HAL
open science

Cost and dimension of words of zero topological entropy

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

To cite this version:

Julien Cassaigne, Anna E. Frid, Svetlana Puzynina, Luca Q. Zamboni. Cost and dimension of words of zero topological entropy. Bulletin de la société mathématique de France, 2019, 147 (4), pp.639-660. 10.24033/bsmf.2794 . hal-02375077

HAL Id: hal-02375077

https://hal.science/hal-02375077v1

Submitted on 29 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cost and dimension of words of zero topological entropy

Julien Cassaigne[a], Anna E. Frid[b], Svetlana Puzynina[c,1], Luca Q. Zamboni[d]

[a]*CNRS, Institut de Mathématiques de Marseille, France*
[b]*Aix-Marseille Université, Institut de Mathématiques de Marseille, France*
[c]*LIP, ENS de Lyon, Université de Lyon, France and Sobolev Institute of Mathematics, Novosibirsk, Russia*
[d]*Institut Camille Jordan, Université Lyon 1, France*

arXiv:1607.04728v1 [math.CO] 16 Jul 2016

Abstract

Let $\mathbb{A}^*$ denote the free monoid generated by a finite nonempty set $\mathbb{A}$. In this paper we introduce a new measure of complexity of languages $L \subseteq \mathbb{A}^*$ defined in terms of the semigroup structure on $\mathbb{A}^*$. For each $L \subseteq \mathbb{A}^*$, we define its *cost* $c(L)$ as the infimum of all real numbers α for which there exist a language $S \subseteq \mathbb{A}^*$ with $p_S(n) = O(n^\alpha)$ and a positive integer k with $L \subseteq S^k$. We also define the *cost dimension* $d_c(L)$ as the infimum of the set of all positive integers k such that $L \subseteq S^k$ for some language S with $p_S(n) = O(n^{c(L)})$. We are primarily interested in languages L given by the set of factors of an infinite word $x = x_0x_1x_2\cdots \in \mathbb{A}^\mathbb{N}$ of zero topological entropy, in which case $c(L) < +\infty$. We establish the following characterisation of words of linear factor complexity: Let $x \in \mathbb{A}^\mathbb{N}$ and $L = \mathrm{Fac}(x)$ be the set of factors of x. Then $p_x(n) = \Theta(n)$ if and only $c(L) = 0$ and $d_c(L) = 2$. In other words, $p_x(n) = O(n)$ if and only if $\mathrm{Fac}(x) \subseteq S^2$ for some language $S \subseteq \mathbb{A}^+$ of bounded complexity (meaning $\limsup p_S(n) < +\infty$). In general the cost of a language L reflects deeply the underlying combinatorial structure induced by the semigroup structure on $\mathbb{A}^*$. For example, in contrast to the above characterisation of languages generated by words of sub-linear complexity, there exist non factorial languages L of complexity $p_L(n) = O(\log n)$ (and hence of cost equal to 0) and of cost dimension $+\infty$. In this paper we investigate the cost and cost dimension of languages defined by infinite words of zero topological entropy. We establish the existence of words of cost zero and finite cost dimension having arbitrarily high polynomial complexity. In contrast we also show that for each $\alpha > 2$ there exist infinite words x of positive cost and of complexity $p_x(n) = O(n^\alpha)$.

Keywords: Symbolic dynamics, complexity.
2010 MSC: 37B10

Email addresses: `julien.cassaigne@math.cnrs.fr` (Julien Cassaigne), `anna.e.frid@gmail.com` (Anna E. Frid), `s.puzynina@gmail.com` (Svetlana Puzynina), `zamboni@math.univ-lyon1.fr` (Luca Q. Zamboni)

[1]Supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

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 $\mathrm{Fac}(x) = \{x_ix_{i+1}\cdots x_j \,|\, 0 \leq i \leq j\}$. First introduced by Hedlund and Morse in their seminal 1938 paper [13] under the name of *block growth*,[2] 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 [13] states that every aperiodic (meaning non-ultimately periodic) word contains at least $n+1$ distinct factors of each length n. 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 [14] on the classification of the factor complexities of 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 [9] 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. Thus given a language $L \subseteq \mathbb{A}^*$ (for instance consisting of all factors of some infinite word $x \in \mathbb{A}^{\mathbb{N}}$) one may ask whether L is contained in a finite product of the form S^k where S is a language of strictly lower complexity. Consider for example the *Thue-Morse infinite word*

$$x = 01101001100101101001 0\cdots$$

where for each $n \geq 0$, the n'th 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 [15, 16] 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 computed by Brlek [3] and independently by de Luca and Varricchio [7], is given by $p_x(1) = 2, p_x(2) = 4$ and for $n \geq 3$

$$p_x(n) = \begin{cases} 6 \cdot 2^{r-1} + 4q & 0 < q \leq 2^{r-1} \\ 2^{r+2} + 2q & 2^{r-1} < q \leq 2^r \end{cases}$$

where r and q are uniquely determined by the equation $n = 2^r + q + 1$, $r \geq 0$ and $0 < q \leq 2^r$. For each $n \geq 0$, let $t_n = \varphi^n(0)$ and $\bar{t}_n = \varphi^n(1)$. Then both t_n and $\bar{t}_n$ are factors of x of length

[2] In [8], Ehrenfeucht, Lee, and Rozenberg adopted the term *subword complexity.*

2^n. Let $S \subseteq \{0,1\}^*$ be the set consisting of all prefixes and suffixes (including ε) of t_n and $\bar{t}_n$ for each $n \geq 0$. Since $t_{n+1} = \varphi^{n+1}(0) = \varphi^n(01) = t_n\bar{t}_n$ and similarly $\bar{t}_{n+1} = \bar{t}_n t_n$, 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 $\varepsilon, 0$ and 1, we may suppose $|u| \geq 2$. Consider the least $n \geq 0$ such that u is a factor of t_{n+1} or a factor of $\bar{t}_{n+1}$. If u is a factor of t_{n+1}, by minimality of n we can write $u = vw$ where v is a non empty suffix of t_n and w a non empty prefix of $\bar{t}_n$. Whence $u \in S^2$. A similar argument applies in case u is a factor of $\bar{t}_{n+1}$. Thus while $\mathrm{Fac}(x)$ is of linear complexity, it is contained in a product S^2 where S is a language of bounded complexity. With some care, this construction may be generalized to fixed points of arbitrary (primitive) substitutions $\tau : \mathbb{A} \to \mathbb{A}^+$ by letting S be the collection of all prefixes and suffixes of all $\tau^n(u)$ $(n \geq 0)$ where u ranges over all factors of $\tau(a)$ for each $a \in \mathbb{A}$.

As another example, let $L = \mathrm{Fac}(x)$ denote the set of factors of a Sturmian word $x \in \{0,1\}^{\mathbb{N}}$ (for instance we may take $x = 0100101001001010010\cdots$ the Fibonacci word defined as the fixed point of the substitution $0 \mapsto 01, 1 \mapsto 0$). We recall that Sturmian words are infinite words having exactly $n+1$ distinct factors of each length $n \geq 1$. In view of the Morse-Hendlund theorem, Sturmian words 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. 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). In other words, $l_x(n)$ ($r_x(n)$, respectively) occurs in x preceded (followed, respectively) by both 0 and 1. See for instance Chapter 2 of [11]. Set $S = \{\varepsilon\} \cup \{r_x(n)0 \,|\, n \geq 0\} \cup \{1l_x(n) \,|\, n \geq 0\}$. Then S consists of precisely 2 words of each given length $n \geq 1$. One can prove that $\mathrm{Fac}(x) \subseteq S^2$ (see Example 3.4). It turns out that this is optimal in the sense that if x is an infinite word and $\mathrm{Fac}(x) \subseteq S^2$ for some language S with $\limsup p_S(n) \leq 1$, then x is ultimately periodic (see [17]).

The above examples are only special cases of the following more general result proved herein:

Theorem 1. *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 language $S \subseteq \mathbb{A}^*$ of bounded complexity (i.e., $\limsup p_S(n) < +\infty$).*

It turns out that Theorem 1 is very specific to languages defined by infinite words. In fact, there exist non-factorial languages L of complexity $p_L(n) = O(\log n)$ which are not contained in any finite product of the form S^k where S is a language of bounded complexity and k a positive integer.

Our aim here is to express and study these ideas in greater generality. Given a language L of low complexity, meaning $\limsup_{n\to\infty} \frac{\log p_L(n)}{n} = 0$, we define the *cost* of L, denoted $c(L)$, as the infimum of all real numbers α for which there exist a language S with $p_S(n) = O(n^\alpha)$ and a positive integer k such that $L \subseteq S^k$. More precisely, for each real number $\alpha \in [0, +\infty)$, we define

the α-*dimension* $d_\alpha(L)$ by

$$d_\alpha(L) = \inf\{k \geq 1 \mid L \subseteq S^k \text{ for some language } S \in \mathcal{L}(\alpha)\},$$

where $\mathcal{L}(\alpha)$ denotes the collection of all languages $L \subseteq \mathbb{A}^*$ (over some finite non empty alphabet $\mathbb{A}$) whose complexity $p_L(n) = O(n^\alpha)$. If x is an infinite word and $L = \mathrm{Fac}(x)$, then, by the Morse-Hedlund theorem, L belongs to $\mathcal{L}(0)$ if and only if x is ultimately periodic. While if x is a Sturmian word or if x is generated by a primitive substitution, then L belongs to $\mathcal{L}(1)$. Then the cost $c(L)$ is given by

$$c(L) = \inf\{\alpha \in [0, +\infty) \mid d_\alpha(L) < +\infty\}.$$

In each case above we take the convention that $\inf \emptyset = +\infty$. If $c = c(L) < +\infty$, then we call $d_c(L) \in \{1, 2, 3, \ldots\} \cup \{\infty\}$ the *cost dimension* of L. In case $L = \mathrm{Fac}(x)$ for some infinite word x, then we write $c(x)$ ($d_c(x)$, respectively) in lieu of $c(L)$ ($d_c(L)$, respectively). Thus, the Morse-Hedlund theorem states that an infinite word $x \in \mathbb{A}^\mathbb{N}$ is ultimately periodic if and only if $c(x) = 0$ and $d_0(x) = 1$, i.e., x is of cost equal to 0 and cost dimension equal to 1. Similarly, Theorem 1 asserts that x is of linear complexity (i.e., $p_x(n) = \Theta(n)$) if and only if x is of cost equal to 0 and cost dimension equal to 2. The above definitions may be adapted to other measures of complexity as we do herein for the so-called *accumulative complexity* $p^*_L(n)$ which counts the number of words in L of length less than or equal to n.

A fundamental question, to which a substantial portion of the paper is devoted, is to what extent does the complexity of a language determine its cost and cost dimension and vice versa. A first basic observation is that languages L of positive entropy have cost equal to $+\infty$. For this reason we restrict our attention to languages and words of zero topological entropy. Via a straightforward counting argument, it is shown that for each $\alpha \geq 0$, if $d_\alpha(L) = k$ for some $1 \leq k < +\infty$, then $L \in \mathcal{L}(k(\alpha + 1) - 1)$. It follows from this that $c(L)$ is finite if and only if the complexity of L is bounded above by a polynomial. We further show by direct construction that for each positive integer $k \geq 1$ there exists an infinite word x of complexity $p_x(n) \in \Omega(n^{k-1})$ with $d_0(x) = k$. In other words, we establish the existence of words of cost zero and of arbitrarily high polynomial complexity.

Conversely, given the complexity of a language, what can be said of its cost and cost dimension. We already mentioned two results in this direction: first the obvious fact that for languages L of bounded complexity we have $d_0(L) = 1$. Second, that if L is the set of factors of an aperiodic infinite word, then L is of linear complexity if and only if its cost $c(L) = 0$ and its cost dimension $d_0(L) = 2$. However in general, the cost and cost dimension of a given language depend only in part on its complexity. In fact, both reflect deeply the underlying combinatorial structure of the language. For instance, we already mentioned that non-factorial languages are in general very far from satisfying any result along the lines of Theorem 1. But even in the case of languages defined

by infinite words, the characterisation of Theorem 1 does not seem to extend nicely to higher complexities. For instance, we prove that the word $x = \prod_{i=1}^{\infty} ab^i = ababbabbb\cdots$ generated by the (non-primitive) substitution $a \mapsto ab, b \mapsto b, c \mapsto ca$, considered by Pansiot in [14] and of complexity $p_x(n) = \Theta(n^2)$, verifies $d_0(u) > 3$. On the other hand we also show that $d_0(x) \leq 6$ which in particular implies is of cost zero. We do not know whether there exist words of sub-quadratic complexity and positive cost. However, we prove that for every real number $\alpha \in (0, 1)$ there exists an infinite word x with complexity $p_x(n) \in O(n^{2+\alpha})$ and cost $c(x) \geq \alpha$. In other words, there exist words of positive cost having relatively low (sub-cubic) complexity. This should be contrasted with the result mentioned earlier on the existence of words of arbitrarily high polynomial complexity having cost equal to zero. These results suggest that the cost of a word measures something beyond its factor complexity which makes it of independent interest.

The paper is structured as follows: In §2 we briefly recall some of the basic terminology and notions arising in the study of infinite words. For a more detailed exposition, the reader is referred to one of the standard texts in combinatorics on words such as the Lothaire books [10, 11, 12]. Also in §2, 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 and yet relevant in what follows, in particular used 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 and local periods. In §3 we define the key notions of cost and cost dimension of a language in the context of the factor complexity as well as the accumulative complexity. Also in this section we establish various fundamental results linking the cost of a language to its complexity and relations between the cost $c(L)$ defined in terms of the factor complexity and the cost $c^*(L)$ defined in terms of the accumulative complexity. In §4 we study the cost and cost dimension of words of sub-linear complexity. We begin §4 by introducing 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. This decomposition enables us to obtain what we regard to be the main result of the paper (see Theorem 4.7), and which gives a complete characterisation of words of linear complexity in terms of cost and cost dimension: An infinite word x is of linear complexity, i.e., $p_x(n) = \Theta(n)$ if and only if the cost $c(x) = 0$ and the cost dimension $d_0(x) = 2$. Theorem 4.7 is actually a consequence of a more general result given by Theorem 4.4 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 §5 we study the cost and cost dimension of words of sub-quadratic complexity. We begin the section with another consequence of Theorem 4.4 which yields a non-trivial bound on the cost of words x of complexity $p_x(n) = O(n^\alpha)$ for $\alpha \in (1, 2)$. We estimate the cost complexity of the fixed point x of the substitution $a \mapsto ab, b \mapsto b, c \mapsto ca$ which is known to have quadratic complexity and prove that $4 \leq d_0(x) \leq 6$. In particular this shows that the result of Theorem 4.7 already breaks down for words of quadratic complexity. In §6 we investigate the cost and cost dimension of words of greater than quadratic complexity and prove

that every real number $\alpha \in (0,1)$ there exists an infinite word x with complexity $p_x(n) \in O(n^{2+\alpha})$ and cost $c(x) \geq \alpha$ (see Corollary 6.3). Finally in §7 we exhibit an example of a non-factorial language L of complexity $p_L(n) = O(\log n)$ (and hence of cost zero) having infinite cost dimension i.e., $d_0(L) = +\infty$. This is yet another illustration of how the main result of Theorem 4.7 depends strongly on the assumption that the language L be defined by an infinite 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 more details we refer the reader to [11]. We also introduce the new notions of internal and extremal occurrences of factors in finite and infinite words which are defined by their virtual occurrences and local periods.

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 $\mathbb{A}^+ = \mathbb{A}^* \setminus \{\varepsilon\}$. For each $u \in \mathbb{A}^*$ and $a \in \mathbb{A}$, we let $|u|_a$ denote the number of occurrences of a in u. For $u = u_0u_1 \cdots u_{n-1} \in \mathbb{A}^+$ we define

$$\mathrm{Fac}(u) = \{u_i \cdots u_j : 0 \leq i \leq j \leq n-1\} \cup \{\varepsilon\}.$$

A subset $L \subseteq \mathbb{A}^*$ is called a *language*. 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

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

and its accumulative complexity $p_L^* : \mathbb{N} \to \mathbb{N}$ by

$$p_L^*(n) = \sum_{i=0}^{n} p_L(i).$$

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 if x. We will frequently use the notation $x[i,j]$ for $x_i \cdots x_j$. 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 $a, b \in \mathbb{A}$. Let $p_x : \mathbb{N} \to \mathbb{N}$ (resp., $p_x^* : \mathbb{N} \to \mathbb{N}$) denote the *factor complexity* (resp., *accumulative factor complexity*) of x defined by:

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

and

$$p_x^*(n) = \sum_{i=0}^{n} p_x(i).$$

We say $x \in \mathbb{A}^{\mathbb{N}}$ (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 non-empty words $u, v \in \mathbb{A}^*$. An infinite word is said to be *aperiodic* if it is not ultimately periodic. It follows that 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 periodic if it is eventually periodic to both the left and the right, i.e., if x admits a prefix of the form $\cdots uuu$ and a suffix of the form $vvv\cdots$ for some $u, v \in \mathbb{A}^+$. Otherwise x is said to be aperiodic.

Definition 2.1. Let $u = u_1u_2\cdots u_n$ and v belong to $\mathbb{A}^+$ and fix $1 \le i \le n$. We say there there is a *virtual occurrence* of v in u beginning (ending, respectively) at position i if the shorter of v and $u_i\cdots u_n$ ($u_1\cdots u_{i-1}$, respectively) is a prefix (suffix, respectively) of the other. That is $v\mathbb{A}^* \cap u[i,n]\mathbb{A}^* \neq \emptyset$ ($\mathbb{A}^*v \cap \mathbb{A}^*u[1,i-1] \neq \emptyset$, respectively).

Definition 2.2. For $u = u_1u_2\cdots u_n$ and $1 \le i \le n$, 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 at position $i = 3$ (witnessed by $v = 01$) as well as a virtual square of length 3 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 at position 1, and of lengths 3 and 5 at position 2.

Definition 2.3. For $v = v_1v_2\cdots v_n \in \mathbb{A}^+$. Define the (least) *period* of v, denoted $\pi(v)$, to be the least positive integer m such that $v_i = u_{i+m}$ for all $1 \le i \le n-m$.

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 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 a virtual square 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 virtual square of length $\pi(v)$ at position $i+n$. For instance, if $x = 01001010100\cdots$, then the occurrence of $v = 010$ at position 0 is not initial since x has a virtual square of length $2 = \pi(v)$ centered at position 0. Instead this occurrence is final (even if it is immediately followed by another occurrence of v) since x does not 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 aperiodic, then each factor v of x admits a final occurrence in x.

Throughout the paper we make use of the usual Landau notations O, Ω, Θ, and o. We adopt the following definition of Ω which is more commonly used in computer science: Given functions $f, g : \mathbb{N} \to \mathbb{R}^+$, we write

$$f(n) = \Omega(g(n)) \text{ if } \quad \exists K > 0,\, \exists N,\, \forall n \geq N :\, f(n) \geq K g(n).$$

3. Dimension and cost: definitions, examples and general properties

For each real number $\alpha \in [0, +\infty)$, we denote by $\mathcal{L}(\alpha)$ (resp., $\mathcal{L}^*(\alpha)$) the collection of languages $L \subseteq \mathbb{A}^*$ (over some finite non empty alphabet $\mathbb{A}$) with $p_L(n) = O(n^\alpha)$ (resp., $p_L^*(n) = O(n^\alpha)$). Analogously, we denote by $\mathcal{W}(\alpha)$ (resp., $\mathcal{W}^*(\alpha)$) the collection of infinite words $x \in \mathbb{A}^\mathbb{N}$ (over some finite non empty alphabet $\mathbb{A}$) such that $\mathrm{Fac}(x) \in \mathcal{L}(\alpha)$ (resp., $\mathrm{Fac}(x) \in \mathcal{L}^*(\alpha)$). The set $\mathbb{A}^*$ is considered as a free monoid, and thus for each $S \subseteq \mathbb{A}^*$ the set S^k is just the set of all concatenations of k elements of S.

Definition 3.1. Let $L \subseteq \mathbb{A}^*$. For each real number $\alpha \in [0, +\infty)$, we define the *α-dimension* $d_\alpha(L)$ by

$$d_\alpha(L) = \inf\{k \geq 1 \mid L \subseteq S^k \text{ for some language } S \in \mathcal{L}(\alpha)\},$$

and the *cost* $c(L)$ by

$$c(L) = \inf\{\alpha \in [0, +\infty) \mid d_\alpha(L) < +\infty\}.$$

If $c = c(L) < +\infty$, we call $d_c(L) \in [1, +\infty]$ the *cost dimension* of L.

By convention $\inf \emptyset = +\infty$. Definition 3.1 extends naturally to infinite words $x \in \mathbb{A}^\mathbb{N}$ by replacing L by $\mathrm{Fac}(x)$ so we define accordingly $d_\alpha(x)$ and $c(x)$. Replacing $\mathcal{L}(\alpha)$ by $\mathcal{L}^*(\alpha)$ we define analogously the *α-accumulative dimension* $d_\alpha^*(L)$ and the *accumulative cost* $c^*(L)$.

We observe that in our definition of $d_\alpha(L)$, we may replace S^k by $S_1 \cdots S_k$ for some languages $S_1, \ldots, S_k \in \mathcal{L}(\alpha)$. The following lemma is an immediate consequence of the definition:

Lemma 3.2. *Suppose $L \in \mathcal{L}(\alpha_0)$ (resp., $L \in \mathcal{L}^*(\alpha_0)$) for some $\alpha_0 \geq 0$. Then $d_\alpha(L) = 1$ (resp., $d_\alpha^*(L) = 1$) for each $\alpha \geq \alpha_0$ and hence $c(L) \leq \alpha_0$ (resp., $c^*(L) \leq \alpha_0$).*

Lemma 3.3. *For each language $L \subseteq \mathbb{A}^*$, we have $d_0(L) = 1$ if and only if L is of bounded complexity. For each infinite word $x \in \mathbb{A}^\mathbb{N}$, we have $d_0(x) = 1$ if and only if x is ultimately periodic.*

Proof. The first statement is clear from Definition 3.1. As for the second, if x is ultimately periodic, then its complexity is bounded, whence $d_0(x) = 1$. Conversely if $d_0(x) = 1$, then the complexity of x is bounded, and hence by the Morse-Hedlund theorem x is ultimately periodic. $\square$

Example 3.4 (Sturmian words)**.** Here we prove that for every Sturmian word x we have $d_0(x) = 2$. To see this, we show that for each Sturmian word $x \in \{0,1\}^{\mathbb{N}}$, there exist sets S, T with $p_S(n), p_T(n) \equiv 1$ (for each $n \geq 0$) such that $\mathrm{Fac}(x) \subseteq ST$. Combined with Lemma 3.3, this implies that $d_0(x) = 2$. 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, $p_S(n), p_T(n) \equiv 1$. It remains to show that $\mathrm{Fac}(x) \subseteq ST$. To this end we 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 [2]). We claim that for each $n \geq 1$, $w(n)$ contains $n + 1$ distinct factors of length n. Assuming for a moment this claim, it follows that each factor of x of length n is a factor of $w(n)$ and hence $\mathrm{Fac}(x) \subseteq ST$ as required. 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 $u = r_x(n)0$, for otherwise the word u' obtained by deleting the last letter of u would occur twice in $w(n)$, a contradiction. Similarly, if $u \neq 1l_x(n)$, then the word u'' obtained by deleting the first letter of u would occur twice in $w(n)$, a contradiction. Thus $u = r_x(n)1 = 0l_x(n)$, which is impossible since, as $r_x(n)$ and $l_x(n)$ are reversals of one another, we have that $r_x(n)1$ and $0l_x(n)$ do not contain the same number of $0'$s and $1'$s.

The next proposition illustrates the basic relations between the dimension d_α and the accumulative dimension d^*_α. It is stated in terms of languages $L \subseteq \mathbb{A}^*$ but the same inequalities hold for infinite words $x \in \mathbb{A}^{\mathbb{N}}$.

Proposition 3.5. *For each $\alpha \geq 0$ and language $L \subseteq \mathbb{A}^*$ we have*

1. $d_\alpha(L) \leq d^*_\alpha(L)$,
2. $d^*_{\alpha+1}(L) \leq d_\alpha(L) \leq 2d^*_{\alpha+1}(L)$.

Proof. We begin by showing that $d_\alpha(L) \leq d^*_\alpha(L)$. The result is clear if $d^*_\alpha(L) = +\infty$. Thus assume $d^*_\alpha(L) = k$ for some positive integer k. Then $L \subseteq S^k$ for some language $S \in \mathcal{L}^*(\alpha)$. Hence $S \in \mathcal{L}(\alpha)$ whence $d_\alpha(L) \leq k = d^*_\alpha(L)$ as required. Next we show that $d^*_{\alpha+1}(L) \leq d_\alpha(L)$. Again the result is clear if $d_\alpha(L) = +\infty$, thus we may suppose $d_\alpha(L) = k$ for some positive integer k. Then $L \subseteq S^k$ for some language $S \in \mathcal{L}(\alpha)$. In other words, $p_S(n) = O(n^\alpha)$. Thus $p^*_S(n) = O(n^{\alpha+1})$, i.e., $S \in \mathcal{L}^*(\alpha + 1)$, and hence $d^*_{\alpha+1}(L) \leq k = d_\alpha(L)$. In order to prove the remaining inequality, we will need the following lemma:

Lemma 3.6. *Let $T \subseteq \mathbb{A}^*$. If $T \in \mathcal{L}^*(\alpha+1)$, then $T \subseteq S^2$ for some $S \in \mathcal{L}(\alpha)$.*

Proof. Since $T \in \mathcal{L}^*(\alpha+1)$, there exists a constant $K > 0$ such that $p_T^*(n) \leq Kn^{\alpha+1}$ for each $n \geq 1$. We order $T = \{v_1, v_2, v_3, \ldots\}$ so that $|v_m| \leq |v_{m+1}|$ for each $m \geq 1$. Thus for each $m \geq 2$ we have

$$m \leq p_T^*(|v_m|) \leq K|v_m|^{\alpha+1}. \tag{1}$$

(For $m = 1$, we may have $v_1 = \varepsilon$, and thus the latter inequality will not hold.)

Pick M such that

$$M > \max\{K(\alpha+1)2^{\alpha+2}; 2\}.$$

We now show that there exists a language $S \subset \mathbb{A}^*$ with $p_S(n) \leq \lceil Mn^\alpha \rceil$ for each $n \geq 1$, and $T \subseteq S^2$. To prove this we define inductively a nested sequence of sets $S_1 \subseteq S_2 \subseteq S_3 \subseteq \cdots$ with $S_m \subseteq \mathbb{A}^*$ such that for each $m \geq 1$ the following three conditions are satisfied:

i) $\mathrm{Card}(S_m) \leq 2m$,

ii) $p_{S_m}(n) \leq \lceil Mn^\alpha \rceil$ for each $n \geq 1$,

iii) $\{v_1, v_2, \ldots, v_m\} \subseteq S_m^2$.

For $m = 1$, we consider the factorization $v_1 = \varepsilon \cdot v_1$ and put $S_1 = \{\varepsilon, v_1\}$. Then clearly S_1 satisfies each of the conditions i), ii) and iii) above. For the inductive step, suppose for $m \geq 1$ we have constructed sets $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_m$ with the required properties. We say that $n \geq 1$ is a forbidden length if $p_{S_m}(n) = \lceil Mn^\alpha \rceil$, i.e., in constructing S_{m+1} from S_m we cannot add to S_m any word of forbidden length without violating condition ii) at level $m+1$. Note that 0 is never a forbidden length since there exists only one word of length 0, ε, and nothing else can be added to the set of words of length 0.

Let F denote the set of all forbidden lengths. For each $0 \leq i \leq |v_{m+1}|$ we can factor v_{m+1} as $v_{m+1} = x_i y_i$, with $|x_i| = i$. We claim that there exists $0 \leq j \leq \left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1$ such that neither $|x_j|$ nor $|y_j|$ belongs to F.So, we can take $S_{m+1} = S_m \cup \{x_j, y_j\}$. To prove the claim, suppose to the contrary that for each $0 \leq i \leq \left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1$ there exists $n_i \in \{i, |v_{m+1}| - i\} \cap F$. Then summing up the number of elements in S_m of forbidden lengths we obtain:

$$\mathrm{Card}(S_m) \;\geq\; \sum_{n \in F} \lceil Mn^\alpha \rceil \geq \sum_{i=0}^{\left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1} \lceil Mn_i^\alpha \rceil \geq \sum_{i=0}^{\left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1} Mn_i^\alpha \geq \sum_{i=1}^{\left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1} Mi^\alpha + M|v_{m+1}|^\alpha.$$

The latter inequality holds since 0 is never a forbidden length, and thus $n_0 = |v_{m+1}|$. Continuing

the chain of inequalities, we see that

$$\begin{aligned}
\mathrm{Card}(S_m) &\geq \sum_{i=1}^{\left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1} Mi^\alpha + M|v_{m+1}|^\alpha > \sum_{i=1}^{\left\lceil \frac{|v_{m+1}|}{2} \right\rceil} Mi^\alpha \geq \int_0^{\frac{|v_{m+1}|}{2}} Mx^\alpha \, dx \\
&\geq \frac{M}{(\alpha+1)} \left(\frac{|v_{m+1}|}{2} \right)^{\alpha+1} > \frac{K(\alpha+1)2^{\alpha+2}}{(\alpha+1)} \left(\frac{|v_{m+1}|}{2} \right)^{\alpha+1} \\
&\geq 2K|v_{m+1}|^{\alpha+1} \geq 2K|v_m|^{\alpha+1} \geq 2m,
\end{aligned}$$

where the last inequality follows from (1)), contradicting i). This completes the inductive step. Having defined the nested sequence $(S_m)_{m\geq 1}$, we set $S = \bigcup_{m\geq 1} S_m$. Then $p_S(n) = O(n^\alpha)$ and $T \subseteq S^2$. □

We now return to the proof of Proposition 3.5 and establish the remaining inequality $d_\alpha(L) \leq 2d^*_{\alpha+1}(L)$. Let us assume $d^*_{\alpha+1}(L) = k$ for some positive integer k. Then $L \subseteq T^k$ for some $T \in \mathcal{L}^*(\alpha+1)$. By Lemma 3.6 there exists $S \in \mathcal{L}(\alpha)$ such that $T \subseteq S^2$. Thus $L \subseteq S^{2k}$ whence $d_\alpha(L) \leq 2k = 2d^*_{\alpha+1}(L)$ as required. □

The next statement follows immediately from the second double inequality of Proposition 3.5.

Corollary 3.7. *For any language $L \subseteq \mathbb{A}^*$,*

1. *if $c(L) > 0$, then $c^*(L) = c(L) + 1$;*
2. *if $c(L) = 0$, then $0 \leq c^*(L) \leq 1$.*

The next proposition establishes a first relationship between d_α and complexity:

Proposition 3.8. *Let $\alpha \geq 0$ and $L \subseteq \mathbb{A}^*$. If $d_\alpha(L) = k$ for some positive integer k, then $L \in \mathcal{L}(k(\alpha+1)-1)$. In particular, if $x \in \mathbb{A}^\mathbb{N}$ and $L = \mathrm{Fac}(x)$, then by taking $\alpha = 0$ we have that if $d_0(x) = k$, then $x \in \mathcal{W}(k-1)$.*

Proof. It suffices to prove the proposition for languages L. The result is clear in case $k = 1$. So let us fix $k \geq 2$, and let $L \subseteq S^k$ for some $S \in \mathcal{L}(\alpha)$. Then there exists a positive integer C such that $p_S(n) \leq Cn^\alpha$ 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 $C|v_i|^\alpha$ choices for each v_i. Thus $p_L(n) = O(n^{k(\alpha+1)-1})$ as required. □

As an immediate consequence we get:

Corollary 3.9. *For each language $L \subseteq \mathbb{A}^*$ (resp., infinite word $x \in \mathbb{A}^{\mathbb{N}}$) we have $c(L) < +\infty$ if and only if $L \in \mathcal{L}(\alpha)$ (resp., $x \in \mathcal{W}(\alpha)$) for some $\alpha \geq 0$.*

Proof. If $c(L) < +\infty$, then $d_\alpha(L) < +\infty$ for each $\alpha > c(L)$. Fix $\alpha > c(L)$ and a positive integer k such that $d_\alpha(L) = k$. Then by Proposition 3.8 $L \in \mathcal{L}(k(\alpha+1)-1)$. The converse follows from Lemma 3.2. □

In view of the next corollary, we restrict ourselves henceforth to languages and words of entropy zero.

Corollary 3.10. *Languages of positive entropy have cost equal to $+\infty$.*

Proposition 3.8 suggests that a priori there is no polynomial bound on the complexity of infinite words of cost equal to 0. The following proposition shows that for each $k \geq 1$ there exists a word x of complexity $\Omega(n^{k-1})$ with $d_0(x) = k$ and hence in particular $c(x) = 0$.

Proposition 3.11. *For each $k \geq 1$ there exists a word x of complexity $\Omega(n^{k-1})$ of cost 0 and cost dimension k.*

Proof. For $k = 1$ we may simply take the constant word $x = a^\omega$, and for $k = 2$ it suffices to take x to be any Sturmian word (see Example 3.4). Thus we may assume that $k \geq 3$. We construct a word x on the alphabet $\{0, 1, \ldots, k-2\}$ as follows: We enumerate

$$\{1, \ldots, k-2\}^+ = \{t_1, t_2, t_3, \ldots\}$$

where the t_i are listed in increasing order, where $\{1, \ldots, k-2\}^+$ is ordered by $t_i < t_j$ if and only if either $|t_i| < |t_j|$ or in case $|t_i| = |t_j|$ then t_i is less than t_j relative to the lexicographic order. So the sequence $t_1, t_2, \ldots$ looks like $1, 2, \ldots, k-2, 11, 12, \ldots$ Then $x \in \{0, 1, \ldots, k-2\}^{\mathbb{N}}$ is defined by

$$x = t_0 t_1 t_0 t_2 t_0 t_1 t_0 t_3 \ldots,$$

where $t_0 = 0$. In other words x is obtained as the limit of a sequence (w_n) defined by $w_0 = t_0$, $w_{n+1} = w_n t_{n+1} w_n$ for all $n \geq 0$. We claim that the complexity of x is $\Omega(n^{k-1})$. Indeed, let us restrict ourselves to factors of x of length n which contain a complete factor $0t_p0$, where the length of t_p is at least $n/2$. Such a factor of x exists for each $t_p = 1^{j_1}2^{j_2}\ldots(k-2)^{j_{k-2}}$ (that is, for each $j_1, \ldots, j_{k-2}$ under the condition $j_1 + \cdots + j_{k-2} = |t_p| \geq n/2$), and for each starting point of that occurrence of t_p, which is any number between 1 and $n - |t_p| - 1$. So, we have $k-1$ degrees of freedom, and thus the complexity of x is at least $O(n^{k-1})$. On the other hand, take a factor w of x and find in it a word t_p, where p is maximal possible. Here incomplete intersections count: we just fix an occurrence of w to x, see what words t_p it intersects and choose the greatest p. If

t_p is completely in w, it is followed in it by a prefix of x. Denote the set of prefixes of x by S_{k-1}. Symmetrically, just before t_p in w, if it is taken from the beginning, there is a suffix of some word w_m (and w_m are suffixes one of another). We denote the set of these suffixes by S_0. As for t_p itself, it belongs to the concatenation of $1^* = S_1$, $2^* = S_2$, etc.; so,

$$w \in S_0 S_1 \dots S_{k-2} S_{k-1}, \tag{2}$$

where the complexity of each S_i is 1.

If t_p is not completely contained in w, three situations are possible. Either $w = t's$, where t' is a suffix of t_p; then $t' \in i^*(i+1)^* \cdots (k-2)^*$ for some $i \in \{1, \dots, k-2\}$, s is a prefix of x, and thus $w \in S_i \cdots S_{k-2} S_{k-1} \subset S_0 S_1 \dots S_{k-2} S_{k-1}$. Or, symmetrically, $w = pt''$, where t'' is a prefix of t_p; then $t'' \in 1^* 2^* \cdots i^*$ for some $i \in \{1, \dots, k-2\}$, p is a suffix of some w_m, and thus $w \in S_0 S_1 \cdots S_i \subset S_0 S_1 \dots S_{k-2} S_{k-1}$. Or, at last, w is a factor of t_p, and then $w \in i^*(i+1)^* \cdots j^*$ for some $i, j \in \{1, \dots, k-2\}$, $i \leq j$, and thus $w \in S_i S_{i+1} \cdots S_j \subset S_0 S_1 \dots S_{k-2} S_{k-1}$. In all the cases, (2) holds. □

While the definition of x in the previous proposition is on a alphabet size which varies with k, by applying to x the morphism $f : i \to 1^i 0^{k-i}$ we obtain an infinite binary word satisfying the same required properties.

We end this section by noting that the set S in Definition 3.1 is not assumed to be factorial. In fact, as the following proposition shows, this is too strong of a condition:

Proposition 3.12. *Let $x \in \mathbb{A}^{\mathbb{N}}$. Suppose $\mathrm{Fac}(x) \subseteq S^k$ for some factorial language S and positive integer k. Then there exists a suffix y of x such that $\mathrm{Fac}(y) \subseteq S$. In particular, for each positive integer $\alpha \geq 0$, if $S \in \mathcal{L}(\alpha)$ then $x \in \mathcal{W}(\alpha)$.*

Proof. We remark that if S is factorial, then so is S^k for each $k \geq 1$. Let $k \geq 1$ be the least positive integer such that $\mathrm{Fac}(x) \subseteq S^k$. The result is clear in case $k = 1$, so we may suppose $k > 1$. By minimality of k, there exists a factor u of x not belonging to S^{k-1}. Pick $y \in \mathbb{A}^{\mathbb{N}}$ such that uy is a suffix of x. We claim $\mathrm{Fac}(y) \subseteq S$. Since S is factorial, it suffices to show that every prefix of y belongs to S. So let $z \in \mathbb{A}^*$ be a prefix of y. Then we can write $uz = v_1 v_2 \cdots v_k$ for some $v_i \in S$. Since S^{k-1} is factorial and $u \notin S^{k-1}$, it follows that $v_1 v_2 \cdots v_{k-1}$ is a proper prefix of u and hence z is a proper suffix of v_k. Thus $z \in S$ as required. □

4. A characterisation of words of linear complexity in terms of cost dimension

In this section we characterize words of linear complexity in terms of the cost dimension. Let $x \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^{\mathbb{Z}}$. For each $n \geq 0$, let $\mathcal{R}_x(n)$ denote the set of right special factors of x of length n and $\mathcal{R}_x = \bigcup_{n \geq 0} \mathcal{R}_x(n)$.

Definition 4.1. Let D be a positive integer. A subset $M \subseteq \mathbb{A}^*$ is called a *D-marker set* for x if for each $n \geq 1$ and each factor u of x of length $|u| \geq Dn$ we have $\mathrm{Fac}(u) \cap M \cap \mathbb{A}^n \neq \emptyset$. The elements of M are called D-markers.

Lemma 4.2. *Let C be a positive integer. Then for each aperiodic word $x \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^{\mathbb{Z}}$ 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 periodic, a contradiction. It follows that the suffix w'' of w' of length n belongs to $\mathcal{R}_x(n)$. □

The following proposition gives an alternative and more general method for constructing marker sets whose complexity is related to the complexity of the underlying word:

Proposition 4.3. *For each aperiodic word $x \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^{\mathbb{Z}}$ 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 $M = \cup_{n \geq 1} M_n$. 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, 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 $2n - j' + 1 \geq n - j + 1$ 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

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

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. □

We now state and prove the most general result of this section.

Theorem 4.4. *Assume either $y \in \mathbb{A}^{\mathbb{Z}}$, or $y \in \mathbb{A}^{\mathbb{N}}$ and is recurrent. Let D be a positive integer and assume that M is a D-marker set for y. Then there exist languages $S, T \subseteq \mathbb{A}^*$ such that $\mathrm{Fac}(y) \subseteq ST$ and for each $n \geq 2D$ we have*

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

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

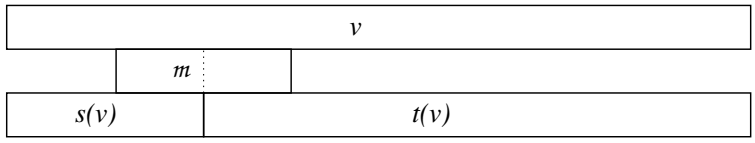

Figure 1: Building elements of S and T from a word v and an occurrence of a marker to it

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

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

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

It follows immediately from the definitions that $\mathrm{Fac}(y) \subseteq ST$. It remains to show that complexities of S and T satisfy (3). 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

$$\frac{n}{2D} < 2^k \leq 2n, \tag{4}$$

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 (4) 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 in which 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

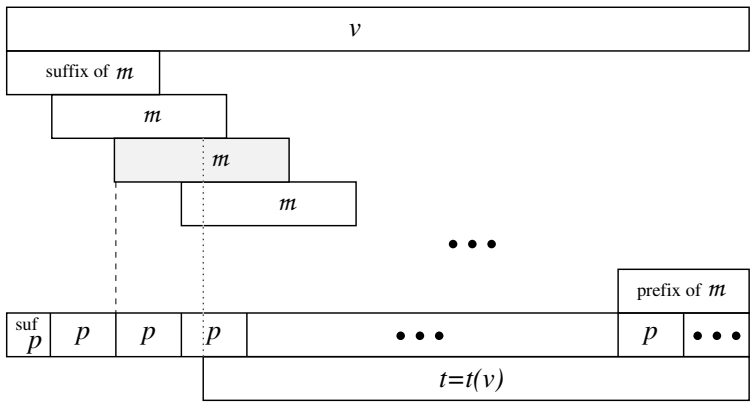

Figure 2: The case of an internal occurrence, the unique t is determined by $\mathfrak{m}$ and the length

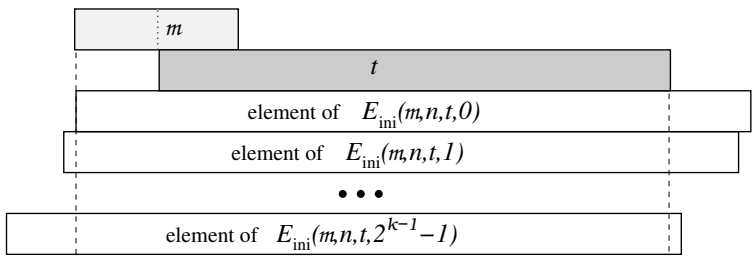

Figure 3: The sets $E_{\text{ini}}(\mathfrak{m}, n, t, i)$. The parts between dashed lines are common for all elements

$\mathfrak{m}$ 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 $\mathrm{Card}(T_{\text{int}}(\mathfrak{m}, n)) = 1$.

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

Lemma 4.5. *For each $n \geq 2D$ we have*

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

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

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:

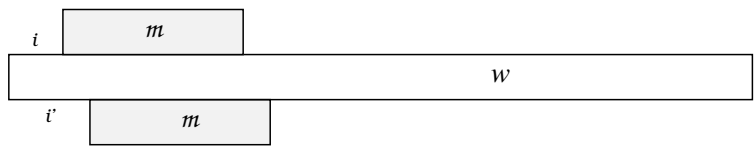

Figure 4: Proof of Claim 4.5.1. The lower occurrence of $\mathfrak{m}$ is not initial

Claim 4.5.1. *For each $t, t' \in T_{\text{ini}}(\mathfrak{m}, n)$ and $0 \leq i, i' < 2^{k-1}$, where $t \neq t'$ or $i < i'$, we have*

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

Proof of Claim 4.5.1. 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 $0 \leq i < i' < 2^{k-1}$. 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. Since $\mathfrak{m}$ is $(i' - i)$-periodic, it follows that $\pi(\mathfrak{m}) \leq i' - i < 2^{k-1} < |\mathfrak{m}|/2$ and hence $w[i, i' + 2^k - 1]$ is $\pi(\mathfrak{m})$-periodic contradicting that the occurrence of $\mathfrak{m}$ at position i' 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 $i \in \{0, \ldots, 2^{k-1} - 1\}$ 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,

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

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

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

as required. □

A similar argument applies to $T_{\text{fin}}(\mathfrak{m}, n))$ 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 4.4. □

We recall the following result due to the first author from [4] (see also [6]):

Theorem 4.6. *Let C be a positive integer. Then for each aperiodic word $x \in \mathbb{A}^{\mathbb{N}}$ with $p_x(n) \leq Cn$ for $n \geq 1$, there exists a constant K (which is a polynomial function in C) such that $\text{Card}(\mathcal{R}_x(n)) \leq K$ for each $n \geq 0$.*

We next establish the following classification of words of linear complexity:

Theorem 4.7. *Let $x \in \mathbb{A}^{\mathbb{N}}$. Then $d_0(x) = 2$ if and only if $p_x(n) = \Theta(n)$. In particular, each $x \in \mathcal{W}(1)$ has cost equal to 0.*

Proof. One direction follows immediately from Proposition 3.8. In fact, if $d_0(x) = 2$, then applying Proposition 3.8 with $\alpha = 0$ and $k = 2$ we deduce that $x \in \mathcal{W}(1)$, i.e., $p_x(n) = O(n)$. On the other hand, by Lemma 3.3 we also have that x is aperiodic, and thus by Morse-Hedlund, $p_x(n) \geq n + 1$ for each n. Hence, $p_x(n) = \Theta(n)$ as required.

For the converse, suppose $x \in \mathbb{A}^{\mathbb{N}}$ and $p_x(n) = \Theta(n)$. Then x is aperiodic for otherwise $p_x(n) = O(1)$. Since there x is not assumed to be recurrent, to apply Theorem 4.4 we will need to replace x by 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, then so is y. We now apply Theorem 4.4 to show that there exist languages S and T 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 4.2, M is a D-marker set for y where $D = C + 1$. By Theorem 4.4 there exist languages S and T with p_S,p_T satisfying (3) where $M = \mathcal{R}_y$ and $D = C + 1$.

Since $\mathcal{R}_y(n) = \mathcal{R}_x(n) \cup \{a^n\}$ for each $n \geq 0$, by Theorem 4.6 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) 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. Since $\mathrm{Fac}(x) \subseteq \mathrm{Fac}(y) \subseteq ST$, it follows that $d_0(x) \leq 2$. But since x is aperiodic, Lemma 3.3 implies $d_0(x) \geq 2$. Hence $d_0(x) = 2$ as required. □

Remark 4.8. The general Theorem 4.4 can also be extended to non-recurrent one-sided infinite words by the same extension argument as Theorem 4.7.

Remark 4.9. Since the complexity of a Sturmian word is linear, Theorem 4.7 applies. However, the general result gives a poorer upper bound on the complexity of S than the one obtained in Example 3.4.

5. Cost and dimension of words of sub-quadratic complexity

We begin this section with another corollary of Theorem 4.4 which yields a non-trivial bound on the cost for words of complexity $o(n^2)$ (see Corollary 5.2).

Corollary 5.1. *Assume either $x \in \mathbb{A}^{\mathbb{Z}}$ and is aperiodic, or $x \in \mathbb{A}^{\mathbb{N}}$ and is both recurrent and aperiodic. Then there exist languages $S, T \subseteq \mathbb{A}^*$ with $\mathrm{Fac}(x) \subseteq ST$ and*

$$p_S(n), p_T(n) \leq \frac{12p_x(8n)}{n} + \frac{192p_x(8n)p_x(3n)}{n^2}$$

for each $n \geq 6$.

Proof. Fix $x \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^{\mathbb{Z}}$. Since x is aperiodic, by Proposition 4.3, there exists a 3-marker set M with $p_M(n) \leq \frac{p_x(4n)}{n}$. By Theorem 4.4 there exist languages $S, T \subseteq \mathbb{A}^*$ verifying (3) for $n \geq 6$ where $I_n = (\log_2\left(\frac{n}{6}\right), \log_2(2n)]$. Thus for each n, there are at most 4 possible values for k (say $k_0 < k_1 < k_2 < k_3$) and each verifies $2^{k_i} > 2^i \frac{n}{6}$ or equivalently $\frac{1}{2^{k_i}} < 2^{-i}\frac{6}{n}$. For each $0 \leq i \leq 3$ we bound the term $p_M(2^{k_i})$ by

$$p_M(2^{k_i}) \leq \frac{p_x(4 \cdot 2^{k_i})}{2^{k_i}} \leq \frac{p_x(8n)}{2^{k_i}}.$$

Thus from (3) we have

$$\begin{aligned}
p_S(n), p_T(n) &\leq \sum_{k \in I_n \cap \mathbb{N}} p_M(2^k)\left(1 + \frac{4p_x(3n)}{2^k}\right) \leq \sum_{i=0}^{3} p_M(2^{k_i})\left(1 + \frac{4p_x(3n)}{2^{k_i}}\right) \\
&\leq \sum_{i=0}^{3} \frac{p_x(8n)}{2^{k_i}} + \sum_{i=0}^{3} \frac{4p_x(8n)p_x(3n)}{2^{2k_i}} \leq p_x(8n)\sum_{i=0}^{3}\frac{1}{2^{k_i}} + 4p_x(8n)p_x(3n)\sum_{i=0}^{3}\frac{1}{2^{2k_i}} \\
&\leq \frac{6p_x(8n)}{n}\sum_{i=0}^{3}\frac{1}{2^i} + \frac{144p_x(8n)p_x(3n)}{n^2}\sum_{i=0}^{3}\frac{1}{2^{2i}} \\
&= \frac{15}{8} \cdot \frac{6p_x(8n)}{n} + \frac{85}{64} \cdot \frac{144p_x(8n)p_x(3n)}{n^2} \leq \frac{12p_x(8n)}{n} + \frac{192p_x(8n)p_x(3n)}{n^2}.
\end{aligned}$$

□

As an immediate consequence we have:

Corollary 5.2. *Let $\alpha \geq 1$. Then for each $x \in \mathcal{W}(\alpha)$ we have $c(x) \leq \min\{\alpha, 2\alpha - 2\}$.*

Proof. The result is clear in case x is ultimately periodic since $c(x) = 0$. Thus we may assume x is aperiodic. Clearly since $p_x(n) = O(n^\alpha)$, it follows that $c(x) \leq \alpha$. If x is recurrent, then by Corollary 5.1 taking $p_x(n) = O(n^\alpha)$, there exists languages S, T such that $\mathrm{Fac}(x) \subseteq ST$ and $p_S(n), p_T(n) = O(n^{2\alpha-2})$. Thus $c(x) \leq 2\alpha - 2$. If x is not recurrent, then as in the proof of

Theorem 4.7, we may replace x by an aperiodic bi-infinite word y with $p_y(n) = p_x(n) + n$. Since $\alpha \geq 1$, it follows that $p_y(n) = O(n^\alpha)$ and so we may apply Corollary 5.1 to y to deduce the existence of languages S, T with $\mathrm{Fac}(x) \subseteq \mathrm{Fac}(y) \subseteq ST$ and with $p_S(n), p_T(n) = O(n^{2\alpha-2})$. Whence again $c(x) \leq 2\alpha - 2$. □

As another consequence of Corollary 5.1 we have:

Corollary 5.3. *Let $x \in \mathbb{A}^\mathbb{N}$ be a pure morphic word (see [1]). Then except if the complexity of x is in $\Theta(n^2)$, we have $d_\alpha(x) \leq 2$ for each $\alpha > 0$ and hence $c(x) = 0$.*

Proof. By a celebrated result of Pansiot in [14], see also [6], if x is a pure morphic word, then $p_x(n) = \Theta(c_n)$ where $c_n \in \{1, n, n \log\log n, n \log n, n^2\}$. Applying Corollary 5.1 to each choice of c_n except $c_n = n^2$, gives $\mathrm{Fac}(x) \subseteq ST$ where $p_S(n), p_T(n) = O(n^\alpha)$ for each $\alpha > 0$. Whence $d_\alpha(x) \leq 2$ for each $\alpha > 0$ and hence $c(x) = 0$. □

We suspect that $c(x) = 0$ even for fixed points of complexity $O(n^2)$ although we are unable to prove it.

We saw that $d_0(x) = 1$ if and only if a word x is ultimately periodic, while $d_0(x) = 2$ if and only if $p_x(n) = \Theta(n)$. We now show that Theorem 4.7 does not extend to infinite words of quadratic complexity by exhibiting an infinite word u of complexity $p_u(n) = \Theta(n^2)$ for which $d_0(u) > 3$. But for this same word, we will show that $d_0(u) \leq 6$.

Theorem 5.4. *Let $u = \prod_{i=1}^\infty ab^i = ababbabbb\cdots$. Then $p_u(n) = \Theta(n^2)$ and $4 \leq d_0(u) \leq 6$.*

Proof. We begin by observing that the factor complexity of u is quadratic: u is the second shift of the fixed point beginning in c of the (non-primitive) morphism $a \mapsto ab, b \mapsto b, c \mapsto ca$, considered by Pansiot in [14] (see Theorem 4.1 and Example 1 therein). To show that $d_0(u) > 3$, we actually prove something stronger:

Lemma 5.5. $d_1^*(u) > 3$.

Proof. Suppose to the contrary that $d_1^*(u) \leq 3$. Then there exist languages $X, Y, Z \subseteq \{a, b\}^*$ with $p_X^*(n), p_Y^*(n), p_Z^*(n) = O(n)$ and such that $\mathrm{Fac}(u) \subseteq XYZ$. Thus each factor v of u admits a factorization $v = x(v)y(v)z(v)$ with $x(v) \in X$, $y(v) \in Y$ and $z(v) \in Z$.

For each $k, l \geq 1$ set $w_{k,l} = ab^l ab^{l+1} \cdots ab^{l+k-1}a$. Then each $w_{k,l}$ is a factor of u of length

$$|w_{k,l}| = k\left(l + \frac{k+1}{2}\right) + 1. \tag{5}$$

Claim 5.5.1. *Let*

$$E(n) = \{(k, l) \mid |w_{k,l}| \leq n,\ k \geq 3,\ l \geq \sqrt{n}\}.$$

Then $\mathrm{Card}(E(n)) = \Theta(n \log n)$.

Proof of Claim 5.5.1. Using (5), we see that the condition $|w_{k,l}| \leq n$ is equivalent to

$$l \leq \frac{n-1}{k} - \frac{k+1}{2}.$$

Thus,

$$\mathrm{Card}(E(n)) = \sum_{k=3}^{\infty} \mathrm{Card}\left(\left\{l \in \mathbb{N} : \sqrt{n} \leq l \leq \frac{n-1}{k} - \frac{k+1}{2}\right\}\right).$$

All but finite number of terms of this sum are null. In particular, they are null for $k \geq \sqrt{n}$: in that case,

$$\frac{n-1}{k} - \frac{k+1}{2} \leq \frac{n}{\sqrt{n}} - \frac{\sqrt{n}+1}{2} < \sqrt{n}.$$

A term number k of the sum is bounded from above by $\frac{n-1}{k}$ and from below by $\frac{n-1}{k} - \frac{k+1}{2} - \sqrt{n} - 1$ (this expression can be negative, so the kth term is not always equal to it). So,

$$\mathrm{Card}(E(n)) \leq \sum_{k=3}^{\lfloor\sqrt{n}\rfloor} \frac{n-1}{k} = \Theta\left(n \log n\right) \text{ and}$$

$$\mathrm{Card}(E(n)) \geq \sum_{k=3}^{\lfloor\sqrt{n}\rfloor} \left(\frac{n-1}{k} - \frac{k+1}{2} - \sqrt{n} - 1\right) = \Theta\left(n \log n\right).$$

□

We say that a factor v of u is *of type* (k,l) if $v = b^i w_{k,l} b^j$ for some $i, j \geq 0$. Clearly, each factor v of u is either of type (k,l) or contains at most one occurrence of the symbol a.

Claim 5.5.2. *Denote by $F(n)$ the subset of $E(n)$ of pairs (k,l) for which there exists a factor v of u of type (k,l) with $|v| \leq n$ whose decomposition $v = x(v)y(v)z(v)$ satisfies $|x(v)|_a \leq 1$ and $|z(v)|_a \leq 1$. Set $H(n) = E(n) \setminus F(n)$. Then* $\mathrm{Card}(H(n)) = \Theta(n \log n)$.

Proof of Claim 5.5.2. Consider the mapping $\varphi_n : F(n) \to Y$ defined as follows: For each $(k,l) \in F(n)$, there exists a factor v of u of type (k,l) with $|v| \leq n$, $|x(v)|_a \leq 1$ and $|z(v)|_a \leq 1$. Set $\varphi_n((k,l)) = y(v) \in Y$. Since $|v|_a = |w_{k,l}|_a = k+1 \geq 4$, we have that $|y(v)|_a \geq k-1 \geq 2$. It follows therefore that $y(v)$ is either of type (k,l), or of type $(k-1,l+1)$, or of type $(k-1,l)$, or of type $(k-2,l+1)$. This implies that for each $y \in Y$ in the image of φ_n, there are at most four pairs $(k,l) \in F(n)$ which map to y. But by assumption the total number of words in Y of length at most n is $p_Y^*(n) = O(n)$. Thus $\mathrm{Card}(F(n)) \leq 4p_Y^*(n) = O(n)$. On the other hand by Claim 5.5.1, we have $\mathrm{Card}(E(n)) = \Theta(n \log n)$. Thus $\mathrm{Card}(H(n)) = \Theta(n \log n)$. □

The next claim gives the asymptotic growth of the number of such factors v of u of type $(k,l) \in H(n)$.

Claim 5.5.3. *Let $s(n)$ denote the number of distinct factors v of u of length $|v| \leq n$ whose type belongs to $H(n)$. Then $s(n) = \Omega(n^2 \log n)$.*

Proof of Claim 5.5.3. In view of Claim 5.5.2, it suffices to show that for each type $(k,l) \in H(n)$ there are at least n factors v of u of length $|v| \leq n$ and of type (k,l). So fix a type $(k,l) \in H(n)$. Then v is of type (k,l) if and only if $v = b^i w_{k,l} b^j = b^i a b^l a b^{l+1} \cdots a b^{l+k-1} a b^j$ where $0 \leq i \leq l-1$ and $0 \leq j \leq l+k$. Thus there are at least l choices for each of i and j. But since $l \geq \sqrt{n}$, we have at least n choices for such v. □

Let v be a factor of u of length $|v| \leq n$ whose type belongs to $H(n)$. Then by definition of $H(n)$, writing $v = x(v)y(v)z(v)$ we have either $|x(v)|_a \geq 2$ or $|z(v)|_a \geq 2$. In case $|x(v)|_a \geq 2$, then v is uniquely determined by its length and $x(v)$. Thus the number of such words is bounded above by $np^*_X(n) = O(n^2)$. Similarly, if $|z(v)|_a \geq 2$, then v is uniquely determined by its length and $z(v)$, and hence the number of such words is also bounded above by $np^*_Z(n) = O(n^2)$. Thus $s(n) = O(n^2)$ in contradiction with Claim 5.5.3. This completes our proof of Lemma 5.5. □

Having established that $d^*_1(u) > 3$ it follows from Proposition 3.5 that $d_0(u) > 3$ as required. We next show that $d_0(u) \leq 6$.

Proposition 5.6. *Let $u = \prod_{i=1}^{\infty} ab^i$. Then there exist languages S_1, S_2, S_3 and S_4 with $S_1, S_4 \in \mathcal{L}(0)$ and $S_2, S_3 \in \mathcal{L}^*(1)$ such that* $\mathrm{Fac}(u) \subseteq S_1 S_2 S_3 S_4$.

Combined with Lemma 3.6 and Lemma 5.5, Proposition 5.6 yields:

Corollary 5.7. $d^*_1(u) = 4$ *and* $d_0(u) \leq 6$.

Proof of Proposition 5.6. Given a positive integer n, let $\nu_2(n)$ denote the 2-adic valuation of n defined as the largest exponent r such that 2^r divides n. Given positive integers $k \leq l$, there exists a unique $k \leq j \leq l$ such that $\nu_2(j) \geq \nu_2(i)$ for each $k \leq i \leq l$.

Every factor v of u containing at least two occurrences of the letter a is necessarily of the form $b^i a b^k a b^{k+1} a \cdots b^l a b^{i'} = b^i w_{l-k+1,k} b^{i'}$ for some $1 \leq k \leq l$, $0 \leq i \leq k-1$ and $0 \leq i' \leq l+1$. Given such a v we factor it as follows:

$$\underbrace{b^i}\underbrace{ab^k ab^{k+1} a \ldots ab^{j-1} a}\underbrace{b^j a \ldots ab^l a}\underbrace{b^{i'}} = \underbrace{b^i}\underbrace{w_{j-k,k}}\underbrace{b^j w_{l-j,j+1}}\underbrace{b^{i'}}$$

where j is the unique number between k and l of maximal 2-adic valuation. Here by convention $w_{0,k} = a$ for all k. Writing $j = 2^r(2m+1)$, where $r = \nu_2(j) \geq 0$ and $m \geq 0$, we have $k > j - 2^r = 2^{r+1}m$ and $l < j + 2^r = 2^{r+1}(m+1)$. Thus

$$\mathrm{Fac}(u) \subseteq S_1 S_2 S_3 S_4,$$

where $S_1 = S_4 = \{b^n \mid n \geq 0\}$, and

$$\begin{aligned} S_2 &= \{\varepsilon, a\} \cup \{ab^k a \cdots ab^{2^r(2m+1)-1} a \mid r \geq 0,\, m \geq 0,\, 2^r(2m+1) - 1 \geq k > 2^{r+1}m\}, \\ S_3 &= \{\varepsilon\} \cup \{b^{2^r(2m+1)} a \cdots ab^l a \mid r \geq 0,\, m \geq 0,\, 2^r(2m+1) \leq l < 2^{r+1}(m+1)\}. \end{aligned}$$

Note that by adding ε to both S_2 and S_3 allows us to also decompose factors of u containing fewer than two occurrences of the letter a. So for instance, $b^i a b^{i'}$ factors as $b^i a b^{i'} = b^i \cdot a \cdot \varepsilon \cdot b^{i'}$ and b^i as $b^i = b^i \cdot \varepsilon \cdot \varepsilon \cdot \varepsilon$. Also note that $ab^k a \cdots ab^{j-1} a \in S_2$ if and only if $\nu_2(j) = \max\{\nu_2(i) \,|\, k \leq i \leq j\}$, and similarly $b^j a \cdots ab^l a \in S_3$ if and only if $\nu_2(j) = \max\{\nu_2(i) \,|\, j \leq i \leq l\}$.

Clearly $p_{S_1}(n) = p_{S_4}(n) = 1$ for each $n \geq 0$, whence $S_1, S_4 \in \mathcal{L}(0)$. Thus it remains to show that S_2 and S_3 are each in $\mathcal{L}^*(1)$, i.e., each has linear accumulative complexity.

Claim 5.7.1. *Let s be a positive integer. Then for each fixed $r \geq 0$ and $m \geq 0$,*

$$\mathrm{Card}(\{v = ab^k a \cdots ab^{2^r(2m+1)-1} a \in S_2 \,|\, 2 \leq |v| \leq 2^s + 1\}) \leq \min\left\{2^r, \frac{2^s+1}{2^{r+1}m+1}\right\},$$

$$\mathrm{Card}(\{v = b^{2^r(2m+1)} a \cdots ab^l a \in S_3 \,|\, 2 \leq |v| \leq 2^s + 1\}) \leq \min\left\{2^r, \frac{2^s+1}{2^r(2m+1)+1}\right\}.$$

Proof of Claim 5.7.1. From the definition of S_2, if $v = ab^k a \cdots ab^{2^r(2m+1)-1} a \in S_2 \backslash \{\varepsilon, a\}$, then k ranges between $2^{r+1}m + 1$ and $2^r(2m+1) - 1$. Thus the number of such v is bounded above by $2^r(2m+1) - 1 - (2^{r+1}m+1) + 1 = 2^r - 1 < 2^r$. Similarly, if $v = b^{2^r(2m+1)} a \cdots ab^l a \in S_3 \backslash \{\varepsilon, a\}$, then l ranges between $2^r(2m+1)$ and $2^{r+1}(m+1) - 1$, thus the number of such v is bounded above by $2^{r+1}(m+1) - 1 - 2^r(2m+1) + 1 = 2^{r+1} - 2^r = 2^r$. The second estimate in each case takes into account the restriction on $|v|$ and is obtained by replacing the elements in each set by their lengths. In the case of S_2, we are estimating the cardinality of a set of natural numbers whose biggest element is at most $2^s + 1$, smallest element is $2^r(2m+1) + 1$, and the smallest difference between two elements is $2^{r+1}m + 2$ (corresponding to the smallest allowable value of k). Thus the cardinality of the set is bounded above by $\frac{(2^s+1)-(2^r(2m+1)+1)}{2^{r+1}m+2} + 1 < \frac{2^s+1}{2^{r+1}m+1}$. A similar argument yields the second estimate in the case of S_3. □

Claim 5.7.2. *Let s be a positive integer. Then $p^*_{S_2}(2^s+1) \leq 2 + 2^s(3+\sqrt{2})$.*

Proof of Claim 5.7.2. Let s be a positive integer. Let $v \in S_2$ with $|v| \leq 2^s + 1$. Then either $v = \varepsilon$ or $v = a$, or $v = ab^k a \cdots ab^{2^r(2m+1)-1} a$ in which case in particular $2^r(2m+1) + 1 \leq 2^s + 1$. This implies that $0 \leq r \leq s$ and $m < 2^{s-r-1}$. Thus either $0 \leq r < s$ and $m < 2^{s-r-1}$, or $s = r$ and $m = 0$. In the latter case, $v = ab^{2^s-1} a$ and hence this case contributes just one element to $p^*_{S_2}(2^s+1)$. Thus, adding $v = \varepsilon$ and $v = a$, we obtain the estimate

$$p^*_{S_2}(2^s+1) \leq 3 + \mathrm{Card}(\{v = ab^k a \cdots ab^{2^r(2m+1)-1} a \,|\, |v| \leq 2^s+1, 0 \leq r < s \text{ and } m < 2^{s-r-1}\}).$$

Applying Claim 5.7.1 for the number of words $v \in S_2$ of the form $v = ab^k a \cdots ab^{2^r(2m+1)-1} a$ for each parameter value (r, m) yields

$$p^*_{S_2}(2^s+1) \leq 3 + \sum_{r=0}^{s-1} \sum_{m=0}^{2^{s-r-1}} \min\left\{2^r, \frac{2^s+1}{2^{r+1}m+1}\right\}. \tag{6}$$

We extract for each value of r the term corresponding to $m = 0$. Since $\min\{2^r, 2^s + 1\} = 2^r$, the contribution to $p^*_{S_2}(2^s + 1)$ of all pairs $(r, 0)$ is bounded by $\sum_{r=0}^{s-1} 2^r = 2^s - 1$. Hence

$$p^*_{S_2}(2^s + 1) \leq 2 + 2^s + \sum_{r=0}^{s-1} \sum_{m=1}^{2^{s-r-1}} \min\left\{2^r, \frac{2^s + 1}{2^{r+1}m + 1}\right\}$$

Since $m < 2^{s-r-1}$, we have $2^{r+1}m < 2^s$ and hence $\frac{2^s+1}{2^{r+1}m+1} < \frac{2^s}{2^{r+1}m}$. Moreover since for all positive x, y we have $\min(x, y) \leq \sqrt{xy}$, we obtain

$$\begin{aligned} p^*_{S_2}(2^s + 1) &\leq 2 + 2^s + \sum_{r=0}^{s-1} \sum_{m=1}^{2^{s-r-1}} \min\left\{2^r, \frac{2^s}{2^{r+1}m}\right\} \\ &\leq 2 + 2^s + \sum_{r=0}^{s-1} \sum_{m=1}^{2^{s-r-1}} 2^{\frac{s-1}{2}} \frac{1}{\sqrt{m}} \\ &= 2 + 2^s + 2^{\frac{s-1}{2}} \sum_{r=0}^{s-1} \sum_{m=1}^{2^{s-r-1}} \frac{1}{\sqrt{m}}. \end{aligned}$$

Since

$$\sum_{m=1}^{2^{s-r-1}} \frac{1}{\sqrt{m}} \leq \int_0^{2^{s-r-1}} \frac{dx}{\sqrt{x}} = 2\sqrt{2^{s-r-1}} = 2^{\frac{s-r+1}{2}}$$

we obtain

$$p^*_{S_2}(2^s + 1) \leq 2 + 2^s \left(1 + \sum_{r=0}^{s-1} 2^{-r/2}\right) \leq 2 + 2^s \left(1 + \sum_{r=0}^{\infty} \left(\frac{1}{\sqrt{2}}\right)^r\right) = 2 + 2^s(3 + \sqrt{2})$$

as required. □

Claim 5.7.3. *For each positive integer n we have $p^*_{S_2}(n) \leq 2 + n(6 + 2\sqrt{2})$.*

Proof of Claim 5.7.3. For $n < 1$, the bound is obvious. Fix a positive integer $n \geq 2$ and pick $s \geq 1$ such that $2^{s-1} < n \leq 2^s$, so that $2^s < 2n$. Using Claim 5.7.2 together with the fact that $p^*_{S_2}$ is a non-decreasing function, we obtain

$$p^*_{S_2}(n) \leq p^*_{S_2}(2^s + 1) \leq 2 + 2^s(3 + \sqrt{2}) \leq 2 + 2n(3 + \sqrt{2}) \leq 2 + n(6 + 2\sqrt{2})$$

as required. □

It remains to find a linear bound for $p^*_{S_3}(n)$.

Claim 5.7.4. *Let s be a positive integer. Then $p^*_{S_3}(2^s + 1) \leq 2 + 2^s(3 + \sqrt{2})$. And hence as in Claim 5.7.3 we have $p^*_{S_3}(n) \leq 2 + n(6 + 2\sqrt{2})$.*

Proof of Claim 5.7.4. The proof for S_3 is analogous to that of S_2. Fix a positive integer s. Let $v \in S_3$ with $|v| \leq 2^s + 1$. Then either $v = \varepsilon$ or $v = b^{2^r(2m+1)}a \cdots ab^l a$ in which case $2^r(2m+1)+1 \leq 2^s+1$. As before this implies either $0 \leq r < s$ and $m < 2^{s-r-1}$, or $s = r$ and $m = 0$. In the latter case, $v = b^{2^s}a$ and hence this case contributes just one element to $p^*_{S_2}(2^s+1)$. Thus, combined with $v = \varepsilon$, we obtain the estimate

$$p^*_{S_3}(2^s+1) \leq 2 + \mathrm{Card}(\{v = b^{2^r(2m+1)}a \cdots ab^l a \,|\, |v| \leq 2^s+1, 0 \leq r < s \text{ and } m < 2^{s-r-1}\}).$$

Applying Claim 5.7.1 for the number of words corresponding to each parameter value (r, m) gives

$$p^*_{S_3}(2^s+1) \leq 2 + \sum_{r=0}^{s-1} \sum_{m=0}^{2^{s-r-1}} \min\left\{2^r, \frac{2^s+1}{2^r(2m+1)+1}\right\}. \tag{7}$$

The claim now follows by observing that the righthand side of (7) is less than the righthand side of (6). □

Claim 5.7.4 completes the proof of Proposition 5.6. □

This concludes our proof of Theorem 5.4. □

6. Positive cost for greater than quadratic complexity

At the moment, we do not know if the cost of a word of quadratic complexity can be greater than 0. However, the next theorem states that for any growth of complexity function which is faster than Cn^2, this is possible.

Theorem 6.1. *Let $f(n)$ be any non-decreasing integer function satisfying $f(1) = 1$, $f(n) \leq n$ and $\lim_{n\to\infty} f(n) = +\infty$. Then there exists an infinite word $x \in \{a,b\}^{\mathbb{N}}$ of complexity $O(n^2 f(n))$ such that if $\mathrm{Fac}(x) \subseteq S^k$ for some $S \subseteq \{a,b\}^*$ and $1 \leq k < +\infty$, then*

$$p^*_S(n) = \Omega\Big(\sum_{p=1}^{\frac{n-2}{2(2k-1)}} f(p)\Big).$$

Proof. Fix a function $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ satisfying $g(1,1) \geq 1$, $g(p,q) \leq g(p,q+1)$, $g(p, f(p)) \leq g(p+1, 1)$ for all $p, q \in \mathbb{N}$ and $\lim_{p\to\infty} g(p,1) = +\infty$. For instance, we can take $g(p,q) = p^{f(p)} + q$.

Define $x \in \{a,b\}^{\mathbb{N}}$ as follows:

$$x = \prod_{p=1}^{\infty} \prod_{q=1}^{f(p)} (a^p b^q)^{g(p,q)}.$$

Fix $k \geq 1$, and suppose $\mathrm{Fac}(x) \subseteq S^k$ for some language $S \subseteq \{a,b\}^*$.

Claim 6.1.1. *For every triple of positive integers n, p, q verifying $(p+q)(2k-1) \leq n-2$, $q \leq f(p)$ and $g(p,q) \geq 2k-1$, the set S contains a factor $s_{p,q}$ of $b(a^p b^q)^{2k-1}a$ of length $|s_{p,q}| \leq n$ containing $ba^p b^q a$ as a factor. Moreover, $s_{p,q} \neq s_{p',q'}$ whenever $(p,q) \neq (p',q')$.*

Proof of Claim 6.1.1. Since $g(p,q) \geq 2k-1$ and $q \leq f(p)$, the word $b(a^p b^q)^{2k-1}a$ is a factor of x. Moreover since $(p+q)(2k-1) \leq n-2$, we have that $|b(a^p b^q)^{2k-1}a| \leq n$. Given any factorization $b(a^p b^q)^{2k-1}a = u_1 u_2 \cdots u_k$ with $u_i \in \{a,b\}^*$, we see that of $2k$ occurrences of ba, at most $k-1$ lie accross boundaries of u_i. It remains $k+1$ occurrences of ba, and so two of them lie in the same u_j. This means that u_j contains $ba^p b^q a$ as a factor and we can take $s_{p,q} = u_j$. □

Let

$$P(n) = \{(p,q) \mid (p+q)(2k-1) \leq n-2,\ q \leq f(p),\ g(p,q) \geq 2k-1\}.$$

By Claim 6.1.1, there exists an injection

$$P(n) \hookrightarrow \{v \in S \mid |v| \leq n\}$$

given by $(p,q) \mapsto s_{p,q}$. We now estimate, for each n sufficiently large, the cardinality of the set $P(n)$. Since the function $g(p,q)$ is non-decreasing on p and q, and $g(p,1) \to +\infty$, there exists a positive integer p_0 such that $g(p,q) \geq 2k-1$ for all $p \geq p_0$ and all q. Since $f(p) \leq p$ for all p, for any $q \leq f(p)$ we have $p+q \leq p+f(p) \leq 2p$. In other words, any p between p_0 and $\frac{n-2}{2(2k-1)}$ satisfies the conditions $(p+q)(2k-1) \leq n-2$ and $g(p,q) \geq 2k-1$. Since for each such p there are $f(p)$ possible values for the second coordinate q, for all n sufficiently large we have

$$p_S^*(n) \geq \mathrm{Card}(P(n)) \geq \sum_{p=p_0}^{\frac{n-2}{2(2k-1)}} f(p).$$

Whence

$$p_S^*(n) = \Omega\Big(\sum_{p=0}^{\frac{n-2}{2(2k-1)}} f(p)\Big).$$

It remains to show that the factor complexity of x is $O(n^2 f(n))$. For this purpose we partition the factors of x into four groups and estimate the number of factors of length n in each group. Each factor v of x belongs to one or more of the following groups:

- group 1: factors of a block of the form $(a^p b^q)^j$ for some p, q and j.
- group 2: factors of a block of the form $(a^p b^q)^{k_1}(a^p b^{q+1})^{k_2}$.
- group 3: factors of a block of the form $(a^p b^{f(p)})^{k_1}(a^{p+1}b)^{k_2}$.
- group 4: factors containing some complete block $(a^p b^q)^{g(p,q)}$ as a factor.

We note that some of these groups overlap, which is not a problem since we seek only an upper bound on the factor complexity. We estimate the number of words of length n in each group.

In group 1, we have $O(n)$ words of the form $a^i b^{n-i}$ or $b^i a^{n-i}$, plus $O(n^2)$ words of the form $a^i b^q a^{n-q-i}$ (uniquely determined by $i \geq 1, q < n$) or $b^i a^p b^{n-p-i}$ (uniquely determined by $i \geq 1, p < n$), plus words containing factors of the form $ba^p b^q a$ or $ab^q a^p b$. These last set of words are uniquely determined by $p < n$, $q \leq f(p)$ and the position of the first occurrence of a^p, which takes values between 0 and $p + q - 1 < n$. Thus, the number of such words (and thus of all the words in group 1) is $O(n^2 f(n))$.

Words in group 2 which do not belong to group 1 contain factors of the form $ab^q a^p b^{q+1}$. Such a word is uniquely determined by $p < n$, $q \leq f(p) - 1$ and the position of the first occurrence of b^{q+1}, which takes values between 0 and $n - q - 1 < n$. Hence the number of such words is also $O(n^2 f(n))$.

An analogous counting argument applies to group 3. Words in group 3 which have not yet been accounted for are uniquely determined by $p < n$ and the first position of a^{p+1}, whence their number is $O(n^2)$.

Finally, for each word v in group 4, we consider the first complete block $u = (a^p b^q)^{g(p,q)}$ contained in v. Then v is uniquely determined by p, q and the position of u in v, hence the number of such words is again $O(n^2 f(n))$.

Thus, the complexity $p_x(n) = O(n^2 f(n))$ as required. This completes the proof of Theorem 6.1. □

Corollary 6.2. *For each non-decreasing integer function $f(n)$ verifying $f(1) = 1$, $f(n) \leq n$ and $\lim_{n\to\infty} f(n) = +\infty$, there exists an infinite word $x \in \{a, b\}^{\mathbb{N}}$ of complexity $O(n^2 f(n))$ with $d_0(x) = d_1^*(x) = +\infty$.*

Proof. Let x be as in Theorem 6.1. Due to result of the theorem, if $\mathrm{Fac}(x) \subseteq S^k$ for some language S, then $p_S^*(n) = \Omega(\sum_{p=1}^{\frac{n-2}{2(2k-1)}} f(p))$. Given any positive M, we can find p_0 such that $f(p_0) > M$; then, since $f(n)$ is non-decreasing,

$$\sum_{p=1}^{\frac{n-2}{2(2k-1)}} f(p) > \sum_{p=p_0}^{\frac{n-2}{2(2k-1)}} f(p) \geq M\left(\frac{n-2}{2(2k-1)} - p_0\right) > \frac{M}{4k}n + d$$

for an appropriate constant d not depending on n. So, $p_S^*(n)$ grows faster than linearly. This means exactly that $d_1^*(x) = +\infty$; and $d_0(x) = +\infty$ due to Proposition 3.5. □

Corollary 6.3. *For each $0 < \alpha < 1$, there exists an infinite word $x \in \{a, b\}^{\mathbb{N}}$ of complexity $O(n^{2+\alpha})$ such that $c(x) \geq \alpha$.*

Proof. Fix $0 < \alpha < 1$. Then applying Theorem 6.1 to $f(n) = \lfloor n^\alpha \rfloor$, we have that there exists a word $x \in \{a,b\}^{\mathbb{N}}$ of complexity $O(n^{2+\alpha})$ such that if $\mathrm{Fac}(x) \subseteq S^k$ for some $S \subseteq \{a,b\}^*$ and $1 \le k < +\infty$, then

$$p_S^*(n) = \Omega(n^{\alpha+1}).$$

Thus $c^*(x) \ge \alpha + 1$, and hence $c(x) \ge \alpha$. □

7. Non-factorial languages

Positive results of previous sections concern mostly languages of factors of infinite words. In this section, we show that for a general non-factorial language, low complexity does not imply $d_0(L) < +\infty$.

Theorem 7.1. *There exists a non-factorial language L of complexity $p_L(n) = O(\log n)$ (and hence of cost zero) such that $d_0(L) = +\infty$.*

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 $|y_n| \log_2 |y_n| \le n$. Thus for example $y_2 = 10$ since $2\log_2 2 \le 2 < 3\log_2 3$ and $y_{65} = 1000001210000012$ since $16\log_2 16 \le 65 < 17\log_2 17$. Finally, define $L = \{y_n | n \ge 1\}$.

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

$$|y_n| \le \frac{n}{\log_2 |y_n|} \le n. \tag{8}$$

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

$$|y_n| + 1 > \frac{n}{\log_2(|y_n|+1)} \ge \frac{n}{\log_2(n+1)}, \tag{9}$$

so $|y_n| = \Omega(n/\log n)$. Combining the (8) and (9) yields

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

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 show that $d_0(L) = +\infty$. Indeed, suppose by contrary that $L \subseteq S^k$ for some $k \in \mathbb{Z}$ 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 $|y_n| \ge (k+1)|x_n|$. This means that for all $n > n_0$, the word y_n contains at least $k+1$ occurrences of 2, and 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 $n \geq n_0$ are pairwise distinct.

Now for each $n \geq n_0$ 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 accumulative complexity of S grows faster than linearly, which is impossible if its usual complexity is bounded.

It remains to prove that $p_L(n) = \Theta(\log n)$. Indeed,

$$p_L(n) = \#\{m : |y_m| = n\}.$$

In other words,

$$p_L(n) = \#\{m : n \log_2 n \leq m < (n+1)\log_2(n+1)\}.$$

Whence,

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

This completes the proof of Theorem 7.1. □

The language L in Theorem 7.1 provides an example of a language of cost equal to 0 and having infinite cost dimension. We do not know whether there exists an infinite word x with $c(x) = 0$ and $d_0(x) = +\infty$.

References

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

[2] 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.

[3] S. Brlek, Enumeration of factors in the Thue-Morse word, *Disc. Appl. Math.* **24** (1989), pp. 83–96.

[4] J. Cassaigne, Special factors of sequences with linear subword complexity, DLT **1995**, pp. 25–34, World Sci. Publishing, Singapore, 1996.

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

[6] J. Cassaigne, F. Nicolas, Factor complexity, Combinatorics, automata and number theory, Encyclopedia Math. Appl., **135**, Cambridge Univ. Press, 2010, pp. 163–247.

[7] 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.

[8] A. Ehrenfeucht, K.P. Lee, G. Rozenberg, Subword complexities of various deterministic developmental languages without interactions. *Theoret. Comput. Sci.* **1** (1975) pp. 59–76.

[9] J. Leroy, Some improvements of the S-adic conjecture, *Adv. in Appl. Math.* **48** (2012), no. 1, pp. 79–98.

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

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

[12] M. Lothaire, Applied combinatorics on words, Cambridge University Press, 2005.

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

[14] J.-J. Pansiot. Complexité des facteurs des mots infinis engendrés par morphismes itérés, in Paredaens, J. (ed.) ICALP **1984**, LNCS, vol. 172, Springer, Heidelberg, 1984, pp. 380–389.

[15] A. Thue, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat-Nat. Kl.* **7** (1906), pp. 1–22.

[16] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat-Nat. Kl.* **1** (1912), pp. 1–67.

[17] L.Q. Zamboni, On the character of words of sub-linear complexity, in preparation 2016.