On the character of words of sub-linear complexity

Luca Q. Zamboni

On the character of words of sub-linear complexity

Luca Q. Zamboni

Université de Lyon, Université Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 boulevard du 11 novembre 1918, F69622 Villeurbanne Cedex, France

Abstract

Let $\mathbb{A}^*$ denote the free monoid generated by a finite nonempty set $\mathbb{A}$. For each infinite word $x = x_0x_1x_2\cdots \in \mathbb{A}^\omega$, the factor complexity $p_x(n)$ counts the number of distinct blocks $x_ix_{i+1}\cdots x_{i+n-1}$ of length n occurring in x. In other words, the factor complexity of x is the complexity of the language of its factors $\mathrm{Fac}_x = \{x_ix_{i+1}\cdots x_j \,|\, 0 \leq i \leq j\}$. Our starting point in this paper is the following characterisation of infinite words of sub-linear factor complexity obtained recently by the author together with J. Cassaigne, A. Frid and S. Puzynina: Let $x \in \mathbb{A}^\omega$. Then $p_x(n) = O(n)$ if and only if $\mathrm{Fac}(x) \subseteq S^2$ for some $S \subseteq \mathbb{A}^*$ with $\limsup p_S(n) < +\infty$. In other words, $p_x(n) \leq Cn$ for some constant C if and only if there exists a set S of bounded complexity such that every factor w of x can be factored as $w = uv$ with $u, v \in S$. Given an infinite word $x \in \mathbb{A}^\omega$, we define its *character*, denoted $\chi(x)$, to be the least value for $\limsup p_S(n)$ over all languages S such that $\mathrm{Fac}_x \subseteq S^2$. Clearly $\chi(x) \in [1, +\infty]$ and it follows from the above characterisation that $p_x(n) = O(n)$ if and only if $\chi(x) < +\infty$. We prove that $\chi(x) = 1$ if and only if x is ultimately periodic and that $\chi(x) = 2$ whenever x is a Sturmian word.

Keywords: Combinatorics on words, complexity.
2010 MSC: 68R15

1. Introduction

Let $\mathbb{A}$ be a finite non-empty set. For each infinite word $x = x_0x_1x_2\cdots \in \mathbb{A}^\omega$, 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. First introduced by Hedlund and Morse in their seminal 1938 paper [20] under the name of *block growth*, the factor complexity measures the extent of randomness of x : Periodic (meaning ultimately periodic) words have bounded factor complexity while digit expansions of normal numbers have maximal complexity. A celebrated theorem of Morse and Hedlund in [20] states that every aperiodic (meaning non-ultimately periodic) word contains at least $n + 1$ distinct factors of each length n. Words having precisely $n + 1$

Email address: zamboni@math.univ-lyon1.fr (Luca Q. Zamboni)

factors of each given length n are called Sturmian words. From the point of view of the Morse-Hedlund theorem, they are those aperiodic words of minimal factor complexity. They arise naturally in various branches of mathematics including combinatorics, algebra, number theory, ergodic theory, dynamical systems and differential equations. In theoretical physics, Sturmian words constitute 1-dimensional models for quasi-crystals, and in theoretical computer science they are used in computer graphics as digital approximation of straight lines. Perhaps the most studied Sturmian word is the so-called Fibonacci word

$$x = 0100101001001010010\cdots$$

defined as the fixed point of the substitution $0 \mapsto 01, 1 \mapsto 0$.

There are several variations and extensions of the Morse-Hedlund theorem associated with other notions of complexity including Abelian complexity [9, 24], Abelian first returns [23], maximal pattern complexity [15], palindrome complexity [1], cyclic complexity [5] and group complexity [8] to name just a few. In most cases, these alternative notions of complexity may be used to detect (and in some cases characterize) ultimately periodic words. Generally, amongst all aperiodic words, Sturmian words have the lowest possible complexity, although in some cases they are not the only ones. For instance, a restricted class of Toeplitz words is found to have the same maximal pattern complexity as Sturmian words (see section 4 of [16]). There have also been numerous attempts at extending the Morse-Hedlund theorem in higher dimensions. A celebrated conjecture of M. Nivat states that any 2-dimensional word having at most mn distinct $m \times n$ blocks must be periodic. In this case, it is known that the converse is not true. To this day the Nivat conjecture remains open although the conjecture has been verified for m or n less or equal to 3 (see [10, 25]). A very interesting higher dimensional analogue of the Morse-Hedlund theorem was obtained by Durand and Rigo in [12] in which they re-interpret the notion of periodicity in terms of Presburger arithmetic.

Beyond Sturmian words, there are many other rich families of infinite words having low (meaning sub-linear) factor complexity. They include all words generated by primitive substitutions and more generally all linearly recurrent words [11], automatic sequences [2], Arnoux-Rauzy sequences [4] and symbolic codings of interval exchange transformations [3]. In [13], S. Ferenczi proved that the language of any uniformly recurrent word having sub-linear factor complexity is S-adic, meaning that for any such x there exist a finite set S of morphisms over some finite alphabet $\mathbb{A}$, $a \in \mathbb{A}$ and an infinite sequence $(\tau_n)_{n\geq 0} \in S^\omega$ such that $x = \lim_{n\to\infty} \tau_0 \cdots \tau_n(a^\omega)$. An important open question along these lines is the so-called S-adic conjecture which involves finding some condition (*) such that for any word x, the factor complexity of x is sub-linear if and only if x is an S-adic word verifying condition (*). See for instance [17] and the references therein.

Our starting point in this paper is a different characterisation of infinite words $x \in \mathbb{A}^\omega$ of sub-linear factor complexity recently obtained by the author together with J. Cassaigne, A. Frid and S. Puzynina in [6]:

Theorem 1 (Theorem 1 in [6]). *Let $\mathbb{A}$ be a finite non-empty set and $x \in \mathbb{A}^\omega$. Then $p_x(n) = O(n)$ if and only if $\mathrm{Fac}_x \subseteq S^2$ for some $S \subseteq \mathbb{A}^*$ with $\limsup p_S(n) < +\infty$.*

In other words, $p_x(n) \leq Cn$ for some constant C if and only if there exists a set S of bounded complexity such that every factor w of x can be factored as $w = uv$ with $u, v \in S$. The proof of Theorem 1 in [6] is constructive in the sense that if $p_x(n) = O(n)$, it describes a way of constructing a set S of bounded complexity such that $\mathrm{Fac}_x \subseteq S^2$. However, it is natural to ask what is the smallest value for $\limsup p_S(n) < +\infty$ over all languages S for which $\mathrm{Fac}_x \subseteq S^2$. More precisely, for each $x \in \mathbb{A}^\omega$, we define the *character* of x by

$$\chi(x) = \inf\{\limsup p_S(n) \,:\, S \subseteq \mathbb{A}^*,\, \mathrm{Fac}_x \subseteq S^2\}$$

where we take the convention that $\inf \emptyset = +\infty$. Thus $\chi(x) \in [1, +\infty] \cap \mathbb{Z}$ and by Theorem 1 we have that $p_x(n) = O(n)$ if and only if $\chi(x) < +\infty$. In this paper we prove that $\chi(x) = 2$ whenever x is a Sturmian word. More precisely, given a Sturmian word $x \in \{0,1\}^\omega$ there exists a set S such that $\mathrm{Card}(S \cap \{0,1\}^n) = 2$ for all $n \geq 1$ and $\mathrm{Fac}_x \subseteq S^2$. Moreover, this is optimal in the sense that $\chi(x) = 1$ if and only if x is ultimately periodic.

The paper is organised as follows: In §2 we recall some fundamental notions concerning finite and infinite words. In §4 we establish the main result of the paper, namely that an infinite word x is ultimately periodic if and only if $\chi(x) = 1$. In §5 we prove that $\chi(x) = 2$ whenever x is a Sturmian word.

2. Preliminaries

In this section we briefly recall some basic definitions and notations concerning finite and infinite words which are relevant to the subsequent sections. For a more detailed exposition, the reader is referred to one of the standard texts in combinatorics on words such as the Lothaire books [18, 19].

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 a finite word $u = a_1 a_2 \dots a_n$ with $n \geq 1$ and $a_i \in \mathbb{A}$, we denote the length n of u by $|u|$. For each $a \in \mathbb{A}$, we let $|u|_a$ denote the number of occurrences of the letter a in u. Let $\mathbb{A}^\omega$ denote the set of (right) infinite words $x = x_0 x_1 x_2 \cdots$ with $x_i \in \mathbb{A}$. Given an infinite word $x \in \mathbb{A}^\omega$, a word $u \in \mathbb{A}^*$ is called a *factor* of x if either $u = \varepsilon$ or if $u = x_i x_{i+1} \cdots x_{i+n}$ for some natural numbers i and n. We denote by $\mathrm{Fac}_x(n)$ the set of all factors of x of length n, and set

$$\mathrm{Fac}_x = \bigcup_{n \in \omega} \mathrm{Fac}_x(n).$$

Let $p_x : \mathbb{N} \to \mathbb{N}$ denote the *factor complexity* of x defined by:

$$p_x(n) = \mathrm{Card}(\mathrm{Fac}_x(n)).$$

An infinite word x is called *ultimately periodic* if $x = uvvv\cdots = uv^\omega$ for some non-empty words $u, v \in \mathbb{A}^*$. An infinite word is said to be *aperiodic* if it is not ultimately periodic. A celebrated theorem of Morse and Hedlund in [20] states that every aperiodic word contains at least $n+1$ distinct factors of each length n. Sturmian words are infinite words having exactly $n+1$ distinct factors of each length $n \geq 1$.

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 $a, b \in \mathbb{A}$. A factor u of x is called *bispecial* if u is both left and right special. It follows that every aperiodic word contains a right and a left special factor of each length. More precisely, let $x \in \mathbb{A}^\omega$ be aperiodic. Given a factor u of x, let $\mathscr{R}(u)$ ($\mathscr{L}(u)$, respectively) denote the shortest right (left, respectively) special factor of x beginning (ending, respectively) in u. We note that if u is right (left, respectively) special, then $\mathscr{L}(u)$ ($\mathscr{R}(u)$, respectively) is bispecial. Also, since x is aperiodic, for any factors u and v of x with $|u| < |v|$, if u is both a prefix and a suffix of v, then $\mathscr{R}(u)$ is a prefix of v. Otherwise, every occurrence of u in x is an occurrence of v in x and x would be ultimately periodic.

For each finite word u on the alphabet $\mathbb{A}$ we set

$$x\big|_u = \{n \in \omega \mid x_n x_{n+1} \cdots x_{n+|u|-1} = u\}.$$

In other words, $x\big|_u$ denotes the set of all occurrences of u in x. We say x is *recurrent* if for every $u \in \mathrm{Fac}_x$ the set $x\big|_u$ is infinite. We say x is *uniformly recurrent* if for every $u \in \mathrm{Fac}_x$ the set $x\big|_u$ is syndedic, i.e., of bounded gap.

3. A characterisation for periodicity

We begin by establishing some preliminary results needed in the proof of our main result (see Theorem 5). Throughout this section $\mathbb{A}$ will denote a finite nonempty set.

Lemma 2. *Let $C, N \in \mathbb{Z}^+$ and $x \in \mathbb{A}^\omega$ be such that $p_x(n) = n + C$ for all $n \geq N$.* [1] *Then there exist distinct letters $a, b \in \mathbb{A}$ and, for each $n \geq N$, a unique factor $r_n \in \mathrm{Fac}_x(n)$ having exactly two right extensions in x of length $n+1$ given by $r_n a$ and $r_n b$, while all other factors of x of length n have a unique right extension in x to a factor of length $n+1$.*

Proof. Clearly for each $n \geq 0$, each word $u \in \mathrm{Fac}_x(n)$ is a prefix of at least one word $v \in \mathrm{Fac}_x(n+1)$. Now since $p_x(n+1) - p_x(n) = 1$ for each $n \geq N$, it follows that for each $n \geq N$, there exists a unique factor r_n of x of length n having exactly two right extensions $r_n a_n$ and $r_n b_n$ to a factor of x of length $n+1$, where a_n and b_n are distinct letters in $\mathbb{A}$, and all other factors of x of length n have a unique right extension to a factor of length

[1] Infinite words x of complexity $p_x(n) = n + C$ for all n sufficiently large were studied and characterised by A. Heinis in his doctoral thesis. He proved that every such word is a morphic image of a Sturmian word (see also [14]).

$n+1$. Moreover, since r_n is necessarily a suffix of r_{n+1} it follows that $\{a_n, b_n\} = \{a_{n+1}, b_{n+1}\}$, whence there exist distinct letters $a, b \in \mathbb{A}$ such that $\{a, b\} = \{a_n, b_n\}$ for each $n \geq N$. □

Lemma 3. *Let $C, N \in \mathbb{Z}^+$ and $x \in \mathbb{A}^\omega$ be such that $p_x(n) = n + C$ for all $n \geq N$. If x is recurrent, then there exist distinct letters $a, b \in \mathbb{A}$ and, for each $n \geq N$, a unique factor $l_n \in \mathrm{Fac}_x(n)$ having exactly two left extensions in x of length $n+1$ given by al_n and bl_n while all other factors of x of length n have a unique left extension in x to a factor of length $n+1$.*

Proof. The proof is identical to that of Lemma 2 with prefix replaced by suffix and right replaced by left. □

Proposition 4. *Let $C, N \in \mathbb{Z}^+$ and $x \in \mathbb{A}^\omega$ be such that $p_x(n) = n + C$ for all $n \geq N$. Then there exists a uniformly recurrent suffix y of x.*

Proof. We begin by showing that there exists a recurrent suffix y of x. Then we show this same suffix is uniformly recurrent. Let $a, b \in \mathbb{A}$ and $r_n \in \mathrm{Fac}_x(n)$ $(n \geq N)$ be as in Lemma 2. Let

$$R = \{u \in \mathbb{A}^+ : |u| \leq N,\ \mathrm{Card}(x\big|_u) < +\infty\}.$$

In other words, R is the set of all non-recurrent factors u of x with $|u| \leq N$. Pick a suffix y of x such that $y\big|_u = \emptyset$ for all $u \in R$. We claim that y is recurrent. In fact, supposing y were not recurrent, let v be the shortest non-recurrent prefix of y and let y' be a suffix of y such that $y'\big|_v = \emptyset$. Then $|v| > N$ otherwise v belongs to R. Let u be the prefix of v of length $|v| - 1$ so that $v = uc$ for some $c \in \mathbb{A}$. Set $n = |u|$. Then $n \geq N$ and since u is recurrent in y and $v = uc$ is not recurrent in y, it follows that u is a right special factor of x and hence by Lemma 2 $u = r_n$ and $c \in \{a, b\}$. Moreover, since u is the unique right special factor of x of length n and the only right extensions of u in x of length $n+1$ are ua and ub, it follows that the suffix y' of y contains no right special factors of length n. In other words, $p_{y'}(n+1) = p_{y'}(n)$ whence y' is ultimately periodic, a contradiction.

Next we show y is uniformly recurrent. It suffices to show that every prefix u of y of length $|u| \geq N$ occurs in y with bounded gap. Since y is recurrent, every prefix u of y is contained in a right special prefix of y. Thus it suffices to show that every right special prefix u of y with $|u| \geq N$ occurs in y with bounded gap. So let us fix a right special prefix u of y with $|u| \geq N$. Set $n = |u|$ and consider any factor w of y of length $2n + C$. We claim that u occurs in w. In fact, there are $n + C + 1$ occurrences in w of words of length n. Since $p_y(n) \leq p_x(n) = n + C$, there exists some factor v of y of length n occurring at least twice in w. Let z be a factor of w which begins and ends in v and of length $|z| > |v|$. Since y is aperiodic, there exists a right special prefix z' of z of length $|z'| \geq |v| = n$. Then by Lemma 2 we deduce that u is a suffix of z' and hence u occurs in w as required. □

Theorem 5. *A word $x \in \mathbb{A}^\omega$ is ultimately periodic if and only if $\chi(x) = 1$ where*

$$\chi(x) = \inf\{\limsup p_S(n) \ :\ S \subseteq \mathbb{A}^*,\ \mathrm{Fac}_x \subseteq S^2\}$$

Proof. One direction is straightforward. Suppose $x = uv^\omega$ for some $u, v \in \mathbb{A}^+$. Let

$$S = \mathrm{Fac}_{uv} \cup \mathrm{Pref}(v^\omega).$$

We claim $\mathrm{Fac}_x \subseteq S^2$ from which it follows that $\chi(x) = 1$ (since $p_S(n) = 1$ for all $n \geq |uv|+1$). In fact, let $w \in \mathrm{Fac}_x$. If $w \in \mathrm{Fac}_{uv}$, then $w \in S$ and so $w = w \cdot \varepsilon \in S^2$. Otherwise, we can write $w = w'w'' \in S^2$ where w' is a suffix of uv and w'' a prefix of v^ω.

For the reverse implication, let us assume that $x \in \mathbb{A}^\omega$ and $\chi(x) = 1$. Fix $S \subseteq \mathbb{A}^*$ and a positive integer N_0 with $\mathrm{Fac}_x \subseteq S^2$ and $p_S(n) \leq 1$ for all $n \geq N_0$.

Without loss of generality, we may assume that $p_S(n) = 1$ for all $n \geq N_0$. In fact, we could extend S to a set $\tilde{S}$ by adding to S an arbitrary element of $\mathbb{A}^n$ for each n such that $S \cap \mathbb{A}^n = \emptyset$. Then $\mathrm{Fac}_x \subseteq \tilde{S}^2$ and $p_{\tilde{S}}(n) = 1$ for all $n \geq N_0$.

Lemma 6. *Assume $x \in \mathbb{A}^\omega$, $S \subseteq \mathbb{A}^*$ and $N_0 \in \mathbb{Z}^+$ are such that $\mathrm{Fac}_x \subseteq S^2$ and $p_S(n) = 1$ for all $n \geq N_0$. Then either x is ultimately periodic, or there exist positive integers C, N with $N \geq 2N_0$ such that $p_x(n) = n + C$ for all $n \geq N$.*

Proof. Assume x is aperiodic. Let $C_0 = \max\{p_S(n) \,|\, 0 \leq n < N_0\}$. Fix $n \geq 2N_0$. Each $w \in \mathrm{Fac}_x(n)$ may be written as $w = uv$ with $u, v \in S$. Thus an upper bound for $p_x(n)$ is obtained by counting the number words of the form uv with $u, v \in S$ and $|u| + |v| = n$. If u or v is in the range $1 \leq |u|, |v| \leq N_0 - 1$, then we obtain at most C_0 distinct words of the form uv with $u, v \in S$. While if u and v are in the range $N_0 \leq |u|, |v| \leq n - N_0$ or $\{|u|, |v|\} = \{0, n\}$, then we obtain at most 1 word of the form uv with $u, v \in S$. Whence

$$p_x(n) \leq 2C_0(N_0 - 1) + n - 2N_0 + 3. \tag{1}$$

Thus setting $C_1 = 2C_0(N_0 - 1) - 2N_0 + 3$ we have that $p_x(n) \leq n + C_1$ for all $n \geq 2N_0$. Let $C \leq C_1$ be the largest positive integer for which we have $p_x(n) = n + C$ for some $n \geq 2N_0$, and pick $N \geq 2N_0$ such that $p_x(N) = N + C$. Then it is readily checked by induction that $p_x(n) = n+C$ for all $n \geq N$. In fact, if for $n \geq N$ we have $p_x(n) = n+C$, then by maximality of the choice of C we have $p_x(n + 1) \leq n + 1 + C$. On the other hand since x is aperiodic, $p_x(n + 1) \geq p_x(n) + 1 = n + 1 + C$. Thus $p_x(n + 1) = n + 1 + C$ as required. □

Returning to the proof of Theorem 5, we wish to show that if $x \in \mathbb{A}^\omega$, $S \subseteq \mathbb{A}^*$ and $N_0 \in \mathbb{Z}^+$ are such that $\mathrm{Fac}_x \subseteq S^2$ and $p_S(n) = 1$ for all $n \geq N_0$, then x is ultimately periodic. In view Lemma 6 combined with Proposition 4, short of replacing x by a suffix of x, we may assume without loss of generality that x itself is uniformly recurrent. So henceforth in proof of Theorem 5, we will assume that $x \in \mathbb{A}^\omega$ is uniformly recurrent, $S \subseteq \mathbb{A}^*$ and $N_0 \in \mathbb{Z}^+$ are such that $\mathrm{Fac}_x \subseteq S^2$ and $p_S(n) = 1$ for all $n \geq N_0$. As before let $C_0 = \max\{p_S(n) \,|\, 0 \leq n < N_0\}$. We note $C_0 \geq 1$.

We propose to show that x is ultimately periodic. So suppose to the contrary that x is aperiodic. Let N and C be as in Lemma 6. Also set

$$d = 2(N_0 - 1)(C_0 - 1). \tag{2}$$

We will show that for each $K \geq \max\{4, N, 6d+1\}$ there exists $z \in \mathbb{A}^2$ such that $z^{\lfloor \frac{K}{2} \rfloor} \in \mathrm{Fac}_x$. As x is assumed to be uniformly recurrent, this will give the desired contradiction.

For each $i \geq N_0$, let s_i denote the unique element of S of length i. For each $n \geq N$, let

$$d_n = \mathrm{Card}\{i : N_0 \leq i \leq n - N_0,\ s_i s_{n-i} \notin \mathrm{Fac}_x\}.$$

Recall that each factor w of x of length n is a product uv with $u, v \in S$. Thus d_n corresponds to the number of factorisations uv with $N_0 \leq |u|, |v| \leq n - N_0$ which are not needed in generating all factors of x of length n. Hence, for each $n \geq N$, a refinement of the estimation used in the proof of Lemma 6 yields:

$$2C_0(N_0 - 1) + n - 2N_0 + 3 - d_n \geq p_x(n) \geq n + 1 \tag{3}$$

where the last inequality is a consequence of the Morse-Hedlund theorem as we are assuming that x is aperiodic. It follows from (2) and (3) that $d_n \leq d$.
For $n \geq N$, let

$$U(n) = \{u \in S \,|\, N_0 \leq |u| \leq n - N_0, \text{ and } \exists v \in S \text{ with } \{uv, vu\} \subseteq \mathrm{Fac}_x(n)\}.$$

We note that our assumption on $p_S(n)$ implies that if $u \in U(n)$, then there exists a unique $v \in S$ with $\{uv, vu\} \subseteq \mathrm{Fac}_x(n)$ and v too belongs to $U(n)$. In other words, if $s_i s_{n-1} \notin \mathrm{Fac}_x(n)$ for $N_0 \leq i \leq n - N_0$, then $s_i \notin U(n)$ and $s_{n-1} \notin U(n)$. Thus

$$\mathrm{Card}\, U(n) \geq n - 2N_0 + 1 - 2d_n \geq n - 2N_0 + 1 - 2d. \tag{4}$$

Let $\mathcal{B}_x = \{B_0, B_1, B_2, \ldots\}$ denote the infinite set of bispecial factors of x and set $b_i = |B_i|$. We order $\mathcal{B}_x$ so that $b_m < b_{m+1}$ for each $m \geq 0$. Recall that by Lemma 6, we have $p_x(n) = n + C$ for each $n \geq N$ and hence by application of Lemma 2 and Lemma 3 it follows that x has a unique right and left special factor, denoted r_n and l_n, of each length $n \geq N$. Hence given any two bispecial factors of x each of length greater or equal to N, the shorter of the two is both a prefix and a suffix of the other. Since x is uniformly recurrent, it follows that $\liminf_{m\to\infty}(b_{m+1} - b_m) = +\infty$. In fact, if $b_m \geq N$ then setting $\pi_m = b_{m+1} - b_m$ we have that π_m is a period of B_{m+1}.

Let K be any positive integer verifying $K \geq \max\{4, N, 6d + 1\}$. We will show that for each such K there exists $z \in \mathbb{A}^2$ such that $z^{\lfloor \frac{K}{2} \rfloor} \in \mathrm{Fac}_x$. Pick a positive integer m such that

$$b_{m+1} - 6K > b_m > N,$$

and set $n = b_m + 2K$. For convenience, we summarise the inclusions thus far

$$1 \leq N_0 < N < b_m < b_m + K < n = b_m + 2K < b_m + 6K < b_{m+1}.$$

We claim there exists

$$u \in U(n) \cap U(n+1) \cap U(n+2)$$

with

$$b_m < |u| \leq b_m + K. \tag{5}$$

In fact, consider all elements $u \in S$ of length $b_m < |u| \leq b_m + K$. Then by (4) at most $2d$ of them do not belong to $U(n)$. Of those which belong to $U(n)$, at most $2d$ do not belong to $U(n+1)$. Of those which belong to $U(n) \cap U(n+1)$, at most $2d$ do not belong to $U(n+2)$. Since $K \geq 6d+1$ the result follows. Thus there exists

$$\{u, v^{(0)}, v^{(1)}, v^{(2)}\} \subseteq S$$

with $b_m < |u| \leq b_m + K$ and $\{uv^{(i)}, v^{(i)}u\} \subseteq \mathrm{Fac}_x(n+i)$ for each $0 \leq i \leq 2$. We note that $|v^{(i+1)}| = |v^{(i)}| + 1$ and since $|v^{(0)}| = n - |u| = b_m + 2K - |u|$ it follows from (5) that

$$2K > |v^{(0)}| \geq K. \tag{6}$$

We now consider three possible cases: *Case 1*: $uv^{(2)}$ is a prefix of $\mathscr{R}(u)$ and $v^{(2)}u$ is a suffix of $\mathscr{L}(u)$. *Case 2*: There exists a proper prefix w of $uv^{(2)}$ of length $|w| \geq |u|$ which is a right special factor of x. *Case 3*: There exists a proper suffix w of $v^{(2)}u$ of length $|w| \geq |u|$ which is a left special factor of x.

In Case 1, every occurrence of u in x is an occurrence of $uv^{(2)}$. Hence $v^{(1)}$ is a prefix $v^{(2)}$. Similarly, every occurrence of u in x is immediately preceded by $v^{(2)}$ whence $v^{(1)}$ is also a suffix of $v^{(2)}$. Hence $v^{(2)} = c^k$ for some $c \in \mathbb{A}$. Since $k = |v^{(2)}| = |v^{(0)}| + 2 \geq K + 2$, by taking $z = c^2$ we have that $z^{\lfloor \frac{K}{2} \rfloor} \in \mathrm{Fac}_x$.

In Case 2, we first claim that w is unique, i.e., $w = \mathscr{R}(u)$. In fact, suppose that w and w' are both proper right special prefixes of $uv^{(2)}$ with $|w'| > |w| \geq |u|$. Then as w is also a suffix of w' it follows that w occurs twice in $uv^{(2)}$. Then $\mathscr{L}(w)$ is a bispecial factor of x contained in $uv^{(2)}$ whence

$$b_m < |u| \leq |w| \leq |\mathscr{L}(w)| \leq |uv^{(2)}| = n + 2 = b_m + 2K + 2 < b_{m+1}$$

a contradiction since b_m and b_{m+1} are the lengths of consecutive bispecial factors of x. We next claim that at least two $v^{(i)}$ are prefixes of one another. In fact, if $uv^{(0)}$ is not a prefix of $uv^{(1)}$, then for some choice of distinct letters $a, b \in \mathbb{A}$ we would have that wa is a prefix of $uv^{(0)}$ and wb is a prefix of $uv^{(1)}$. We recall that by Lemma 2, w has only two right extensions in x of length $|w| + 1$, namely wa and wb. Thus if in addition $uv^{(0)}$ is not a prefix of $uv^{(2)}$,

then wb is a prefix of $uv^{(2)}$. But then, since wb is a both a prefix of $uv^{(1)}$ and $uv^{(2)}$ and every occurrence in x of wb is an occurrence in x of $uv^{(2)}$, it follows that $uv^{(1)}$ is a prefix of $uv^{(2)}$. Finally we claim that no suffix w' of $v^{(2)}u$ with $|w'| \geq |u|$ is left special. In fact, if $\mathscr{L}(u)$ were a suffix of $v^{(2)}u$, then $\mathscr{L}(w) = \mathscr{L}(\mathscr{R}(u))$ is a bispecial factor of x contained in $v^{(2)}uv^{(2)}$, whence

$$b_m < |u| \leq |\mathscr{L}(\mathscr{R}(u))| \leq |v^{(2)}uv^{(2)}| = |u|+2|v^{(2)}| < b_m+K+2(2K+2) = b_m+5K+4 < b_{m+1}$$

again a contradiction since b_m and b_{m+1} are the lengths of consecutive bispecial factors of x. Thus in summary we have that each of the $v^{(i)}$ are suffixes of one another and at least two of the $v^{(i)}$ are prefixes of one another. Whence either $v^{(1)}$ or $v^{(2)}$ has period equal to 1 or 2. In either case we have $z^{\lfloor \frac{K}{2} \rfloor}$ is a factor of x for some $z \in \mathbb{A}^2$.

Case 3 works similarly to Case 2. Using Lemma 3 we argue that each of the $v^{(i)}$ are prefixes of one another and at least two of the $v^{(i)}$ are suffixes of one another and hence either $v^{(1)}$ or $v^{(2)}$ has period equal to 1 or 2. This completes our proof of Theorem 5. □

4. The character of Sturmian words

Let $x = x_0x_1x_2\cdots \in \{0,1\}^\omega$ be a Sturmian word, i.e., $p_x(n) = n+1$ for each $n \geq 0$. We will also make use of the *balance* characterisation of Sturmian words due to Morse and Hedlund in [21] and Coven and Hedlund in [9]: $x \in \{0,1\}^\omega$ is Sturmian if and only if x is aperiodic and for any two factors u and v of x of equal length we have $||u|_0 - |v|_0| \leq 1$. The slope $s(x)$ is defined as the frequency of the symbol 1 in x, i.e.,

$$s(x) = \lim_{n\to\infty} \frac{|x_0x_1\cdots x_{n-1}|_1}{n}.$$

In [21], Hedlund and Morse showed that this limit always exists and that x is the symbolic coding of the orbit of a point $\rho(x)$, called the intercept, on the unit circle under a rotation by an irrational angle $s(x)$, where the circle is partitioned into two complementary intervals, one of length $s(x)$ and the other of length $1-s(x)$. Conversely each such coding defines a Sturmian word. Moreover, the set of factors of a Sturmian word depends only on the slope: Let x, y be two Sturmian words. Then $\mathrm{Fac}_x = \mathrm{Fac}_y$ if and only if x and y have the same slope. Let $\mathcal{O}(x) = \{y \in \mathbb{A}^\omega : \mathrm{Fac}_y = \mathrm{Fac}_x\}$.

The condition $p_x(n) = n+1$ implies that x admits a unique left and right special factor of each length n denoted l_n and r_n respectively. For each $n \geq 0$ we have that l_n is a prefix of l_{n+1} and r_n is a suffix of r_{n+1}. Moreover, l_n and r_n are reversals of one another (see for instance Chapter 2 of [19]). Let $x^* = \lim_{n\to\infty} l_n$ denote the characteristic word of slope $s(x)$, i.e., x^* is the unique word all of whose prefixes are left special factors of x.

Theorem 7. *Let $x \in \{0,1\}^{\mathbb{N}}$ be a Sturmian word. Then $\chi(x) = 2$.*

Proof. Fix a Sturmian word $x \in \{0,1\}^\omega$ of slope α. As x is aperiodic, it follows from Theorem 5 that $\chi(x) \geq 2$. Thus it suffices to show that $\mathrm{Fac}_x \subseteq S^2$ for some language $S \subseteq \{0,1\}^*$ with $\limsup p_S(n) = 2$. Set

$$S = \{\varepsilon\} \cup \{r_n 0 \,|\, n \geq 0\} \cup \{1l_n \,|\, n \geq 0\}.$$

Then S consists of precisely 2 words of each given length $n \geq 1$. It remains to show that $\mathrm{Fac}_x \subseteq S^2$.

Lemma 8. *Let x^* denote the characteristic Sturmian word of slope $s(x)$. Then for each $n \geq 0$ we have that $r_n 01x^*$ and $r_n 10x^*$ belong to $\mathcal{O}(x)$.*

Proof. Since each prefix of x^* is a left special factor of x it follows that each of $0x^*$ and $1x^*$ belong to $\mathcal{O}(x)$ and moreover each has a unique infinite left extension (otherwise x^* would be a suffix of itself and x^* would be periodic). We recall that $0x^*$ ($1x^*$, respectively) is the lexicographically smallest (largest, respectively) element of $\mathcal{O}(x)$ (see for instance [22]). It follows immediately from this lexicographic property that $0x^*$ must extend to the left by 1 and that $1x^*$ must extend to the left by 0 so that $10x^*, 01x^* \in \mathcal{O}(x)$. We now claim that $10x^*, 01x^*$ have the same infinite left extension i.e., there exists a left infinite word w such that each suffix of $w10x^*$ and of $w01x^*$ belongs to $\mathcal{O}(x)$. It follows immediately that each suffix of w is a right special factor of x. To establish the claim let z (z', respectively) be the unique infinite left extensions of $10x^*$ ($01x^*$, respectively). We will show that $z = z'$. Assume to the contrary that $z \neq z'$. Then there exists $u \in \mathbb{A}^*$ such that au is a suffix of z and bu is a suffix of z' where $\{a,b\} = \{0,1\}$. Since $au1$ and $bu0$ are each factors of x, by the balance property, we must have that $a = 0$ and $b = 1$. Since u is a left special factor of x we must have that either $u0$ or $u1$ is a prefix of x^*. If $u0$ is a prefix of x^*, then $1u0$ is a prefix of $1x^*$ and hence $1u0$ is the lexicographically largest factor of x to its length. It follows that $1u01x^*$ is lexicographically larger than $1x^*$, a contradiction. If on the other hand $u1$ is a prefix of x^*, then $0u1$ is a prefix of $0x^*$ and hence $0u1$ is the lexicographically smallest factor of x to its length. It follows that $0u10x^*$ is lexicographically smaller than $0x^*$, a contradiction. This completes the proof of Lemma 8. □

Returning to the proof of Theorem 7, it follows from Lemma 8 that the word $w(n) = r_{n-1}01l_{n-1}$ is a factor of x of length $2n$. We claim that for each $n \geq 1$, $w(n)$ contains $n+1$ distinct factors of length n. To prove the claim, we proceed by induction on n. For $n = 1$, we have $w(1) = 01$ which contains 2 factors of length 1. For the inductive step, let $n \geq 1$, and assume $w(n)$ contains $n+1$ distinct factors of length n. We wish to show that $w(n+1)$ contains $n+2$ distinct factors of length $n+1$. Suppose to the contrary that some word u of length $n+1$ occurs twice in $w(n+1)$. We claim $u = r_n 0$, for otherwise the word u' obtained by deleting the last letter of u would occur twice in $w(n)$, a contradiction. Similarly, if $u \neq 1l_n$, then the word u'' obtained by deleting the first letter of u would occur twice in

$w(n)$, a contradiction. Thus $u = r_n 1 = 0 l_n$, which is a contradiction since, as r_n and l_n are reversals of one another, $|r_n 1|_1 > |0 l_n|_1$. Having established the claim, it follows that each factor of x of length n is a factor of $w(n)$ and hence $\mathrm{Fac}_x \subseteq S^2$ as required. □

5. Potential generalisations

It would be interesting to determine the character of other classes of infinite words of sub-linear complexity including automatic words, words generated by primitive substitutions and Arnoux-Rauzy words. The precise determination of the character of a Sturmian word x consisted in showing that on one hand $\chi(x) \geq 2$ (Theorem 5) and on the other hand that $\chi(x) \leq 2$ (Lemma 8). Given a word x of sub-linear complexity, the proof of Theorem 1 in [6] yields an explicit method for constructing a set S of bounded complexity for which $\mathrm{Fac}_x \subseteq S^2$. This construction, which makes use of so-called *D-markers* (see §4.1 in [6]), yields an upper bound on the character of a given word of sub-linear complexity. The difficulty in general is in obtaining a lower bound on the character. For example, let $x \in \{a_1, a_2, \ldots, a_k\}^{\mathbb{N}}$ be an Arnoux-Rauzy word on a k-letter alphabet. Let $\tilde{x}$ denote the corresponding characteristic word, i.e., the unique element of the subshift generated by x each of whose prefixes is a left special factor of x. It follows that for each $n \geq 1$ there exist words $u_1(n), u_2(n), \ldots, u_k(n)$ of length n such that each $u_i(n)$ terminates in the letter a_i and $u_i(n)\tilde{x}$ belongs to the subshift generated by x. In other words, $u_i(n)\tilde{x}$ are the k-left extensions of length n of $\tilde{x}$. It is easy to see that each factor w of x may be written in the form $w = uv$ where u is either empty or $u = u_i(n)$ for some $1 \leq i \leq k$ and $n \geq 1$ and where v is either empty or a prefix of $\tilde{x}$. This decomposition of w is obtained by setting v (possibly empty) equal to the longest suffix of w which is a left special factor of x. It follows that $\mathrm{Fac}_x \subseteq S^2$ where

$$S = \{u_i(n) \,:\, 1 \leq i \leq k\,;\; n \geq 1\} \cup \mathrm{Pref}(\tilde{x}) \cup \{\varepsilon\}$$

and hence $\chi(x) \leq k + 1$. For example, if $\mathbf{t}$ is the Tribonacci word fixed by the substitution $a \mapsto ab$, $b \mapsto ac$, and $c \mapsto a$ we obtain $2 \leq \chi(\mathbf{t}) \leq 4$, where the lower bound is a consequence of Theorem 5. We note that in case x is Sturmian, the above upper bound yields $\chi(x) \leq 3$ which as we saw earlier is not optimal. Similarly, in the case of the Tribonacci word, a more careful analysis yields the upper bound $\chi(\mathbf{t}) \leq 3$. We suspect that in fact $\chi(\mathbf{t}) = 3$ but we do not know how to improve the lower found on the character. It would be reasonable to conjecture in general that if x is an Arnoux-Rauzy word on a k-letter alphabet then $\chi(x) = k$.

A similar analysis applies to the Thue-Morse word $x = 0110100110010110\cdots$ fixed by the uniform substitution $\varphi : 0 \mapsto 01$ and $1 \mapsto 10$. For each $n \geq 0$ let $t_n = \varphi^n(0)$ and $t'_n = \varphi^n(1)$ and let S be the set of all prefixes and all suffixes of t_n and t'_n for all $n \geq 0$. Since $t_{n+1} = t_n t'_n$ and $t'_{n+1} = t'_n t_n$ it follows that S consists of at most 4 words of each given length. Finally it is easy to see that $\mathrm{Fac}_x \subseteq S^2$ and hence $\chi(x) \leq 4$.

For words of higher than linear complexity, things become even more difficult. For example, in [6] it is shown that if x is a fixed point of the (non-primitive) morphism $a \mapsto ab$, $b \mapsto b$ and $c \mapsto ca$ (which is known to have quadratic complexity), then there does not exist a language S of bounded complexity such that $\mathrm{Fac}_x \subseteq S^3$. On the other hand it is shown that there exists a language S of bounded complexity such $\mathrm{Fac}_x \subseteq S^k$ for some $k \in \{4, 5, 6\}$. It is also shown in [6] that for every real number $\alpha \in (0, 1)$, there exists an infinite word x of complexity $O(n^{2+\alpha})$ such that Fac_x is not contained in S^k for any language S of bounded complexity and any positive integer k.

[1] J.-P. Allouche, M. Baake, J. Cassaigne and D. Damanik, *Palindrome complexity,* Selected papers in honor of Jean Berstel, Theoret. Comput. Sci. 292 (2003), 9–31.

[2] J.-P. Allouche, J. Shallit, *Automatic Sequences, Theory, Applications, Generalizations,* Cambridge University Press, 2003.

[3] V.I. Arnold: *Small denominators and problems of stability of motion in classical and celestial mechanics,* Usp. Math. Nauk. 18 (1963), 91–192, (in Russian) translated in Russian Math. Surveys 18 (1963), p. 86–194.

[4] P. Arnoux, G. Rauzy, *Représentation géométrique de suites de complexité* $2n + 1$, Bull. Soc. Math. France, 119 (1991), 199–215.

[5] J. Cassaigne, G. Fici, M. Sciortino, L.Q. Zamboni, *Cyclic complexity of words,* J. Combin. Theory Ser. A, 145 (2017), 35–56.

[6] J. Cassaigne, A. Frid, S. Puzynina, L.Q. Zamboni, *Cost and dimension of words of zero topological entropy,* preprint 2016, http://arxiv.org/abs/1607.04728.

[7] J. Cassaigne, A. Frid, S. Puzynina, L.Q. Zamboni, *Subword complexity and decomposition of the set of factors,* Proceedings of MFCS 2014, LNCS 8634, Springer, 147–158.

[8] E. Charlier, S. Puzynina, L.Q. Zamboni, *On a group theoretic generalization of the Morse-Hedlund theorem*, PAMS 145 (2017), 3381–3394.

[9] E. Coven, G. Hedlund, *Sequences with minimal block growth,* Math. Systems Theory, 7 (1973), 138–153.

[10] V. Cyr, B. Kra, *Complexity of short rectangles and periodicity,* European J. Combin., 52 (2016), 146–173.

[11] F. Durand, B. Host, C. Skau, *Substitutional dynamical systems, Bratteli diagrams and dimension groups,* Ergodic Theory & Dynam. Systems, 19 (1999), 953–993.

[12] F. Durand and M. Rigo, *Multidimensional extension of the Morse-Hedlund theorem,* European J. Combin., 34 (2013), 391–409.

[13] S. Ferenczi, *Rank and symbolic complexity,* Ergodic Theory & Dynam. Systems, 16 (1996), 663–682.

[14] A. Heinis, *Languages under substitutions and balanced words,* Journal de Théorie des Nombres de Bordeaux 16 (2004), 151–172.

[15] T. Kamae and L.Q. Zamboni, *Sequence entropy and the maximal pattern complexity of infinite words,* Ergodic Theory & Dynam. Systems, 22 (2002), 1191–1199.

[16] T. Kamae and L.Q. Zamboni, *Maximal pattern complexity for discrete systems,* Ergodic Theory & Dynam. Systems, 22 (2002), 1201–1214.

[17] J. Leroy, *Some improvements of the S-adic conjecture,* Adv. in Appl. Math., 48 (2012), 79–98.

[18] M. Lothaire, *Combinatorics on words,* Addison-Wesley Publishing Co., Reading, Mass., 1983.

[19] M. Lothaire, *Algebraic combinatorics on words,* Cambridge University Press, 2002.

[20] M. Morse, G. Hedlund, *Symbolic dynamics,* Amer. J. Math., 60 (1938), 815–866.

[21] M. Morse, G. Hedlund, *Symbolic dynamics II. Sturmian trajectories,* Amer. J. Math., 62 (1940), 1–42.

[22] G. Pirillo, *Inequalities characterizing standard Sturmian and episturmian words,* Theoret. Comput. Sci., 341 (2005), 276–292.

[23] S. Puzynina, L.Q. Zamboni, *Abelian returns in Sturmian words,* J. Combin. Theory Ser. A, 120 (2013), 390–408.

[24] G. Richomme, K. Saari and L.Q. Zamboni, *Abelian complexity of minimal subshifts,* J. Lond. Math. Soc. (2) 83 (2011), 79–95.

[25] J. Sander and R. Tijdeman, *The rectangle complexity of functions on two-dimensional lattices,* Theoret. Comput. Sci., 270 (2002), 857–863.