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 $u = u_0 u_1 \cdots u_{n-1}(\cdots)$ 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. □