Zeitschrift: L'Enseignement Mathématique

Herausgeber: Commission Internationale de l'Enseignement Mathématique

Band: 44 (1998)

Heft: 1-2: L'ENSEIGNEMENT MATHÉMATIQUE

Artikel: THREE DISTANCE THEOREMS AND COMBINATORICS ON WORDS

Autor: Alessandri, Pascal / Berthé, Valérie

Kapitel: 2.2 The graph of words

DOI: https://doi.org/10.5169/seals-63900

affine complexity is not necessarily obtained as a coding of rotation. Didier gives in [23] a characterization of codings of rotations. See also [46], where Rote studies the case of sequences of complexity $p(n) = 2n$, for every n. However, if the complexity of a sequence u has the form $p(n) = n + k$, for n large enough, then u is the image of a Sturmian sequence by a morphism, up to a prefix of finite length (see for instance [22] or [1]).

2.2 THE GRAPH OF WORDS

The aim of this section is to introduce the Rauzy graph of words of a sequence, in order to obtain results concerning the frequencies of factors of this sequence. This follows an idea of Dekking, who expressed the block frequencies for the Fibonacci sequence by using the graph of words (see [20] and also [8]). Note that Boshernitzan also introduces in [8] a graph for interval exchange maps (homeomorphic to the Rauzy graph of words) in order to prove Theorem 3, which can be seen as a result on frequencies.

Let us note that precise knowledge of the frequencies of a sequence with values in a finite alphabet $\mathcal{A}$ allows a precise description of the measure associated with the dynamical system $(\overline{\mathcal{O}(u)}, T)$: here T denotes the one-sided shift which associates to a sequence $(u_n)_{n\in\mathbf{N}}$ the sequence $(u_{n+1})_{n\in\mathbf{N}}$ and $\overline{\mathcal{O}(u)}$ is the orbit closure under the shift T of the sequence u in $\mathcal{A}^{\mathbf{N}}$, equipped with the product of the discrete topologies (it is easily seen that $\overline{\mathcal{O}(u)}$ is the set of sequences of factors belonging to the set of factors of u). Indeed, we define a probability measure μ on the Borel sets of $\overline{\mathcal{O}(u)}$ as follows: the measure μ is the unique T-invariant measure defined by assigning to each cylinder $[w]$ corresponding to the sequences of $\overline{\mathcal{O}(u)}$ of prefix w, the frequency of w, for any finite block w with letters from $\mathcal{A}$. Let us note that a dynamical system obtained via a coding of irrational rotation is *uniquely ergodic*, i.e., there exists a unique T-invariant probability measure on this dynamical system, which is thus determined by the block frequencies.

The Rauzy graph Γ_n of words of length n of a sequence with values in a finite alphabet is an oriented graph (see, for instance, [41]), which is a subgraph of the de Bruijn graph of words. Its vertices are the factors of length n of the sequence and the edges are defined as follows: there is an edge from U to V if V follows U in the sequence, i.e., more precisely, if there exists a word W and two letters x and y such that $U = xW$, $V = Wy$ and xWy is a factor of the sequence (such an edge is labelled by xWy). Thus there are $p(n+1)$ edges and $p(n)$ vertices, where $p(n)$ denotes the complexity function. A sequence is said to be *recurrent* if every factor appears at least twice, or equivalently if every factor appears an infinite number of times in this sequence. For instance, codings of rotations are recurrent. Note that the Rauzy graphs of words of a sequence are always connected; furthermore, they are strongly connected if and only if this sequence is recurrent.

If B is a factor, then a letter x such that Bx (respectively, xB) is also a factor is called *right extension* (respectively, *left extension*). Let U be a vertex of the graph. Denote by U^+ the number of edges of Γ_n with origin U and U^- the number of edges of Γ_n with end vertex U. In other words, U^+ (respectively, U^-) counts the number of right (respectively, left) extensions of U. Note that

$$p(n+1) - p(n) = \sum_{U \in V(\Gamma_n)} (U^+ - 1) = \sum_{U \in V(\Gamma_n)} (U^- - 1),$$

where $V(\Gamma_n)$ is the vertex set of Γ_n.

In this section we restrict ourselves to sequences with values in a finite alphabet, for which the frequencies exist. Note that the function which associates to an edge labelled by xWy the frequency of the factor xWy is a *flow*. Indeed, it satisfies Kirchhoff's current law: the total current flowing into each vertex is equal to the total current leaving the vertex. This common value is equal to the frequency of the word corresponding to this vertex. Let us see how to deduce, from the topology of a graph of words, information on the number of frequencies for factors of given length. We will use the following obvious result.

LEMMA 2. *Let U and V be two vertices joined by an edge such that $U^+ = 1$ and $V^- = 1$. Then the two factors U and V have the same frequency.*

A *branch* of the graph Γ_n is a maximal directed path of consecutive vertices $(U_1, \ldots, U_m)$ (possibly $m = 1$), satisfying

$$U_i^+ = 1, \text{ for } i < m, \quad U_i^- = 1, \text{ for } i > 1 .$$

Therefore, the vertices of a branch have the same frequency and the number of frequencies of factors of given length is bounded by the number of branches of the corresponding graph, as expressed below (for a proof of this result due to Boshernitzan, see [8]).

THEOREM 6. *For a recurrent sequence of complexity function $(p(n))$, the frequencies of factors of given length, say n, take at most $3(p(n+1) - p(n))$ values.*

REMARK. In fact, one can prove the following: the frequencies of factors of length n take at most $p(n+1)-p(n)+r_n+l_n$ values, where r_n (respectively, l_n) denotes the number of factors having more than one right (respectively, left) extension.

We deduce from this theorem that if $p(n+1)-p(n)$ is uniformly bounded with n, the frequencies of factors of given length take a finite number of values. Indeed, using a theorem of Cassaigne quoted below (see [10]), we can easily state the following corollary.

THEOREM 7. *If the complexity $p(n)$ of a sequence with values in a finite alphabet is sub-affine then $p(n+1)-p(n)$ is bounded.*

COROLLARY 1. *If a sequence over a finite alphabet has a sub-affine complexity, then the frequencies of its factors of given length take a finite number of values.*

Examples of sequences with sub-affine complexity function include the fixed point of a uniform substitution (i.e., of a substitution such that the images of the letters have the same length), the coding of a rotation or the coding of the orbit of a point under an interval exchange map with respect to the intervals of the interval exchange map.

2.3 FREQUENCIES OF FACTORS OF STURMIAN SEQUENCES

In particular, in the Sturmian case ($p(n) = n+1$, for every integer n), Theorem 6 implies the following result (see [3]).

THEOREM 8. *The frequencies of factors of given length of a Sturmian sequence take at most three values.*

Consider a Sturmian sequence of angle α. We have seen in Section 2.1 that the frequency of a factor $w_1 \cdots w_n$ of u is equal to the length of the interval

$$I(w_1, \ldots, w_n) = \bigcap_{j=0}^{n-1} R^{-j}(I_{w_{j+1}}),$$

and that these sets $I(w_1, \ldots, w_n)$ are exactly the intervals bounded by the points

$$0,\ \{1-\alpha\}, \ldots, \{n(1-\alpha)\}\ .$$