Doctopic: Algorithms, automata, complexity and games ARTICLE IN PRESS TCS:11664

Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

An algorithm for the word entropy ☆

Sébastien Ferenczi [a,b,*], Christian Mauduit [c], Carlos Gustavo Moreira [d]

[a] IMPA – CNRS UMI 2924, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, RJ, Brazil
[b] Institut de Mathématiques de Marseille, UMR 7373 CNRS, 163, avenue de Luminy, 13288 Marseille Cedex 9, France
[c] Université d'Aix-Marseille and Institut Universitaire de France, Institut de Mathématiques de Marseille, UMR 7373 CNRS, 163, avenue de Luminy, 13288 Marseille Cedex 9, France
[d] Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, RJ, Brazil

ARTICLE INFO

Article history:
Received 5 March 2018
Received in revised form 20 June 2018
Accepted 26 June 2018
Available online xxxx

Keywords:
Combinatorics on words
Symbolic dynamics
Entropy

ABSTRACT

For any infinite word w on a finite alphabet A, the complexity function p_w of w is the sequence counting, for each non-negative n, the number $p_w(n)$ of words of length n on the alphabet A that are factors of the infinite word w and the entropy of w is the quantity $E(w) = \lim_{n\to\infty} \frac{1}{n} \log p_w(n)$. For any given function f with exponential growth, Mauduit and Moreira introduced in [10] the notion of word entropy $E_W(f) = \sup\{E(w), w \in \mathbb{A}^{\mathbb{N}}, p_w \le f\}$ and showed its links with fractal dimensions of sets of infinite sequences with complexity function bounded by f. The goal of this work is to give an algorithm to estimate with arbitrary precision $E_W(f)$ from finitely many values of f.

1. Introduction

This work concerns the little-explored field of word combinatorics in positive entropy, which means the study of infinite words on a finite alphabet with a complexity function (see Definition 2.2) of exponential growth. There are not many results on this topic, besides the well-known one of Grillenberger [6] who built symbolic systems of any given entropy.

Mauduit and Moreira introduced in [10] new notions in this context with the arithmetic motivation to study sets of numbers from the interval $[0, 1]$ whose expansion (in a given base q) has a complexity function bounded by a given function f. The determination of the Hausdorff dimension of these sets gave rise to a new quantity $E_W(f)$, called *word entropy* of f, which turns to be equal to the topological entropy of the shift on the set of corresponding expansions.

The computation of $E_W(f)$ is trivial when $E_0(f)$, the *exponential growth rate* of f (defined in (1)), is equal to zero or if f is itself a complexity function. Otherwise, results can be surprising, even when f is very regular: for example, in [10] it is shown that for the function f defined for any non-negative integer n by $f(n) = 3^{\lceil \frac{n}{2} \rceil}$, we have $E_W(f) = \log(\frac{1+\sqrt{5}}{2})$. Another striking result (see Theorem 2.9 from [11]) says that if f verifies the quite natural conditions $(\mathcal{C}^*)$ (see Definition 3.2), then the ratio $E_W(f)/E_0(f)$ lies always in the interval $]\frac{1}{2}, 1]$ and moreover we have

$$\inf\{\frac{E_W(f)}{E_0(f)}), f \text{ satisfies } (\mathcal{C}^*)\} = \frac{1}{2}.$$

☆ This work was supported by CNPq, FAPERJ and the Agence Nationale de la Recherche project ANR-14-CE34-0009 MUDERA. The authors wish to thank the referee for his very careful scrutiny.

* Corresponding author.
E-mail address: ssferenczi@gmail.com (S. Ferenczi).

https://doi.org/10.1016/j.tcs.2018.06.041
0304-3975/© 2018 Published by Elsevier B.V.

Please cite this article in press as: S. Ferenczi et al., An algorithm for the word entropy, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2018.06.041

Indeed, in the overwhelming majority of cases, we do not have access to an exact value of the word entropy. Thus in this work we propose an algorithm to get an approximate value of the word entropy, using in depth the combinatorial properties of the symbolic system.

2. Definitions and notations

We denote by q a fixed integer greater or equal to 2, by A the finite alphabet $A = \{0, 1, \ldots, q-1\}$, by $A^* = \bigcup_{k\geq 0} A^k$ the set of finite words on the alphabet A and by $A^{\mathbb{N}}$ the set of infinite words (or infinite sequences of letters) on the alphabet A. More generally, if $\Sigma \subset A^*$, we denote by $\Sigma^{\mathbb{N}}$ the set of infinite words obtained by concatenating elements of Σ. If $w \in A^{\mathbb{N}}$ we denote by $L(w)$ the set of finite factors of w:

$$L(w) = \{v \in A^*,\ \exists (v', v'') \in A^* \times A^{\mathbb{N}},\ w = v'vv''\}$$

and, for any non-negative integer n, we write $L_n(w) = L(w) \cap A^n$. For any $Y \subset A^{\mathbb{N}}$ and $n \in \mathbb{N}$ we denote $L_n(Y) = \bigcup_{w\in Y} L_n(w)$. If $w \in A^n, n \in \mathbb{N}$ we denote $|w| = n$ the length of the word w and if S is a finite set, we denote by $|S|$ the number of elements of S. For any $(a, b) \in \mathbb{R}^2$ with $a \leq b$, we denote by $[\![a, b]\!]$ the set $[a, b] \cap \mathbb{Z}$ and for any x real number, we denote $\lfloor x \rfloor = \max\{n \in \mathbb{Z}, n \leq x\}$, $\lceil x \rceil = \min\{n \in \mathbb{Z}, x \leq n\}$ and $\{x\} = x - \lfloor x \rfloor$.

Let us recall the following classical lemma concerning sub-additive sequences due to Fekete [4]:

Lemma 2.1. *If $(a_n)_{n\geq 1}$ is a sequence of real numbers such that $a_{n+n'} \leq a_n + a_{n'}$ for any positive integers n and n', then the sequence $\left(\frac{a_n}{n}\right)_{n\geq 1}$ converges to $\inf_{n\geq 1} \frac{a_n}{n}$.*

Definition 2.2. The *complexity function* of $w \in A^{\mathbb{N}}$ is defined for any non-negative integer n by $p_w(n) = |L_n(w)|$.

For any $w \in A^{\mathbb{N}}$ and for any $(n, n') \in \mathbb{N}^2$ we have $L_{n+n'}(w) \subset L_n(w)L_{n'}(w)$ so that $p_w(n + n') \leq p_w(n)p_w(n')$ and it follows from Lemma 2.1 that for any $w \in A^{\mathbb{N}}$, the sequence $\left(\frac{1}{n} \log p_w(n)\right)_{n\geq 1}$ converges to $\inf_{n\geq 1} \frac{1}{n} \log p_w(n)$. We denote

$$E(w) = \lim_{n\to\infty} \frac{1}{n} \log p_w(n) = h_{top}(X(w), T)$$

the topological entropy of the symbolic dynamical system $(X(w), T)$ where T is the one-sided shift on $A^{\mathbb{N}}$ and $X = \overline{orb_T(w)}$ is the closure of the orbit of w under the action of T in $A^{\mathbb{N}}$ ($A^{\mathbb{N}}$ is equipped with the product topology of the discrete topology on A, i.e. the topology induced by the distance $d(w, w') = \exp(-\min\{n \in \mathbb{N} |\ w_n \neq w'_n\})$).

The complexity function gives information about the statistical properties of an infinite sequence of letters. In this sense, it constitutes one possible way to measure the random behaviour of an infinite sequence: see [14,5,13].

3. Exponential rate of growth and word entropy of a function

For any given function f from $\mathbb{N}$ to $\mathbb{R}^+$, we denote

$$W(f) = \{w \in A^{\mathbb{N}}, p_w(n) \leq f(n), \forall n \in \mathbb{N}\},$$
$$\mathcal{L}_n(f) = \bigcup_{w\in W(f)} L_n(w)$$

and $E_0(f)$ the limiting lower exponential growth rate of f

$$E_0(f) = \liminf_{n\to\infty} \frac{1}{n} \log f(n). \tag{1}$$

For any $(n, n') \in \mathbb{N}^2$ we have $\mathcal{L}_{n+n'}(f) \subset \mathcal{L}_n(f)\mathcal{L}_{n'}(f)$ so that the sequence $\left(\frac{1}{n} \log |\mathcal{L}_n(f)|\right)_{n\geq 1}$ converges to $\inf_{n\geq 1} \frac{1}{n} \log |\mathcal{L}_n(f)|$, which is the topological entropy of the subshift $(W(f), T)$:

$$h_{top}(W(f), T) = \lim_{n\to+\infty} \frac{1}{n} \log |\mathcal{L}_n(f)| = \inf_{n\geq 1} \frac{1}{n} \log |\mathcal{L}_n(f)|.$$

The notion of w-entropy (or word-entropy) of f is defined in [10] as follows:

Definition 3.1. If f is a function from $\mathbb{N}$ to $\mathbb{R}^+$, the *w-entropy* (or *word entropy*) of f is the quantity

$$E_W(f) = \sup_{w\in W(f)} E(w).$$