mathfrak{m}$ of length $|\mathfrak{m}|/2$, we have $n \geq 2^{k-1}$. On the other hand, since k was chosen to be maximal, we have $n < D2^{k+1}$ for otherwise v, which is of length at least n, would contain a marker of order $k+1$. These inequalities combined give

$$\frac{n}{2D} < 2^k \leq 2n, \tag{3.2}$$

which implies that k lies in the interval $I_n = (\log_2\left(\frac{n}{2D}\right), \log_2(2n)]$. For each such integer $k \in I_n$, the number of marker words of length 2^k is equal to $p_M(2^k)$.

We next prove that each marker word $\mathfrak{m}$ of length 2^k with k, n satisfying (3.2) contributes at most $1 + \frac{4p_y(3n)}{2^k}$ elements to $T \cap \mathbb{A}^n$. Let $T(\mathfrak{m}, n)$ be the set of all $u \in T \cap \mathbb{A}^n$ with $u = t(v)$ for some factor v of y cut at an occurrence of the marker $\mathfrak{m}$ in v. We consider separately the three possible types of occurrences of $\mathfrak{m}$: internal, initial and final. Thus let $T_{\mathrm{int}}(\mathfrak{m}, n)$ (resp., $T_{\mathrm{ini}}(\mathfrak{m}, n)$ and $T_{\mathrm{fin}}(\mathfrak{m}, n)$) be the subset of $T(\mathfrak{m}, n)$ arising from internal (resp., initial and final) occurrences of $\mathfrak{m}$. Recall that if $t \in T_{\mathrm{int}}(\mathfrak{m}, n)$, then $t = t(v)$ for some factor v of y such that every occurrence of $\mathfrak{m}$ in v is internal. This implies that v is $\pi(\mathfrak{m})$-periodic and hence t is uniquely determined by $\mathfrak{m}$ and $|t| = n$. More precisely, t is the word of length n occurring at position 2^{k-1} of the periodic word p^ω, where p is the prefix of $\mathfrak{m}$ of length $\pi(\mathfrak{m})$ (see Fig. 2). Thus $\mathrm{Card}(T_{\mathrm{int}}(\mathfrak{m}, n)) = 1$.

Next we estimate $\mathrm{Card}(T_{\mathrm{ini}}(\mathfrak{m}, n))$.

Lemma 3.5. *For each $n \geq 2D$ we have*

$$\mathrm{Card}(T_{\mathrm{ini}}(\mathfrak{m}, n)) \leq \frac{2p_y(3n)}{2^k}$$

Proof. For $t \in T_{\mathrm{ini}}(\mathfrak{m}, n)$, and each $0 \leq i < 2^{k-1}$, let $E_{\mathrm{ini}}(\mathfrak{m}, n, t, i)$ be the collection of all factors w of y of length $n + 2^k$ such that w has an occurrence of $\mathfrak{m}$ at position i and an occurrence of t in position $i + 2^{k-1}$; if $i \geq \pi(\mathfrak{m})$, we require the occurrence of $\mathfrak{m}$ to v to be initial (see Fig. 3).

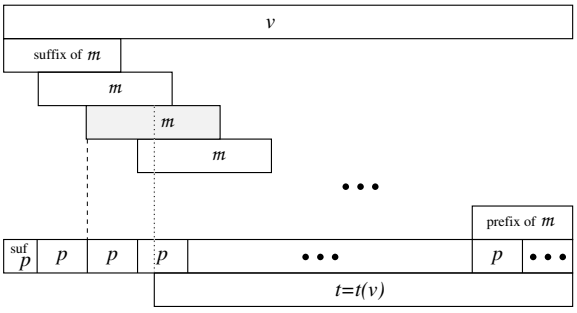


FIGURE 2. The case of an internal occurrence: the unique t is determined by $\mathfrak{m}$ and the length.

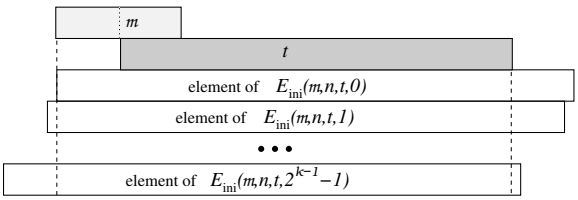


FIGURE 3. The sets $E_{\text{ini}}(\mathfrak{m}, n, t, i)$. The parts between dashed lines are common for all elements.

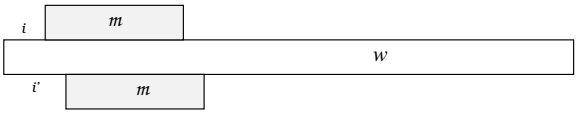


FIGURE 4. Proof of Claim 3.6. The lower occurrence of $\mathfrak{m}$ is not initial

Let v be a factor of y giving rise to t in $T_{\text{ini}}(\mathfrak{m}, n)$, that is, v contains an initial occurrence of $\mathfrak{m}$, and the suffix of v starting in the middle of that occurrence of $\mathfrak{m}$ is t. Since y is assumed either recurrent or bi-infinite, there exists an occurrence of v at the distance more than i from the beginning of the word y. So, $E_{\text{ini}}(\mathfrak{m}, n, t, i)$ is non-empty. Then:

Claim 3.6. *For each $t, t' \in T_{\text{ini}}(\mathfrak{m}, n)$ and $0 \le i, i' < 2^{k-1}$, where $t \ne t'$ or $i \ne i'$, we have*

$$E_{\text{ini}}(\mathfrak{m}, n, t, i) \cap E_{\text{ini}}(\mathfrak{m}, n, t', i') = \emptyset.$$

Proof of Claim 3.6. Suppose $w \in E_{\text{ini}}(\mathfrak{m}, n, t, i) \cap E_{\text{ini}}(\mathfrak{m}, n, t', i')$. First consider the case of $0 \le i < i' < 2^{k-1}$. Then $\mathfrak{m}$ occurs in w in position i and i', and since $i' - i < 2^{k-1} < |\mathfrak{m}|$, it follows that the two occurrences of $\mathfrak{m}$ in w overlap. So, $\mathfrak{m}$ is $(i' - i)$-periodic, and thus $\pi(\mathfrak{m}) \le i' - i < 2^{k-1} < |\mathfrak{m}|/2$. So, $w[i, i' + 2^k - 1]$ is $\pi(\mathfrak{m})$-periodic contradicting that the occurrence of $\mathfrak{m}$ at position $i' > \pi(\mathfrak{m})$ of w was initial (see Fig. 4). So, $i = i'$. But then both t and t' are words of length n occurring in w at position $i + 2^{k-1}$, so, $t = t'$. □

So, each $t \in T_{\text{ini}}(\mathfrak{m}, n)$ and each $i \in \{0, \dots, 2^{k-1} - 1\}$ correspond to at least one factor of y of length $n + 2^k$: the set $E_{\text{ini}}(\mathfrak{m}, n, t, i)$ of all such factors is non-empty,

and for different words t or indices i, these sets do not intersect. So,

$$2^{k-1}\mathrm{Card}(T_{\mathrm{ini}}(\mathfrak{m},n)) \leq \sum_{i=0}^{2^{k-1}-1} \sum_{t\in T_{\mathrm{ini}}(\mathfrak{m},n)} \mathrm{Card}(E_{\mathrm{ini}}(\mathfrak{m},n,t,i)) \leq p_y(n+2^k),$$

and since $n+2^k \leq 3n$ and thus $p_y(n+2^k) \leq p_y(3n)$,

$$\mathrm{Card}(T_{\mathrm{ini}}(\mathfrak{m},n)) \leq \frac{2p_y(3n)}{2^k}$$

as required. □

A similar argument applies to $T_{\mathrm{fin}}(\mathfrak{m},n)$ and gives the same bound. Thus in total each $\mathfrak{m}$ gives rise to at most $1+\frac{4p_y(3n)}{2^k}$ elements in $T\cap\mathbb{A}^n$ as required.

The arguments for the complexity of S are analogous, completing the proof of Theorem 3.4. □

We now state and prove the main result of this paper:

Theorem 3.7. *An infinite word $x\in\mathbb{A}^{\mathbb{N}}$ is of sub-linear complexity if and only if $\mathrm{Fac}(x)\subseteq S^2$ for some $S\subseteq\mathbb{A}^*$ of bounded complexity.*

Proof. Let $x\in\mathbb{A}^{\mathbb{N}}$. First let us suppose that $\mathrm{Fac}(x)\subseteq S^2$ for some $S\subseteq\mathbb{A}^*$ of bounded complexity. The following lemma is stated for general languages L and arbitrary positive integers k, but applied to $L=\mathrm{Fac}(x)$ and $k=2$ implies that $p_x(n)=O(n)$ as required.

Lemma 3.8. *Let $L\subseteq\mathbb{A}^*$. Suppose $L\subseteq S^k$ for some $S\subseteq\mathbb{A}^*$ of bounded complexity. Then $p_L(n)=O(n^{k-1})$.*

Proof. The result is clear in case $k=1$. So let us fix $k\geq 2$, and suppose $L\subseteq S^k$ for some $S\subseteq\mathbb{A}^*$ of bounded complexity. Pick a positive integer C such that $p_S(n)\leq C$ for each $n\geq 0$. Let $u\in L$ and put $n=|u|$. Then u is a concatenation of k elements of S. We claim there are $\binom{n+k-1}{k-1}$ ways of factoring $u=v_1v_2\cdots v_k$ with $|v_i|\geq 0$. In fact, each such factorization of u corresponds to a vector $(n_1,n_2,\ldots,n_k)$ with $n_i\geq 0$ and $n_1+n_2+\cdots n_k=n$. The mapping $(n_1,n_2,\ldots,n_k)\mapsto(n_1+1,n_2+1,\ldots,n_k+1)$ defines a bijection between the sets $A=\{(n_1,n_2,\ldots,n_k)\,|\,n_i\geq 0,\ n_1+n_2+\cdots n_k=n\}$ and $B=\{(m_1,m_2,\ldots,m_k)\,|\,m_i\geq 1,\ m_1+m_2+\cdots m_k=n+k\}$. Since each element of B corresponds to a partition of $n+k$ consecutive points into k non-empty parts, and since each such partition is given by choosing $k-1$ separation points amongst the $n+k-1$ possible separation points, we deduce that $\mathrm{Card}(A)=\mathrm{Card}(B)=\binom{n+k-1}{k-1}$. Having established that there are $\binom{n+k-1}{k-1}=O(n^{k-1})$ ways of factoring $u=v_1v_2\cdots v_k$ with $|v_i|\geq 0$, as each $v_i\in S$, there are at most C choices for each v_i. Thus $p_L(n)\leq C^k\binom{n+k-1}{k-1}$ and thus $p_L(n)=O(n^{k-1})$ as required. □

For the converse, suppose $x\in\mathbb{A}^{\mathbb{N}}$ and $p_x(n)=O(n)$. It suffices to consider the case in which x is aperiodic for otherwise $p_x(n)=O(1)$ and hence we may take $S=\mathrm{Fac}(x)\subseteq S^2$. It follows from the Morse-Hedlund theorem that $p_x(n)=\Theta(n)$. Since we are not assuming that x is recurrent, in order to apply Theorem 3.4 we will need to replace x with a bi-infinite word. Thus, let a be a symbol not belonging to $\mathbb{A}$ and define the bi-infinite word $y=\cdots y_{-2}y_{-1}y_0y_1y_2\cdots\in(\mathbb{A}\cup\{a\})^{\mathbb{Z}}$ by $y_n=x_n$ for $n\geq 0$ and $y_n=a$ for each $n\leq -1$. Note that since $p_y(n)=p_x(n)+n$ and $p_x(n)=\Theta(n)$, it follows that $p_y(n)=\Theta(n)$. Also, since x is aperiodic, the word y

is right aperiodic. We now apply Theorem 3.4 to show that there exist subsets S and T of $\mathbb{A}^*$ of bounded complexity such that $\mathrm{Fac}(y) \subseteq ST$.

Fix a positive integer C such that $p_y(n) \leq Cn$ for each $n \geq 1$. Let $M = \mathcal{R}_y$. By Lemma 3.2, M is a D-marker set for y where $D = C + 1$. By Theorem 3.4 there exist subsets S and T of $\mathbb{A}^*$ such that $\mathrm{Fac}(y) \subseteq ST$ with p_S, p_T satisfying (3.1) where $M = \mathcal{R}_y$ and $D = C + 1$. Since $\mathrm{Fac}(x) \subseteq \mathrm{Fac}(y)$ we have $\mathrm{Fac}(x) \subseteq ST$.

It remains to show that S and T are of bounded complexity. Since $\mathcal{R}_y(n) = \mathcal{R}_x(n) \cup \{a^n\}$ for each $n \geq 0$, by the main result of [5] (see also [7]), there exists a positive integer R such that $p_M(n) \leq R$ for each $n \geq 0$. Moreover $|I_n| = 2 + \log_2 D$ and thus k takes on at most $3 + \log_2 D$ possible values. Furthermore for each such k, we have $\frac{1}{2^k} < \frac{2D}{n}$. Thus, starting with (3.1), we have

$$\begin{aligned} p_S(n), p_T(n) &\leq \sum_{k \in I_n \cap \mathbb{N}} p_M(2^k) \left(1 + \frac{4p_y(3n)}{2^k}\right) \\ &\leq R(3 + \log_2 D) \left(1 + \frac{8Dp_y(3n)}{n}\right) \\ &\leq R(3 + \log_2 D) \left(1 + \frac{24DCn}{n}\right) \\ &= R(3 + \log_2 D)(1 + 24DC) \end{aligned}$$

for each $n \geq 2D$, and hence each of S and T is of bounded complexity as required. □

4. Concluding remarks

Theorem 3.4 may be extended to non-recurrent one-sided infinite words by the same extension argument as in the proof of Theorem 3.7. This is not done in the paper only for the sake of readability.

The upper bound on the complexity of the sets S and T in the proof of Theorem 3.7 is generally far from optimal. In the construction for the Thue-Morse word considered in Example 1.1, we have $p_S(n), p_T(n) = 2$. As another example where the complexity of S and T can be very low, consider Sturmian words.

Example 4.1. For a Sturmian word x, there exist sets S and T of complexity $p_S(n), p_T(n) = 1$ for all n such that $\mathrm{Fac}(x) \subseteq ST$. Indeed, the condition $p_x(n) = n + 1$ implies that x admits a unique left (right, respectively) special factor of each length n denoted $l_x(n)$ ($r_x(n)$, respectively). Moreover, as is well known, $l_x(n)$ and $r_x(n)$ are reversals of one another. Set $S = \{\varepsilon\} \cup \{r_x(n)0 \,|\, n \geq 0\}$ and $T = \{\varepsilon\} \cup \{1l_x(n) \,|\, n \geq 0\}$. Then clearly, $p_S(n), p_T(n) = 1$ for each n. To see that $\mathrm{Fac}(x) \subseteq ST$, recall that for each $n \geq 1$, the word $w(n) = r_x(n-1)01l_x(n-1)$ is a factor of x of length $2n$ (see for instance Exercise 6.1.24 in [3]). It is easily checked that for each $n \geq 1$, $w(n)$ contains $n + 1$ distinct factors of length n. Whence each factor of x of length n is a factor of $w(n)$ and hence $\mathrm{Fac}(x) \subseteq ST$ as required.

In the proof of Theorem 3.7, instead of Lemma 3.2, we could use the set of markers from Proposition 3.3. It would give a different upper bound for the complexity of S and T.

Theorem 3.7 does not extend to arbitrary languages L with $p_L(n) = O(n)$. Here is a counter-example showing it.

Proposition 4.2. *There exists a non-factorial language L of complexity $p_L(n) = O(\log n)$ which is not contained in any finite product of the form S^k where S is a language of bounded complexity.*

Proof. For each positive integer n, define $x_n \in \{0,1,2\}^*$ by $x_n = [n]_2 2$, where $[n]_2$ is the binary representation of n. For example, $x_2 = 102$ and $x_{65} = 10000012$. Clearly, $|x_n| = \lfloor \log_2 n \rfloor + 2$. Next define y_n as the longest prefix of x_n^ω satisfying $|y_n| \log_2 |y_n| \leq n$. Thus for example $y_2 = 10$ since $2\log_2 2 \leq 2 < 3\log_2 3$ and $y_{65} = 1000001210000012$ since $16\log_2 16 \leq 65 < 17\log_2 17$. Finally, define $L = \{y_n | n \geq 1\}$.

We first claim that $|y_n| = \Theta(\frac{n}{\log n})$. Indeed, for $n \geq 2$, $|y_n| \geq 2$ so that $\log_2 |y_n| \geq 1$ and

$$|y_n| \leq \frac{n}{\log_2 |y_n|} \leq n.$$

Since the length $|y_n|$ was chosen to be maximal,

$$|y_n| + 1 > \frac{n}{\log_2(|y_n|+1)} \geq \frac{n}{\log_2(n+1)},$$

so $|y_n| = \Omega(n/\log n)$. Combining the two previous equations yields

$$|y_n| \leq \frac{n}{\log_2(\frac{n}{\log_2(n+1)} - 1)}.$$

Since $\frac{n}{\log_2(n+1)} - 1$ is asymptotically equivalent to $\frac{n}{\log_2 n}$ we deduce $|y_n| = O(\frac{n}{\log n})$. Together with the lower bound above, this gives $|y_n| = \Theta(\frac{n}{\log n})$ as required.

Next we claim that $p_L(n) = \Theta(\log n)$. Indeed,

$$p_L(n) = \mathrm{Card}\{m \,|\, |y_m| = n\}.$$

In other words,

$$p_L(n) = \mathrm{Card}\{m \,|\, n\log_2 n \leq m < (n+1)\log_2(n+1)\}.$$

Whence,

$$p_L(n) \quad = \quad \lceil (n+1)\log_2(n+1) \rceil - \lceil n \log_2 n \rceil = \Theta(\log n).$$

Finally we claim L is not contained in any finite product of the form S^k where S is of bounded complexity. Indeed, suppose to the contrary that $L \subseteq S^k$ for some $k \in \mathbb{N}$ and some set S of bounded complexity. Since

$$\frac{|y_n|}{|x_n|} = \Theta\left(\frac{n}{(\log_2 n)^2}\right),$$

there exists an integer $n_0 > 0$ such that for all $n > n_0$, we have $|y_n| \geq (k+1)|x_n|$. This means that for all $n > n_0$, the word y_n contains at least $k+1$ occurrences of 2, and, by the pigeon hole principle, at least two of them are located in the same word from S, denote it by s_n. Since between two occurrences of 2 in s_n, there is exactly the binary representation of n, all s_n for $n > n_0$ are pairwise distinct.

Now for each $n > n_0$ consider the set $S(n) = \{s_m | n_0 < m \leq n\} \subseteq S$. It contains $n - n_0$ distinct words, and the length of each of them is $o(n)$: indeed, $|s_m| \leq |y_m| = \Theta(n/\log n)$. So the total number of factors of length at most n of S grows faster than linearly, which is impossible if its complexity is bounded. □

References

1. J.-P. Allouche, J. Shallit, *Automatic Sequences, Theory, Applications, Generalizations*, Cambridge University Press, 2003.
2. V. I. Arnold, Small denominators and problems of stability of motion in classical and celestial mechanics, *Usp. Math. Nauk.* **18** (1963), pp. 91–192, (in Russian) translated in *Russian Math. Surveys* **18** (1963), pp. 86–194.
3. P. Arnoux, Chapter 6: Sturmian sequences, in *Substitutions in Dynamics, Arithmetics and Combinatorics*, Lecture Notes in Math. **1794**, Springer Verlag, Berlin, 2002, pp. 143–198.
4. S. Brlek, Enumeration of factors in the Thue-Morse word, *Disc. Appl. Math.* **24** (1989), pp. 83–96.
5. J. Cassaigne, Special factors of sequences with linear subword complexity, in *Proceedings of DLT 1995*, World Sci. Publishing, Singapore, 1996, pp. 25–34.
6. J. Cassaigne, A. Frid, S. Puzynina, L. Zamboni, Subword complexity and decomposition of the set of factors, in *Proceedings of MFCS 2014*, Lect. Notes Comput. Sci. **8634**, Springer, 2014, pp. 147–158.
7. J. Cassaigne, F. Nicolas, Chapter 4: Factor complexity, in *Combinatorics, automata and number theory*, Encyclopedia Math. Appl. **135**, Cambridge Univ. Press, 2010, pp. 163–247.
8. A. de Luca, S. Varricchio, Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups, *Theoret. Comput. Sci.* **63** (1989), pp. 333–348.
9. A. Ehrenfeucht, K. P. Lee, G. Rozenberg, Subword complexities of various deterministic developmental languages without interactions, *Theoret. Comput. Sci.* **1** (1975), pp. 59–76.
10. A. Glen and J. Justin, Episturmian words: a survey, *Theor. Inform. Appl.,* **43** (2009), pp. 403442.
11. J. Leroy, Some improvements of the S-adic conjecture, *Adv. in Appl. Math.* **48** (2012), pp. 79–98.
12. M. Lothaire, *Combinatorics on words*, Addison-Wesley Publishing Co., Reading, Mass., 1983.
13. M. Lothaire, *Algebraic combinatorics on words*, Cambridge University Press, 2002.
14. M. Lothaire, *Applied combinatorics on words*, Cambridge University Press, 2005.
15. M. Morse, G. Hedlund, Symbolic dynamics, *Amer. J. Math.* **60** (1938), pp. 815–866.
16. M. Morse, G. Hedlund, Symbolic dynamics II: Sturmian sequences, *Amer. J. Math.* **62** (1940), pp. 1–42.
17. J.-J. Pansiot. Complexité des facteurs des mots infinis engendrés par morphismes itérés, in *Proceedings of ICALP 1984*, Lect. Notes Comput. Sci. **172**, Springer, 1984, pp. 380–389.
18. M. Queffélec, Substitution Dynamical Systems - Spectral Analysis, Lecture Notes in Math. **1284**, Springer, 1987.
19. G. Rote, Sequences with subword complexity $2n$, *J. Number Theory* **46** (1994), pp. 196–213.
20. A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat-Nat. Kl.* **7** (1906), pp. 1–22.
21. A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat-Nat. Kl.* **1** (1912), pp. 1–67.

CNRS, Aix Marseille Univ, Centrale Marseille, I2M, Marseille, France
E-mail address: julien.cassaigne@math.cnrs.fr

Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France
E-mail address: anna.e.frid@gmail.com

Saint Petersburg State University and Sobolev Institute of Mathematics, Novosibirsk, Russia
E-mail address: s.puzynina@gmail.com

Institut Camille Jordan, Université Lyon 1, France
E-mail address: zamboni@math.univ-lyon1.fr