Bull. Soc. Math. France
147 (4), 2019, p. 639-660

COST AND DIMENSION OF WORDS OF ZERO TOPOLOGICAL ENTROPY

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

Abstract. — The (factor) complexity of a language L is defined as a function $p_L(n)$ which counts for each n the number of words in L of length n. We are interested in whether L is contained in a finite product of the form S^k, where S is a language of strictly lower complexity. In this paper, we focus on languages of zero topological entropy, meaning $\limsup_{n\to\infty} \log p_L(n)/n = 0$. We define the α-dimension of a language L as the infimum of integer numbers k such that there exists a language S of complexity $O(n^\alpha)$ such that $L \subseteq S^k$. We then define the cost $c(L)$ as the infimum of all real numbers α for which the α-dimension of L is finite. In particular, the above definitions apply to the language of factors of an infinite word. In the paper, we search for connections between the complexity of a language (or an infinite word) and its dimension and cost, and show that they can be rather complicated.

Texte reçu le 13 juillet 2018, modifié le 14 janvier 2019, accepté le 1er février 2019.

Julien Cassaigne, Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France • *E-mail :* julien.cassaigne@math.cnrs.fr • *Url :* http://iml.univ-mrs.fr/~cassaign/

Anna E. Frid, Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France • *E-mail :* anna.e.frid@gmail.com • *Url :* http://iml.univ-mrs.fr/~frid/

Svetlana Puzynina, Saint Petersburg State University, 7–9 Universitetskaya emb., 199034 Saint Petersburg, Russia and also Sobolev Institute of Mathematics, 4 Acad. Koptyug avenue, 630090 Novosibirsk, Russia • *E-mail :* s.puzynina@gmail.com • *Url :* http://math.nsc.ru/~puzynina/

Luca Q. Zamboni, Institut Camille Jordan, Université Claude Bernard Lyon 1, 43 boulevard du 11 novembre 1918 F-69622 Villeurbanne Cedex • *E-mail :* zamboni@math.univ-lyon1.fr

Mathematical subject classification (2010). — 68R15, 37B10.

Key words and phrases. — Symbolic dynamics, Factor complexity, Topological entropy.

The work of S. Puzynina has been partially supported by Russian Foundation of Basic Research (grant 18-31-00118).

BULLETIN DE LA SOCIÉTÉ MATHÉMATIQUE DE FRANCE 0037-9484/2019/639/$ 5.00

doi:10.24033/bsmf.2794

Résumé (*Coût et dimension des mots d'entropie topologique nulle*). — La complexité d'un langage L est définie comme la fonction $p_L(n)$ qui compte le nombre de mots de longueur n dans L. Nous nous intéressons à savoir si L est contenu dans un produit fini de la forme S^k, où S est un langage de complexité strictement inférieure. Dans cet article, nous considérons des langages d'entropie topologique nulle, c'est-à-dire $\limsup_{n\to\infty} \log p_L(n)/n = 0$. Nous définissons l'$\alpha$-dimension d'un langage L comme la borne inférieure des nombres entiers k tels qu'il existe un langage S de complexité $O(n^\alpha)$ avec $L \subseteq S^k$. Nous définissons ensuite le coût $c(L)$ comme la borne inférieure de tous les nombres réels α pour lesquels l'α-dimension de L est finie. En particulier, les définitions ci-dessus s'appliquent au langage des facteurs d'un mot infini. Dans l'article, nous cherchons les liens entre la complexité d'un langage (ou d'un mot infini) et sa dimension et son coût, et montrons qu'ils peuvent être assez compliqués.

1. Introduction

Consider a finite non-empty set $\mathbb{A}$ called an alphabet. The complexity or *factor complexity* $p_x(n)$ of an infinite word $x = x_0x_1x_2\cdots \in \mathbb{A}^{\mathbb{N}}$ is 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 1938 [7] under the name *block growth*, the factor complexity provides a useful measure of randomness of x or of the subshift it generates. In particular, periodic words have bounded factor complexity while digit expansions of normal numbers, by the definition of normality, have maximal complexity.

The set $\mathbb{A}^*$ 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 language $L \subseteq \mathbb{A}^*$ (for instance the language $\mathrm{Fac}(x)$ 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, that is, whether each word from L can be represented as a concatenation of at most k words from S. The starting point of this paper is the following characterisation of infinite words $x \in \mathbb{A}^{\mathbb{N}}$ of sub-linear complexity obtained by the authors in [3]:

Theorem 1.1. — *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.,* $\limsup p_S(n) < +\infty$*).*

Our aim here is to express and study these ideas in greater generality, for words and languages of higher but still low complexity. Given a language $L \subseteq \mathbb{A}^*$, we define the *cost* of L, denoted $c(L)$, as the infimum of all real numbers α for which there exists 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 $\alpha \in [0, +\infty)$, denote by $\mathcal{L}(\alpha)$ the collection of all languages $L \subseteq \mathbb{A}^*$ whose complexity satisfies $p_L(n) = O(n^\alpha)$. For example, if x is an infinite word and $L = \mathrm{Fac}(x)$, then, by the Morse-Hedlund theorem [7], L belongs to $\mathcal{L}(0)$ if and only if x is ultimately periodic. If the complexity of x is linear, then L belongs to $\mathcal{L}(1)$, if the complexity is bounded by cn^2, L belongs to $\mathcal{L}(2)$, and so on.

Now 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 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 the case of $L = \mathrm{Fac}(x)$ 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 1.1 together with the Morse-Hedlund theorem 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 $\limsup_{n\to+\infty} \frac{\log p_L(n)}{n}$ have cost equal to $+\infty$. For this reason we restrict our attention to languages and words of zero topological entropy $\limsup_{n\to+\infty} \frac{\log p_L(n)}{n} = 0$. Moreover, as we show by a straightforward argument in Proposition 3.9, $c(L)$ is finite if and only if the complexity of L is bounded above by a polynomial.

Then, in Proposition 3.12, for each positive integer $k \geq 1$ we construct an infinite word x of complexity $p_x(n) = \Omega(n^{k-1})$ 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? Despite the Morse-Hedlund theorem and Theorem 1.1, in general, the cost and cost dimension of a given language depend only in part on its complexity. For instance, languages not closed under taking factors are in general very far from satisfying any result along the lines of Theorem 1.1 (see the last proposition in [3]), but even in the case of languages defined by infinite words, the characterisation of Theorem 1.1 does not seem to extend in an obvious way to higher complexities. For instance, we prove in Theorem 5.1 that the word $x = \prod_{i=1}^{+\infty} ab^i = ababbabbb\cdots$, of complexity $p_x(n) = \Theta(n^2)$, verifies

$d_0(u) > 3$. On the other hand, in the same theorem we show that $d_0(x) \leq 6$, which in particular implies it is of cost zero. We do not know whether there exist words of quadratic complexity and positive cost. However, we prove in Theorem 6.1 that for every real number $\alpha \in (0,1)$ there exists an infinite word x with complexity $p_x(n) = O(n^{2+\alpha})$ and cost $c(x) \geq \alpha$. In other words, there exist words of positive cost and of complexity growing just a bit faster than quadratically. 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.

Some of the results of the paper have been presented at the 2014 MFCS conference [4].

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 [6].

Let $\mathbb{A}$ be a finite non-empty set (the *alphabet*). Let $\mathbb{A}^*$ ($\mathbb{A}^\mathbb{N}$) denote the set of all finite (right infinite) words $u = u_0u_1 \cdots u_{n-1}(\cdots)$ with $u_i \in \mathbb{A}$. The length n of a finite word u is denoted by $|u|$. 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. The set of *factors* of a finite or infinite word u is defined by

$$\mathrm{Fac}(u) = \{u_i \cdots u_j \,|\, 0 \leq i \leq j\} \cup \{\varepsilon\},$$

where $j < |u|$ if u is finite. The factor $u_i \ldots u_j$ can be also denoted by $u[i..j]$.

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 $u \in L$. The *complexity* p_L of a language $L \subseteq \mathbb{A}^*$ is defined by $p_L(n) = \mathrm{Card}(L \cap \mathbb{A}^n)$; its *accumulative complexity* p^*_L is defined by $p^*_L(n) = \sum_{i=0}^n p_L(i)$. For a finite of infinite word x, the complexity (accumulative complexity) of $\mathrm{Fac}(x)$ is denoted simply by $p_x(n)$ (respectively, $p^*_x(n)$).

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

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

For each real number $\alpha \in [0, +\infty)$, we let $\mathcal{L}(\alpha)$ (resp., $\mathcal{L}^*(\alpha)$) denote 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 let $\mathcal{W}(\alpha)$ (resp., $\mathcal{W}^*(\alpha)$) denote 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., $\mathrm{Fac}(x) \in \mathcal{L}^*(\alpha)$). For each $S \subseteq \mathbb{A}^*$, the set S^k denotes 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

$$d_\alpha(L) = \inf\{k \geq 1 \mid \exists S \in \mathcal{L}(\alpha) : L \subseteq S^k\},$$

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. □

Theorem 1.1 from [3] can thus be reformulated as follows: for an infinite word x, we have $d_0(x) = 2$ if and only if x is of linear complexity. Our goal in this paper is to establish further connections between complexity and cost of words and languages. We start from some basic properties.

First of all, the language S in the definition of α-dimension is usually not factorial, as we show in the next proposition.

Proposition 3.4. — *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 real $\alpha \geq 0$, 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 the case of $k = 1$, 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 that $\mathrm{Fac}(y) \subseteq S$. 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. □

Corollary 3.5. — *If for a real $\alpha \geq 0$ and $x \in \mathbb{A}^{\mathbb{N}}$, we set*

$$d_\alpha^{\mathrm{fac}}(x) = \inf\{k \geq 1 \,|\, \mathrm{Fac}(x) \subseteq S^k \textit{ for some factorial language } S \in \mathcal{L}(\alpha)\}$$

and

$$c^{\mathrm{fac}}(x) = \inf\{\alpha \in [0, +\infty) \,|\, d_\alpha^{\mathrm{fac}}(x) < +\infty\},$$

then these functions are degenerate: $d_\alpha^{\mathrm{fac}}(x) < +\infty$ if and only if $x \in \mathcal{W}(\alpha)$ and hence $c^{\mathrm{fac}}(x) = \inf\{\alpha \in [0, +\infty) \,|\, x \in \mathcal{W}(\alpha)\}$.

Proof. — Fix $\alpha \geq 0$. If $x \in \mathcal{W}(\alpha)$, then taking $S = \mathrm{Fac}(x)$ we have $d_\alpha^{\mathrm{fac}}(x) = 1$. Conversely if $d_\alpha^{\mathrm{fac}}(x) < +\infty$, then by Proposition 3.4 we have $x \in \mathcal{W}(\alpha)$. □

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

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

To prove the proposition, we first need the following technical lemma.

Lemma 3.7. — *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 $i \in \{0, \ldots, |v_{m+1}|\}$ we can factor v_{m+1} as $v_{m+1} = x_i y_i$, with $|x_i| = i$. We claim that there exists $j \in \left\{0, \ldots, \left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1\right\}$ 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 $i \in \left\{0, \ldots, \left\lceil \frac{|v_{m+1}|}{2} \right\rceil - 1\right\}$ 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:

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

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 $T \subseteq S^2$. □

Proof of Proposition 3.6. 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 $d_\alpha(L) \leq 2d^*_{\alpha+1}(L)$, 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 3.7 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 3.6.

Corollary 3.8. — *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 3.9. — *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 a language L. The result is clear in the case of $k = 1$. 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. There are $\binom{n+k-1}{k-1}$ ways of factoring $u = v_1v_2\cdots v_k$ with $|v_i| \geq 0$. Here

$\binom{n+k-1}{k-1} = O(n^{k-1})$, and each $v_i \in S$, so that there are at most $C|v_i|^\alpha$ choices for each v_i. Thus $p_L(n) = O(n^{k(\alpha+1)-1})$ as required. □

As an immediate consequence we get:

Corollary 3.10. — *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 let $k = d_\alpha(L)$. Then by Proposition 3.9 $L \in \mathcal{L}(k(\alpha+1)-1)$. 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 3.11. — *Languages of positive entropy have cost equal to $+\infty$.*

Proposition 3.9 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 3.12. — *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 $k = 2$, by Theorem 1.1, it suffices to take x to be any aperiodic word of linear complexity. 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

$$\{1^{j_1} 2^{j_2} \cdots (k-2)^{j_{k-2}} \,|\, j_1 + j_2 + \cdots + j_{k-2} \geq 1\} = \{t_1, t_2, t_3, \ldots\}$$

where the t_i are listed in the increasing radix order on $\{1, \ldots, k-2\}^+$, that is, $u < v$ if and only if either $|u| < |v|$, or $|u| = |v|$ and u is less than v lexicographically. So the sequence $t_1, t_2, \ldots$ looks like $1, 2, \ldots, k-2, 11, 12, \ldots, 1(k-2), 22, \ldots, (k-3)(k-2), (k-2)(k-2), 111, 112, \ldots$ Then $x \in \{0, 1, \ldots, k-2\}^{\mathbb{N}}$ is defined by

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

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 exists $0t_q0$, where the length of t_q is at least $n/2$. Such a factor of x exists for each $t_q = 1^{j_1} 2^{j_2} \cdots (k-2)^{j_{k-2}}$ (that is, for each $j_1, \ldots, j_{k-2}$ under the condition $n/2 \leq j_1 + \cdots + j_{k-2} = |t_q| \leq n-2$), and for each starting point of that occurrence of t_q, which is any number between 1 and $n - |t_q| - 1$. The condition $|t_q| \geq n/2$ ensures that they are all distinct. So, we have $k-1$ degrees of freedom, and thus the complexity of x is $\Omega(n^{k-1})$. On the other hand, take a factor w of x and find in it a word t_q, where q is maximal. Here incomplete intersections count: we just fix an occurrence of w to x, see what words t_q it intersects and choose the greatest q. If t_q is completely in w, it is followed in it by a prefix of x. The set of prefixes of x will be denoted by S_{k-1}. Symmetrically, just before t_q in w, there is a suffix of some word w_m (and w_m are suffixes one of another).The set of these suffixes is denoted by S_0. As for t_q itself, it belongs to the concatenation of $1^* = S_1$, $2^* = S_2$, etc.; so,

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

where the complexity of each S_i is 1.

If t_q is not completely contained in w, three situations are possible. Either $w = t's$, where t' is a suffix of t_q; then $t' \in i^*(i+1)^* \cdots (k-2)^*$ for some $i \in \{1, \ldots, k-2\}$, s is a prefix of x, and thus $w \in S_i \cdots S_{k-2} S_{k-1} \subset S_0 S_1 \cdots S_{k-2} S_{k-1}$. Or, symmetrically, $w = pt''$, where t'' is a prefix of t_q; then $t'' \in 1^* 2^* \cdots i^*$ for some $i \in \{1, \ldots, k-2\}$, p is a suffix of some w_m, and thus $w \in S_0 S_1 \cdots S_i \subset S_0 S_1 \cdots S_{k-2} S_{k-1}$. Or, at last, w is a factor of t_q, and then $w \in i^*(i+1)^* \cdots j^*$ for some $i, j \in \{1, \ldots, k-2\}$, $i \leq j$, and thus $w \in S_i S_{i+1} \cdots S_j \subset S_0 S_1 \cdots S_{k-2} S_{k-1}$. In all the cases, (2) holds. It follows that $d_0(x) \leq k$; by Proposition 3.9, $d_0(x) < k$ would imply complexity $O(n^{k-2})$, so $d_0(x) = k$. □

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

4. Cost and dimension of words of less than quadratic complexity

We begin by recalling a few definitions and results obtained by the authors in [3]. Let $x \in \mathbb{A}^{\mathbb{N}} \cup \mathbb{A}^{\mathbb{Z}}$.

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.

For example, it is shown in [3] that if x is aperiodic and C is a positive integer such that $p_x(n) \leq Cn$ for all $n \geq 1$, then the set $\mathcal{R}_x$ of right special factors of x is a $(C+1)$-marker set for x. In what follows we will need the following proposition also proved in [3]:

Proposition 4.2. — *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 $n \geq 1$.

The following result was proved in [3]. Note that here it is more convenient to consider two different languages S and T instead of their union.

Theorem 4.3. — *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 $I_n = (\log_2(\frac{n}{2D}), \log_2(2n)]$.

As a first consequence of Theorem 4.3, combined with an earlier result of the first author in [2] which gives a uniform bound on the number of right special factors of each length n of an infinite word of linear complexity, we obtain the following reformulation of Theorem 1.1 characterising words of linear complexity:

Theorem 4.4. — *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. — First if $d_0(x) = 2$ then $\mathrm{Fac}(x) \subseteq S^2$ for some $S \subseteq \mathbb{A}^*$ of bounded complexity. Applying Theorem 1.1 together with the fact that x is aperiodic (a consequence of the Morse and Hedlund theorem) we obtain that $p_x(n) = \Theta(n)$ as required. Conversely suppose $x \in \mathbb{A}^{\mathbb{N}}$ and $p_x(n) = \Theta(n)$. Then by Theorem 1.1 in [3] we have that $\mathrm{Fac}(x) \subseteq S^2$ for some $S \subseteq \mathbb{A}^*$ of bounded complexity. Thus $d_0(x) \leq 2$. But since x is aperiodic, Lemma 3.3 implies $d_0(x) \geq 2$. Hence $d_0(x) = 2$ as required. □

We now derive yet another corollary of Theorem 4.3 which yields a nontrivial bound on the cost for words of complexity $o(n^2)$:

Corollary 4.5. — *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 4.2, there exists a 3-marker set M with $p_M(n) \leq \frac{p_x(4n)}{n}$. By Theorem 4.3 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 four 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 $i = 0, \ldots, 3$ 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 4.6. — *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 the case where 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 4.5 with $p_x(n) = O(n^\alpha)$, there exist 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 we may replace x with an aperiodic bi-infinite word $y = \cdots 000x$, where 0 is a new letter, so that $p_y(n) = p_x(n) + n$. Since $\alpha \geq 1$, it follows that $p_y(n) = O(n^\alpha)$ and so we may apply Corollary 4.5 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 4.5, we can calculate cost and complexity of most pure morphic words (see [1]).

Corollary 4.7. — *Let $x \in \mathbb{A}^{\mathbb{N}}$ be a pure morphic word. Then if $p_x(n)$ is not 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 [8], see also [5], 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 4.5 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 conjecture that $c(x) = 0$ even for fixed points of complexity $\Theta(n^2)$. In the next section we give an example of such a word of cost dimension d_0 between 4 and 6. In particular, this example means that Theorem 4.4 does not directly extend to infinite words of quadratic complexity and $d_0(x) = 3$.

5. An example of quadratic complexity

Consider the word $u = \prod_{i=1}^{+\infty} ab^i = ababbabbb\cdots$. Its factor complexity is quadratic, which can be easily proved directly and which was also shown by Pansiot in [8] (see Theorem 4.1 and Example 1 therein) using the fact that u is obtained by erasing the prefix ca from the fixed point beginning with c of the (non-primitive) morphism $a \mapsto ab, b \mapsto b, c \mapsto ca$. This section is devoted to a study of 0-dimension of u and to the proof of the following

Theorem 5.1. — *Let $u = \prod_{i=1}^{+\infty} ab^i = ababbabbb\cdots$. Then $4 \leq d_0(u) \leq 6$.*

Proof. — To show that $d_0(u) > 3$, we actually prove something stronger:

Lemma 5.2. — $d_1^*(u) > 3$.

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

$$|w_{k,l}| = k\left(l + \frac{k+1}{2}\right) + 1. \tag{4}$$

Claim 5.2.1. — *Let*

$$E(n) = \{(k,l) \mid |w_{k,l}| + 2l + k \leq n,\ k \geq 3,\ l \geq \sqrt{n}\}.$$

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

Proof of Claim 5.2.1. — Using (4), we see that the condition $|w_{k,l}|+2l+k \leq n$ is equivalent to

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

Thus,

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

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

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

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

$$\operatorname{Card}(E(n)) \leq \sum_{k=3}^{\lfloor\sqrt{n}\rfloor} \frac{n}{k+2} = \Theta\left(n \log n\right) \text{ and}$$

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

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 5.2.2. — *Let $F(n)$ denote 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 $H(n) = E(n) \backslash F(n)$. Then* $\operatorname{Card}(H(n)) = \Theta(n \log n)$.

Proof of Claim 5.2.2. — 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 $|z(v)|_a \leq 1$. Choose one such factor. 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 $\operatorname{Card}(F(n)) \leq 4p^*_Y(n) = O(n)$. On the other hand, by Claim 5.2.1, we have $\operatorname{Card}(E(n)) = \Theta(n \log n)$. Thus $\operatorname{Card}(H(n)) = \Theta(n \log n)$. $\square$

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

Claim 5.2.3. — *Let $s(n)$ denote the number of distinct factors v of u whose type belongs to $H(n)$. Then $s(n)=\Omega(n^2\log n)$.*

Proof of Claim 5.2.3. — In view of Claim 5.2.2, it suffices to show that for each type $(k,l)\in H(n)$ there are at least n factors v of u of length $|v|\le n$ 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 $v=b^iw_{k,l}b^j=b^iab^lab^{l+1}\cdots ab^{l+k-1}ab^j$ where $0\le i\le l-1$ and $0\le j\le l+k$. Thus there are at least l choices for each of i and j. But since $l\ge\sqrt{n}$, we have at least n choices for such v. □

Let v be a factor of u whose type belongs to $H(n)$. Note that $|v|\le|w_{k,l}|+2l+k-1\le n$. Then by definition of $H(n)$, writing $v=x(v)y(v)z(v)$ we have either $|x(v)|_a\ge 2$ or $|z(v)|_a\ge 2$. In the case of $|x(v)|_a\ge 2$, 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 $|z(v)|_a\ge 2$, 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 $s(n)=O(n^2)$ is in contradiction with Claim 5.2.3. This completes our proof of Lemma 5.2. □

Having established that $d_1^*(u)>3$ it follows from Proposition 3.6 that $d_0(u)>3$ as required.

We next show that $d_0(u)\le 6$.

Proposition 5.3. — *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* $\mathrm{Fac}(u)\subseteq S_1S_2S_3S_4$.

Combined with Lemma 3.7 and Lemma 5.2, Proposition 5.3 yields:

Corollary 5.4. — $d_1^*(u)=4$ *and* $d_0(u)\le 6$.

Proof of Proposition 5.3. — 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 $k\le l$, there exists a unique j such that $k\le j\le l$ and $\nu_2(j)\ge\nu_2(i)$ for each $i=k,\dots,l$.

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

$$\underbrace{b^i}\ \underbrace{ab^kab^{k+1}a\cdots ab^{j-1}a}\underbrace{b^ja\cdots ab^la}\ \underbrace{b^{i'}}\ =\ \underbrace{b^i}\ \underbrace{w_{j-k,k}}\,\underbrace{b^jw_{l-j,j+1}}\,\underbrace{b^{i'}}$$

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 $r=\nu_2(j)\ge 0$ and $m\ge 0$, we have $k>j-2^r=2^{r+1}m$ and $l<j+2^r=2^{r+1}(m+1)$. Thus

$$\mathrm{Fac}(u)\subseteq S_1S_2S_3S_4,$$

where $S_1 = S_4 = \{b^n \,|\, n \geq 0\}$, $S_2 = \{\varepsilon, a\} \cup \bigcup_{r\geq 0} \bigcup_{m \geq 0} S_{2,r,m}$, and $S_3 = \{\varepsilon\} \cup \bigcup_{r\geq 0} \bigcup_{m\geq 0} S_{3,r,m}$, with

$$S_{2,r,m} = \{ab^k a \cdots ab^{2^r(2m+1)-1}a \,|\, 2^{r+1}m < k \leq 2^r(2m+1) - 1\},$$
$$S_{3,r,m} = \{b^{2^r(2m+1)}a \cdots ab^l a \,|\, 2^r(2m+1) \leq l < 2^{r+1}(m+1)\}.$$

Note that 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, $b^i ab^{i'}$ factors as $b^i ab^{i'} = b^i \cdot a \cdot \varepsilon \cdot b^{i'}$ 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 5.4.1. — *Let s be a positive integer. Then for each fixed $r \geq 0$ and $m \geq 0$,*

$$\mathrm{Card}(\{v \in S_{2,r,m} \,|\, |v| \leq 2^s + 1\}) \leq \min\left\{2^r, \frac{2^s+1}{2^{r+1}m+1}\right\},$$
$$\mathrm{Card}(\{v \in S_{3,r,m} \,|\, |v| \leq 2^s + 1\}) \leq \min\left\{2^r, \frac{2^s+1}{2^r(2m+1)+1}\right\}.$$

Proof of Claim 5.4.1. — From the definition of $S_{2,r,m}$ we see that if $v = ab^k a \cdots ab^{2^r(2m+1)-1}a \in S_{2,r,m}$, 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 $v = b^{2^r(2m+1)}a \cdots ab^l a \in S_{3,r,m}$, 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 $2^{r+1}(m+1) - 1 - 2^r(2m+1) + 1 = 2^r$.

The second estimate in each case takes into account the restriction on $|v|$. Observe that $S_{2,r,m}$ and $S_{3,r,m}$ each contain at most one element of a given length. Therefore the elements in each set may be replaced with their lengths. In the case of $S_{2,r,m}$, we are estimating the cardinality of a set of natural numbers whose biggest element is at most $2^s + 1$ and where the difference between two elements, or between 0 and the smallest element (if any), is at least $2^{r+1}m + 1$ (corresponding to the smallest allowable value of k). Thus the cardinality of the set is bounded above by $\frac{2^s+1}{2^{r+1}m+1}$. A similar argument yields the second estimate in the case of $S_{3,r,m}$. □

Claim 5.4.2. — *Let s be a positive integer. Then $p^*_{S_2}(2^s+1) \leq 2 + 2^s(3+\sqrt{2})$.*

Proof of Claim 5.4.2. — 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

$$p^*_{S_2}(2^s+1) \leq 3 + \sum_{r=0}^{s-1} \sum_{m=0}^{2^{s-r-1}} \mathrm{Card}(\{v \in S_{2,r,m} \mid |v| \leq 2^s+1\})$$

and applying Claim 5.4.1 yields

$$p^*_{S_2}(2^s+1) \leq 3 + \sum_{r=0}^{s-1} \sum_{m=0}^{2^{s-r-1}} \min\left\{2^r, \frac{2^s+1}{2^{r+1}m+1}\right\}. \tag{5}$$

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

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

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

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

as required. □

Claim 5.4.3. — *For each positive integer n we have $p^*_{S_2}(n) \leq n(6+2\sqrt{2})$.*

Proof of Claim 5.4.3. — 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 $2^s \leq 2n - 2$. Using Claim 5.4.2 together with the fact that $p^*_{S_2}$ is a non-decreasing function, we obtain

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

as required. □

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

Claim 5.4.4. — *Let n be a positive integer. Then $p^*_{S_3}(n) \leq n(6 + 2\sqrt{2})$.*

Proof of Claim 5.4.4. — The proof for S_3 is analogous to that for S_2 in Claims 5.4.2 and 5.4.3. 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

$$p^*_{S_3}(2^s + 1) \leq 2 + \sum_{r=0}^{s-1} \sum_{m=0}^{2^{s-r-1}} \mathrm{Card}(\{v \in S_{3,r,m} \mid |v| \leq 2^s + 1\})$$

Applying Claim 5.4.1 gives

$$p^*_{S_3}(2^s + 1) \leq 2 + \sum_{r=0}^{s-1} \sum_{m=0}^{2^{s-r-1}} \min\left\{2^r, \frac{2^s + 1}{2^r(2m + 1) + 1}\right\}. \tag{6}$$

The claim now follows by observing that the right hand side of (6) is less than the right hand side of (5). □

Claims 5.4.3 and 5.4.4 complete the proof of Proposition 5.3. □

This concludes the proof of Theorem 5.1. □

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*

$$p_S^*(n) = \Omega\left(\sum_{p=1}^{\left\lfloor \frac{n-2}{2(2k-1)} \right\rfloor} f(p)\right).$$

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

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

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 $(p+q)(2k-1) \leq n-2$, $p \geq 2k-1$, and $q \leq f(p)$, 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 $p \geq 2k-1$ 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 $|b(a^p b^q)^{2k-1} a| \leq n$. As $\mathrm{Fac}(x) \subseteq S^k$, there exists a factorization $b(a^p b^q)^{2k-1} a = u_1 u_2 \cdots u_k$ with $u_i \in S$. Of $2k$ occurrences of ba, at most $k-1$ lie across boundaries of u_i. There 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 $s_{p,q} = u_j$. As $ba^p b^q a$ is not a factor of $b(a^{p'} b^{q'})^{2k-1} a$ when $(p,q) \neq (p',q')$, it follows that $s_{p,q} \neq s_{p',q'}$. □

Let

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

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 $f(p) \leq p$ 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 $2k-1$ and $\frac{n-2}{2(2k-1)}$ satisfies the conditions $(p+q)(2k-1) \leq n-2$ and $p \geq 2k-1$. Since for each such p there are $f(p)$ possible values for the second coordinate q, for all n sufficiently large we have

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

Whence

$$p_S^*(n) = \Omega\left(\sum_{p=1}^{\left\lfloor\frac{n-2}{2(2k-1)}\right\rfloor} f(p)\right).$$

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 $(a^p b^q)^p$ 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 $ab^q a^p b$. Words of this last form 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 $u = (a^p b^q)^p$ 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. From this theorem, if $\mathrm{Fac}(x) \subseteq S^k$ for some language S, then $p_S^*(n) = \Omega\left(\sum_{p=1}^{\left\lfloor \frac{n-2}{2(2k-1)} \right\rfloor} f(p)\right)$. Given any positive M, we can find p_0 such that $f(p_0) \geq M$; then, since $f(n)$ is non-decreasing,

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

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 3.6. □

Corollary 6.3. — *For each $\alpha \in (0,1)$, 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 $\alpha \in (0,1)$. 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 a finite $k \geq 1$, then

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

Thus $c^*(x) \geq \alpha + 1$, and hence $c(x) \geq \alpha$ by Corollary 3.8. □

7. Open questions

Some immediate questions that we cannot answer concern properties of words of quadratic complexity. We have shown in Section 5 that Theorem 1.1 cannot be directly generalized to words of quadratic complexity and concatenation of three languages. But is it true that each word of complexity $O(n^2)$ has zero cost? Is it true at least for morphic words, like the example from Section 5? If it is, then what is an upper bound for the 0-dimension of such a word? These questions could be worthy of a separate study.

BIBLIOGRAPHY

[1] J.-P. Allouche & J. Shallit – *Automatic sequences, theory, applications, generalizations*, Cambridge University Press, 2003.

[2] J. Cassaigne – "Special factors of sequences with linear subword complexity", Proceedings of DLT 1995, World Sci. Publishing, Singapore, 1996, p. 25–34.

[3] J. Cassaigne, A. Frid, S. Puzynina & L. Zamboni – "A characterisation of words of sub-linear complexity", *Bull. AMS.*

[4] ———, "Subword complexity and decomposition of the set of factors", LNCS, no. 8634, Proceedings of MFCS 2014, Springer, 2014, p. 147–158.

[5] J. Cassaigne & F. Nicolas – "Combinatorics, automata and number theory", Encyclopedia Math. Appl., no. 135, ch. Factor complexity, p. 163–247, Cambridge Univ. Press, 2010.

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

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

[8] J.-J. Pansiot – "Complexité des facteurs des mots infinis engendrés par morphismes itérés", LNCS, no. 172, Proceedings of ICALP 1984, Springer, Heidelberg, 1984, p. 380–389.