

# Rank and symbolic complexity

Sébastien Ferenczi  
Institut de Mathématiques de Luminy  
CNRS - UPR 9016  
Case 930 - 163 avenue de Luminy  
F13288 Marseille Cedex 9  
France  
e-mail: ferenczi at iml.univ-mrs.fr

February 7, 1996

## Abstract

We investigate the relation between the complexity function of a sequence, that is the number  $p(n)$  of its factors of length  $n$ , and the rank of the associated dynamical system, that is the number of Rokhlin towers required to approximate it. We prove that if the rank is one, then  $\liminf_{n \rightarrow +\infty} \frac{p(n)}{n^2} \leq \frac{1}{2}$ , but give examples with  $\limsup_{n \rightarrow +\infty} \frac{p(n)}{G(n)} = 1$  for any prescribed function  $G$  with  $G(n) = o(a^n)$  for every  $a > 1$ . We give exact computations for examples of the "staircase" type, which are strongly mixing systems with quadratic complexity. Conversely, for minimal sequences, if  $p(n) < an + b$  for some  $a \geq 1$ , the rank is at most  $2[a]$ , with bounded strings of spacers, and the system is generated by a finite number of substitutions.

Given a dynamical system, there are several notions indicating that it has a simple structure. One is the notion of **rank**, defined in [ORN-RUD-WEI] to formalize some constructions initiated by [CHA]; it is purely measure-theoretic, but leads to symbolic constructions, with systems defined by sequences on a finite alphabet. Another one, which may be tracked back to [HED-MOR1], is the combinatorial notion of **complexity function**, of languages or sequences; since the famous works of Hedlund and Morse, it is

known to provide much information about dynamical systems associated to sequences.

Systems of finite rank, or, better, of rank one, are in some sense the simplest dynamical systems we know, while sequences of sub-affine or at least sub-polynomial complexity are the simplest sequences we can build. Now, these two notions can be used in the same setting, the setting of symbolic topological dynamical systems; in the most famous classes of examples, the substitutions and the interval exchanges, low rank is known to co-exist with slowly growing complexity functions.

We explore here the exact relations between these notions, and show that rank one systems may have rather wild complexity functions: they are always below a fixed polynomial of degree 2 for infinitely many values of  $n$ , but may have peaks with any prescribed sub-exponential growth; that is the quickest we may expect as the topological entropy is zero. This yields a hitherto unknown consequence for rotations of a multidimensional torus. The proof uses an algorithm which gives the exact values of the complexity function for most rank one systems; using it, we show that the famous Smorodinsky - Adams' example has a sub-quadratic complexity; then we build some new examples of rank one systems, to get the above-mentioned peaks, or to get a sub-quadratic complexity function with unbounded second-order differences; we also have some upper bounds valid when the number of spacers (see definition below) added at each stage of the construction does not grow too fast. In the other direction, sub-affine complexity, with the additional property of minimality, implies finite rank, with an explicit bound on the rank, and a stronger property: we show that such systems can be generated by a finite number of substitutions; this number has explicit bounds in the particular case where the differences of the complexity function are bounded by 2.

## 1 Definitions

We have to recall that  $N$  is the set of nonnegative integers and  $N^*$  the set of positive integers.

For a finite alphabet  $A$ , a **word** or **block** is a finite string of elements in  $A$ ; the set of all finite words on  $A$  is denoted by  $A^*$ . The **concatenation** of two words  $v = v_1 \dots v_r$  and  $w = w_1 \dots w_s$  is the word  $vw = v_1 \dots v_r w_1 \dots w_s$ . A word  $v = v_1 \dots v_r$  is said to **occur** in a sequence  $(x_n)$  if there exists  $m$  such that

$x_m = v_1, \dots, x_{m+r-1} = v_r.$

A **symbolic system**  $(X, T)$  is the topological dynamical system consisting of a closed set  $X$  of (in this paper, two-sided) sequences on a finite alphabet  $A$ , and the shift  $T$  on  $X$ , defined by  $(Tx)_n = x_{n+1}$ , equipped with the usual (product) topology; we call the **language of the system** the set of all finite words  $w_1\dots w_k$  in  $A^*$  which occur in elements of  $X$ . We denote by  $L(X)$  this set, and by  $L_n(X)$  its intersection with the set of words of length  $n$ .

**Definition 1** *For a symbolic system  $(X, S)$ , the **symbolic complexity** of  $X$  is the function which associates to each positive integer  $n$  the cardinality of the set  $L_n(X)$ , denoted by  $p_X(n)$ .*

Of course, this function has no reason to be invariant under any notion of isomorphism. However, we are interested in rather slowly growing symbolic complexities; and happily there is a partial result, which is a straightforward consequence of the fact that a topological isomorphism is a finitary isomorphism:

**Lemma 1** *If  $(X, T)$  and  $(Y, T)$  are topologically isomorphic symbolic systems, then there exists a constant  $c$  such that, for all  $n > c$ ,*

$$p_X(n - c) \leq p_Y(n) \leq p_X(n + c);$$

*hence a relation like  $p_X(n) \leq an^k + o(n^k)$  when  $n \rightarrow +\infty$  is preserved by topological isomorphism; and the same is true if we replace  $(p(n) \leq)$  by  $(p(n) \geq)$ , or  $o(n^k)$  by  $O(n^k)$ . The boundedness of first-order or second-order differences of the complexity function is also preserved.*

Furthermore, the knowledge of the complexity function may give access to the structure of the system; we recall the well-known fact that if  $p_X(n) \leq n$  for at least one  $n$ , or if  $p(n + 1) = p(n)$  for at least one  $n$ , then  $X$  is made of sequences of the type  $u^\infty v w^\infty$  for finite words  $u, v, w$ , and in fact  $p_X(n)$  is bounded ([HED-MOR2]). This proves also that some functions, like  $[\log n]$  for example, cannot be complexity functions, and raises the vast open question: which functions from  $N^*$  to  $N^*$  are complexity functions?

The notion of rank one is basically a measure-theoretic notion, and we refer the reader to [ORN-RUD-WEI] for the original definition (see also Proposition 4 below and the following discussion). However, it was shown in [KAL] that every (measure-theoretic) rank one system has a topological model, which we shall call a **standard model of a rank one system**:

**Definition 2** *A standard model of a rank one system is the following symbolic dynamical system  $(X, T)$ : given sequences of positive integers  $q_n, n \in \mathbb{N}$  and  $a_{n,i}, n \in \mathbb{N}, 1 \leq i \leq q_n$ , such that, if  $h_n$  is defined by*

$$h_0 = 1, \quad h_{n+1} = (q_n + 1)h_n + \sum_{j=1}^{q_n} a_{n,j},$$

and  $S_n$  by

$$S_n = \sum_{i=1}^{q_n} a_{n,i} = h_{n+1} - (q_n + 1)h_n,$$

then

$$\sum_{n=1}^{+\infty} \frac{S_n}{(q_n + 1)h_n} < +\infty; \tag{1}$$

and such that, if we define the words  $B_n$  on the alphabet  $(0, 1)$  by

$$B_0 = 0, \quad B_{n+1} = B_n 1^{a_{n,1}} B_n 1^{a_{n,2}} B_n \dots B_n 1^{a_{n,q_n}} B_n, \tag{2}$$

then  $T$  is the shift on the set  $X$  of sequences  $(x_n)$  of  $(0, 1)^{\mathbb{Z}}$  such that for every pair of integers  $s < t$ ,  $(x_s \dots x_t)$  is a subsequence of  $B_n$  for some  $n$ .

Note that the topological system defined above is not necessarily minimal, but the only possible non-dense orbit is the sequence identically equal to 1, and this happens if and only if the numbers  $a_{n,j}$  are not bounded. We recall that the one” symbols added between words  $B_n$  are called **spacers**.

From the measure-theoretic point of view, there is only one non-atomic invariant probability measure, denoted by  $\mu$ : it gives to the cylinder set  $(x_1 = a_1, \dots, x_p = a_p)$  a measure equal to the limit as  $n$  goes to infinity of the frequency of occurrences of the word  $a_1 \dots a_p$  in the word  $B_n$  (this limit exists because of (1)). There may be other invariant measures, such as the measure  $\delta_1$  which gives mass one to the sequence identically equal to 1, and all convex combinations of  $\mu$  and  $\delta_1$ .

Now, we can define a (measure-theoretic) **rank one system** to be any system which is measure-theoretically isomorphic to a standard model of a rank one system (that is: the system in Definition 1 for some sequences  $q_n$  and  $a_{n,i}$  satisfying (1)), equipped with its unique non-atomic invariant probability; a given rank one system is completely defined by giving its standard model and more precisely the recursion formula (2) giving the words (more usually called blocks)  $B_n$ .

By abuse of notation, we shall call **complexity** of a given rank one system, defined by a standard model, the complexity function of its given standard model; of course, the standard model is not unique, and this complexity is in no way a measure-theoretic invariant of the system; but, because of Lemma 1 above, this complexity will yield precise information on any topological system isomorphic to the given standard model of a rank one system, and even, in some cases, may imply some properties of the underlying measure-theoretic system (see Corollary 2 below).

## 2 The mixing rank one of Smorodinsky - Adams

Before trying to estimate the complexity of a general rank one system, we give a complete computation for an example, which will prove in some sense fairly typical.

Among rank one systems, the most striking (and complicated) are the ones which are (**measure-theoretically strongly**) **mixing**: this means that,  $\mu$  being the invariant nonatomic probability, for any measurable subsets  $E$  and  $F$  of  $X$ ,

$$\mu(T^n E \cap F) \rightarrow \mu(E)\mu(F)$$

when  $n \rightarrow \pm\infty$ . The existence of mixing rank one systems has been known since [ORN], but the first explicit examples were given in [ADA-FRI]; the simplest and most famous one was defined a long time ago by Smorodinsky (to our knowledge, this was not published before [ADA-FRI]), but was not proved to be mixing until [ADA]. It is given by the recursion formula

$$B_{n+1} = B_n B_n 1 B_n 1 1 B_n \dots (1)^{n-1} B_n,$$

$B_0 = 0 = B_1$ ,  $B_2 = B_1 B_1$ ,  $B_3 = B_2 B_2 1 B_2$  and we continue as above. So

$h_0 = h_1 = 1, h_2 = 2$ , then  $h_{n+1} = (n + 1)h_n + \frac{n(n-1)}{2}$ .

We want to compute  $p(l)$ , the number of different words of length  $l$  belonging to the language of the system. We look at all possible words  $w$  of length  $l = l(w)$  satisfying  $h_n + 1 \leq l(w) \leq h_{n+1}$ ; such a word  $w$  will be of the form

$$u1^{c_1}b1^{c_2}\dots b1^{c_d}v,$$

where  $b$  is the word  $B_n$ ,  $u$  is a (possibly empty and possibly non-strict) suffix of  $B_n$ ,  $v$  is a (possibly empty and possibly non-strict) prefix of  $B_n$ ,  $d, c_1, \dots, c_d$  are natural integers. We call this expression an **( $n$ -)decomposition** of  $w$ . To eliminate some obvious ambiguities, we specify that  $c_1 \neq 0$  whenever  $u$  is empty and  $c_d \neq 0$  whenever  $v$  is empty; however, even with these restrictions the decomposition has no reason to be unique. Let  $a$  be the length of  $u$ ,  $e$  be the length of  $v$ . We denote by  $C$  the finite sequence  $(c_1, \dots, c_d)$ : this is a sequence of  $d$  integers, with  $1 \leq d \leq n + 1$ , each one lying between 0 and  $l$  a priori; we call it an **( $n$ -)configuration** of  $w$ . It depends on  $n$  as we look only at the lengths of the groups of spacers situated inside  $w$  and between words  $B_n$ . We call  $a, e, d, c_1, \dots, c_d$  or in short  $a, e, d, C$  the **parameters** of this decomposition of  $w$ .

Not every configuration is allowed by the system; their description is complicated by the problem of the beginnings and endings of words. We define first the **( $n$ -)block configurations** of the system;  $C = (c_1, \dots, c_d)$  is an **( $n$ -)block configuration** if it is a subsequence of  $d$  consecutive terms of one of the sequences  $(0, 1, 2, \dots, n - 1, k, 0, 1, 2, \dots, n - 1)$  for any natural integer  $k$ .

Now, from the definition of the system, we see that the possible parameters  $a, e, d, c_1, \dots, c_d$  associated to a word  $w$  with length  $l$  satisfying  $h_n + 1 \leq l \leq h_{n+1}$  must satisfy the conditions:

- (3) there exists an **( $n$ -)block configuration**  $S = (s_1, \dots, s_d)$  such that  $c_1 \leq s_1, c_2 = s_2, \dots, c_{d-1} = s_{d-1}, c_d \leq s_d$ ;
- (4) if  $a \neq 0, c_1 = s_1$ ;
- (5) if  $a = 0, c_1 \neq 0$ ;
- (6) if  $e \neq 0, c_d = s_d$ ;

- (7) if  $e = 0$ ,  $c_d \neq 0$ ;
- (8)  $l = a + e + (d - 1)h_n + c_1 + \dots + c_d$ ;
- (9)  $0 \leq a \leq h_n$ ;
- (10)  $0 \leq e \leq h_n$ .

These conditions are the only ones: for integers  $a, d, e, l$  and a sequence  $C = (c_1, \dots, c_d)$  satisfying (3) to (10) and  $h_n + 1 \leq l \leq h_{n+1}$ , we can define a word of length  $l$  by  $w = u1^{c_1}b\dots b1^{c_d}v$ , where  $u$  is the suffix of length  $a$  of  $B_n$  and  $v$  is the prefix of  $B_n$  of length  $e$ . This word, which we call  $W(l, a, C)$  (the parameter  $e$  can be deduced from the others using (8)), belongs to the language of the system and admits at least one decomposition with parameters  $a, d, e, C$ .

When conditions (3) to (10) are satisfied, we say that  $(l, a, C)$  is an **( $n$ -)productive triple**. We define then an auxiliary quantity  $q(l)$ :

*For a fixed  $l$ , let  $n$  be the integer satisfying  $h_n + 1 \leq l \leq h_{n+1}$ ; we look at all the ( $n$ -)productive triples which are of the form  $(l, a, C)$ , and we say that  $(l, a, C) = (l, a, c_1, \dots, c_d)$  is different from  $(l, a', C') = (l, a', c'_1, \dots, c'_d)$  whenever  $a \neq a'$  or  $d \neq d'$  or  $c_i \neq c'_i$  for some  $1 \leq i \leq d \wedge d'$ ; note that we may have  $(l, a, C)$  different from  $(l, a', C')$  but  $W(l, a, C) = W(l, a', C')$ . We define  $q(l)$  as the number of different ( $n$ -)productive triples of the form  $(l, a, C)$ .*

The previous analysis implies that  $p(l) \leq q(l)$ .

We call  $H(n, l)$  the following hypothesis, which has three equivalent forms:

- each word of length  $l$  has a unique ( $n$ -)decomposition;
- if  $(l, a, C)$  and  $(l, a', C')$  are two different ( $n$ -)productive triples, then  $W(l, a, C) \neq W(l, a', C')$ ;
- $p(l) = q(l)$ .

**Lemma 2**  $H(n, l)$  implies  $H(n, l + 1)$  for  $h_n + 1 \leq l \leq h_{n+1}$ .

**Proof**

Suppose  $H(n, l)$  is true and that  $W(l + 1, a, C) = W(l + 1, a', C')$ . We look at the prefix of length  $l$  of this word; it is clearly equal to  $W(l, a_1, C_1)$ , where  $(l, a_1, C_1)$  is the triple deduced from  $(l + 1, a, C)$  by

- replacing  $e$  by  $e - 1$  if  $e \geq 1$ , except if  $e = 1$  and  $c_d = 0$ ;
- replacing  $e$  by  $h_n$ ,  $d$  by  $d - 1$  and deleting  $c_d$  if  $e = 1$  and  $c_d = 0$  or  $e = 0$  and  $c_d = 1$ ;
- replacing  $c_d$  by  $c_d - 1$  if  $e = 0$  and  $c_d > 1$ ;

all other parameters remain unchanged. Hence  $W(l, a_1, C_1) = W(l, a'_1, C'_1)$ , hence  $(l, a_1, C_1) = (l, a'_1, C'_1)$  by  $H(n, l)$ . But in that case, we look at  $e$  and  $e'$ , looking at the  $3 \times 3 = 9$  possibilities. This leads

- either to  $e = 0$ ,  $d = d'$ ,  $e' = 1$ , and  $c_d = c'_d + 1$  (or the symmetrical case), hence  $W(l + 1, a, C)$  ends by 1 and  $W(l + 1, a', C')$  ends by the first letter of  $B_n$ , which is a zero, and they cannot be identical;
- or to impossible cases like  $e - 1 = h_n$ ;
- or else to  $e = e'$  and all parameters of  $(l + 1, a, C)$  and  $(l + 1, a', C')$  are identical, which is what we want.

Hence  $H(n, l + 1)$  is true. QED

**Lemma 3** *For Smorodinsky - Adams' map,  $H(n, h_n + 1)$  is true if and only if  $p(h_n + 1) = \frac{1}{2}(h_n^2 + 5h_n + 2)$ .*

**Proof**

We have just to compute  $q(h_n + 1)$ . The triple  $(h_n + 1, a, C)$  is a productive triple if and only if

- $0 \leq a \leq h_n$ ,
- $d = 1$ ,
- $0 \leq c_1 \leq h_n + 1 - a$ ,
- $(a, c_1) \neq (0, 0)$ .

Hence

$$q(h_n + 1) = \sum_{a=0}^{h_n} (h_n + 2 - a) - 1 = \sum_{t=2}^{h_n+2} (t) - 1 = \frac{1}{2}(h_n + 2)(h_n + 3) - 2.$$

QED

**Proposition 1** *The complexity function of Smorodinsky - Adams' map is given by the formulas:*

$$\begin{aligned} p(1) &= 2, \\ p(2) &= 4, \\ p(3) &= 7, \\ p(4) &= 12, \\ p(5) &= 18, \\ p(6) &= 26, \\ p(7) &= 35, \end{aligned}$$

$$\begin{aligned} p(l+1) - p(l) &= l+2 \quad \text{for } l = h_n + 1, \\ p(l+1) - p(l) &= l+i+1 \quad \text{for } l = h_n + i, \quad 1 \leq i \leq n-1, \\ p(l+1) - p(l) &= l+n \quad \text{for } h_n + n - 1 \leq l \leq 2h_n + 1, \\ p(l+1) - p(l) &= l+n-i \quad \text{for } 2h_n + 2i \leq l \leq 2h_n + 2i + 1, \quad 1 \leq i \leq n-2, \\ p(l+1) - p(l) &= l+2 \quad \text{for } 2h_n + 2n - 2 \leq l \leq 3h_n + 2n - 3, \\ p(l+1) - p(l) &= l+1 \quad \text{for } 3h_n + 2n - 2 \leq l \leq h_{n+1}, \end{aligned}$$

for all  $n \geq 3$ .

**Proof**

We suppose, as a recursion hypothesis, that  $H(n, h_n + 1)$  is true. We call a **successor** of the word  $W$  any word of the form  $W0$  or  $W1$  which exists in the language of the system. If  $W$  is in the language of the system, it has at least one and at most two successors. The recursion hypothesis allows us to decide, by looking at their  $(n-)$ decomposition, which words of length  $l$  have two successors, for  $h_n + 1 \leq l \leq h_{n+1}$ . The word  $W$  has two successors if and only if it ends with a full word  $B_n$ , or with spacers, and the block configurations allow to put after  $W$  either a full word  $B_n$  (which begins with a zero) or an extra spacer. This happens for:

- every word of the form  $A_i 1^i$ , for  $0 \leq i \leq l$ , where  $A_i$  is the (possibly empty) suffix of length  $l - i$  of  $B_{n+1}$ ;
- every suffix of the word  $1^{n-2} B_n 1^{n-1} B_n B_n$ ; this creates one word with two successors, other than the ones already described, for every  $h_n + 1 \leq l \leq 3h_n + 2n - 3$ ;
- every suffix of  $1^{i-1} B_n 1^i B_n$  for all  $1 \leq i \leq n - 2$ ; this adds one more word with two successors for every  $1 \leq i \leq n - 2$  and every  $h_n + 1 + i \leq l \leq 2h_n + 2i - 1$ .

Now,  $p(h_{n+1} + 1) - p(h_n + 1)$  is equal to the total number of words with two successors with lengths  $h_n + 1 \leq l \leq h_{n+1}$ . Because of the recursion hypothesis, we know already that  $p(h_n + 1) = \sum_{t=2}^{h_n+2} (t) - 1$ , and we can write that

$$p(h_{n+1} + 1) - p(h_n + 1) = \sum_{l=h_n+1}^{h_{n+1}} (l + 1) + 2h_n + 2n - 3 + \sum_{i=1}^{n-2} (h_n + i - 1).$$

So we get that

$$p(h_{n+1} + 1) = \sum_{t=1}^{h_{n+1}} (t + 1) - 1 + (n + 1)h_n + 2n - 1 + \sum_{i=0}^{n-3} (i) = \sum_{t=1}^{h_{n+1}+1} (t + 1) - 1,$$

and this result is equivalent to  $H(n + 1, h_{n+1} + 1)$ ; hence we can continue the recursion.

We can compute by hand the first values of  $p(n)$  and see that  $H(3, 8)$  is true; then the proposition follows from the fact that  $p(l + 1) - p(l)$  is the number of words of length  $l$  which have two successors, and these words are identified in the computation above. QED

The complexity function satisfies

$$p(l) \geq \frac{1}{2}(l^2 + 3l - 2)$$

for all  $l$ , with equality for each  $l = h_n + 1$ ,  $n \geq 3$ . On the other end, after the range of quickest growth, we have

$$p(2h_n + 1) = \frac{(2h_n + 1)(2h_n + 2)}{2} + nh_n + (n - 1)\left(1 - \frac{n}{2}\right).$$

Taking into account that  $n! \leq h_n \leq 2(n!)$ , we get that

$$p(l) \leq \frac{l(l+1) + lK(l)}{2}$$

for all  $n$ , where

$$K(l) \sim \frac{\log l}{\log \log l}$$

when  $l$  is large, and that  $p(l)$  exceeds infinitely often values of the form

$$\frac{1}{2} \left( l^2 + lO \left( \frac{\log l}{\log \log l} \right) \right).$$

Note that the second-order differences of the sequence  $p(l)$  are bounded.

In [FER], we conjectured that the complexity of mixing systems has to be at least super-polynomial; this conjecture is strongly contradicted by Smorodinsky - Adams' example, and we shall see that it cannot be more than partially true for the general Ornstein's example; in view of Corollary 3 below, we hazard a guess that this example realizes the lowest possible complexity of mixing systems, and wait confidently for the next counter-example.

### 3 Algorithm and other examples

#### 3.1

For the general case of a rank one system defined by the recursion formula (2), we can apply the same analysis, with only one difference, the actual expression of the  $(n-)$ block configurations. Here  $C = (c_1, \dots, c_d)$  will be an  $(n-)$ block configuration if  $d \leq q_n + 1$  and  $C$  is a subsequence of  $d$  consecutive terms of one of the sequences  $(a_{n,1}, \dots, a_{n,q_n}, a_{r,s}, a_{n,1}, \dots, a_{n,q_n})$  for any  $r \geq n + 1$  and any  $1 \leq s \leq q_r$ . It is still clear of course that  $p(l) \leq q(l)$ ; the hypothesis  $H(n, l)$  will be the same, and Lemma 2 will hold, for every rank one system.

When  $H(n, l)$  is true for every  $h_n + 1 \leq l \leq h_{n+1}$ , we say that the system is  **$(n-)$ rhythmic** (by analogy with a property of substitutions which has the same purpose). A rank one system which is  $(n-)$ rhythmic for all  $n$  large enough will simply be called **rhythmic**. In fact, most rank one systems will

prove to be rhythmic, the only exceptions being those with obvious symmetries in the recursion formulas.

Hence the method used for Smorodinsky - Adams' example gives an algorithm for checking whether a given rank one system is  $(n-)$ rhythmic for any  $n$  and then, if it is found to be  $(n-)$ rhythmic for all  $n$  large enough, for computing completely its complexity; we may sum it up as follows:

- compute  $q(h_n + 1)$  for every  $n$ ;
- make the recursion hypothesis  $p(h_n + 1) = q(h_n + 1)$ ;
- using Lemma 2 and the recursion hypothesis, compute  $p(l + 1) - p(l)$  for every  $h_n + 1 \leq l \leq h_{n+1}$  by counting the number of productive triples with two successors;
- if the value obtained for  $p(h_{n+1} + 1)$  agrees with the one given by  $q(h_{n+1} + 1)$ , then the system is  $(n + 1-)$ rhythmic and the recursion hypothesis may be carried one step further.

### 3.2

This algorithm applies immediately to the **general staircase rank one** given ([ADA-FRI]) by the recursion formula

$$B_{n+1} = B_n B_n 1 B_n 1 1 B_n \dots (1)^{p_n - 1} B_n,$$

for any sequence  $p_n$  compatible with (1). The reader may check that it is rhythmic, and that the complexity is given by a formula analogous to Proposition 1, where the differences  $p(l + 1) - p(l)$  are allowed to climb as far as  $l + p_n$  for some  $l$  between  $h_n$  and  $2h_n$ ; as the maximal value of  $p_n$  compatible with (1) must be in  $o(h_n)$ , the complexity is still sub-quadratic but, for suitable  $p_n$ , exceeds infinitely often values of the form

$$\frac{1}{2} (l^2 + lO(l^{1-\epsilon}))$$

for any given  $\epsilon > 0$ . The second-order differences of  $p$  are still bounded.

### 3.3

Let us now build the following rank one system: at stage  $n$ ,  $h_n$  being known, let

$$r_n = \left\lceil \left( \frac{h_n}{n^2} \right)^{\frac{1}{3}} \right\rceil,$$

$q_n = (r_n + 1)(r_n)^2$ , and  $a_{n,1}, \dots, a_{n,q_n}$  be the sequence of all pairs  $(x_1, x_2)$ ,  $0 \leq x_i \leq r_n - 1$ , written successively in lexicographical order. We can take the first stages similar to the ones in Smorodinsky - Adams' map, until the formula giving  $r_n$  gives non-trivial values. The definition is compatible with (1), and it is easy, though tedious, to check that the transformation is rhythmic. We check also that

$$p(l) < \frac{1}{2}(l^2 + l^{\frac{5}{3}})$$

for every  $l$  large enough (or simply, by Lemma 4 below, the complexity must be sub-quadratic). It is then immediate to check that for all  $n$  large enough  $(p(h_n + r_n) - p(h_n + r_n - 1)) - (p(h_n + r_n - 1) - p(h_n + r_n - 2)) \geq r_n - 1$ .

This yields the following new result:

**Corollary 1** *There exist rank one systems whose complexity functions are sub-quadratic but have unbounded second-order differences.*

This is particularly noticeable as, on the other side, the question whether there exist, for any system, sub-affine complexity functions with unbounded first-order differences has been very recently answered by Cassaigne ([CAS]), and the answer is negative.

If we take some suitable sequence  $k_n$  going to infinity with  $n$ , some  $r_n$  close to  $\frac{1}{k_n^{k_n+1}}$ , and replace all pairs in the above definition by all  $k_n$ -uples at stage  $n$ , we get some rhythmic rank one systems where  $S_n$  is close to  $h_n$ , and the differences  $p(l+1) - p(l)$  exceed  $l + 1 + l^{1-\epsilon}$  infinitely often, for any  $\epsilon > 0$ ; they will be used in the next section.

### 3.4

For a non-rhythmic system, this algorithm will fail; for what to do in that case, we refer the reader to [FER]: in this paper, we compute the complexity

for rank one systems of the del Junco - Rudolph type, given by the recursion formulas  $B_{n+1} = (B_n)^{p_n} 1 (B_n)^{q_n}$ ; they are the only example we know of famous rank one systems which are not rhythmic. The idea is to guess which words of length  $h_n + 1$  will have more than one decomposition, and to put that in the form of a recursion hypothesis  $p(h_n + 1) = q'(h_n + 1)$ . Then we count the number of words with two successors, either because they have one decomposition with two successors producing two different words, or because they have two decompositions whose successors produce two different words. If the obtained value of  $p(h_{n+1} + 1)$  agrees with  $q'(h_{n+1} + 1)$ , we can continue and the complexity function is known.

There are many examples of rank one system, rhythmic or non-rhythmic, whose complexity function is sub-affine (see [FER]), hence far below the values in Proposition 1.

## 4 General rank one systems

### 4.1

We want now to give some upper bounds of the growth of  $p(l)$  for the general rank one system given by (2).

We take a general rank one system with  $q_n$ ,  $a_{n,i}$ ,  $h_n$  and  $S_n$  given as in the definition.

**Proposition 2** *For any rank one system,*

$$p(l) \leq \frac{1}{2}(l^2 + 3l - 2)$$

*whenever  $l = h_n + 1$  for any  $n \geq 1$ . When all the  $a_{n,i}$  are not greater than  $K$ , we may replace this estimate, for the same values of  $l$ , by*

$$p(l) \leq (K + 1)l - \frac{1}{2}(K^2 - K + 2).$$

#### **Proof**

We estimate  $q(h_n + 1)$ : the triple  $(h_n + 1, a, C)$  is a productive triple if and only if

- $0 \leq a \leq h_n$ ;

- $d = 1$ ;
- $0 \leq c_1 \leq (h_n + 1 - a) \wedge K$ ,  $K$  being taken equal to  $+\infty$  when the  $a_{n,i}$  are unbounded;
- $(a, c_1) \neq (0, 0)$ ;
- if  $a + c_1 \neq h_n + 1$ , then  $c_1$  is a value taken by some  $a_{r,s}$  for  $r \geq n$  and  $1 \leq s \leq q_r$ ;
- if  $a + c_1 = h_n + 1$ , then  $c_1$  is smaller or equal to a value taken by some  $a_{r,s}$  for  $r \geq n$  and  $1 \leq s \leq q_r$ .

Hence their number can be bounded as in Lemma 3, and the proposition follows from  $p(l) \leq q(l)$ . QED

As irrational rotations are of rank one as measure-theoretic systems ([deJ]), Proposition 2 translates into:

**Corollary 2** *For any  $k \geq 1$ , let  $R$  be an irrational rotation of the torus  $T^k$ ; there exists a measurable set  $A$  such that the partition  $(A, X \setminus A)$  is a generating partition for the system, and such that the sequence  $1_A(R^l x)$ ,  $l \in \mathbb{N}$ , has a complexity  $p_x(l)$  such that*

$$\liminf_{l \rightarrow +\infty} \frac{p_x(l)}{l^2} \leq \frac{1}{2}$$

for almost all  $x \in T^k$ .

## 4.2

The following lemma is interesting for the techniques involved as well as for its content. It gives an upper bound of the complexity if we know the number of spacers added at each stage, which is itself limited by (1); its proof describes rather precisely how the complexity function grows, at least for rhythmic rank one systems, and which systems are likely to have the highest complexity.

As we want to give a bound of  $p(l)$ , it is enough to have a bound of  $q(l)$ ; note that a bound of the differences of  $q(l)$  will yield a bound of the sequence  $p$ , but not of its differences, except for rhythmic systems.

**Lemma 4** Let  $D(l) = q(l+1) - q(l)$  be the differences of the sequence  $q$  for a general rank one system given by the recursion formula (2). We have

$$D(l) \leq l + 1 + kS_n$$

when  $kh_n \leq l \leq (k+1)h_n$ . The estimate can be replaced by

$$D(l) \leq K + kS_n$$

for the same values of  $l$  when all the  $a_{n,i}$  are not greater than  $K$ .

**Proof**

$D(l)$  is just the number of productive triples  $(l, a, C)$  with two successors, that is the productive triples satisfying

- either  $e = h_n$  and both  $(l+1, a, C_1)$  and  $(l+1, a, C_2)$  are productive triples, where  $C_1 = (c_1, \dots, c_d, 0)$  and  $C_2 = (c_1, \dots, c_d, 1)$ ,
- or  $e = 0$  and both  $(l+1, a, C_3)$  and  $(l+1, a, C)$  are productive triples, where  $C_3 = (c_1, \dots, c_{d-1}, c_d + 1)$ .

We are interested in the growth of  $D(l)$ , that is the number of new productive triples appearing for a given  $l$ . Now, there will be a productive triple with two successors if and only if there exists an *extensible (n-)word configuration*, that is an  $(n)$ -block configuration  $(c_1, \dots, c_d)$  such that  $(c_1, \dots, c_{d-1}, c_d + z)$  is also an  $(n)$ -block configuration for some integer  $z \geq 1$ . Each extensible  $(n)$ -block configuration  $(c_1, \dots, c_d)$  gives birth to several productive triples  $(l, a, C)$  with two successors, namely:

- any triple with  $e = 0$  and  $C = (c_k, \dots, c_d)$  for any  $1 \leq k \leq d$ , if  $c_d \neq 0$ ;
- any triple with  $e = h_n$  and  $C = (c_k, \dots, c_{d-1})$  for any  $1 \leq k \leq d-1$ , if  $c_d = 0$ ;
- any triple with  $e = 0$ ,  $a = 0$  and  $C = (c, c_{k+1}, \dots, c_d)$  for any  $1 \leq k \leq d$  and  $1 \leq c \leq c_k$ , if  $c_d \neq 0$ ;
- any triple with  $e = h_n$ ,  $a = 0$  and  $C = (c, c_{k+1}, \dots, c_{d-1})$  for any  $1 \leq k \leq d-1$  and  $1 \leq c \leq c_k$ , if  $c_d = 0$ .

Though this may look complicated, the important fact is that for fixed  $l$ , from one extensible word configuration comes at most one productive triple  $(l, a, C)$ .

Among all the  $(n-)$ block configurations, we consider the ones which are equal to  $(a_{n,j}, \dots, a_{n,q_n}, i)$  for any  $1 \leq j \leq q_n$  and  $i \geq 0$ ; we call these the *final*  $(n-)$ block configurations. Let  $F(l)$  be the number of productive triples  $(l, a, C)$  with two successors coming from final extensible block configuration; for fixed  $l$ , there is at most one such triple for each (possible) final block configuration whose last element is  $i$ ,  $0 \leq i \leq l$ ; hence we have

$$F(l) \leq l + 1$$

(this bound does not look very sharp, as in general all these block configurations need not exist or be extensible; note however that for Smorodinsky - Adams' example we had equality). In the same way, when the  $a_{n,i}$  are bounded, we have

$$F(l) \leq K.$$

We must now compute  $E(l) = D(l) - F(l)$ , the number of productive triples with two successors which do not come from any final word configurations; we order the non-final extensible block configurations by looking at their last elements: let  $e_1 < e_2 < \dots$  be all the possible values of  $c_d$  for non-final extensible block configurations (which must have at least two elements), let  $f_{i,1} < f_{i,2} < \dots$  be all the possible values of  $c_{d-1}$  when  $c_d = e_i$ , and so on.

The non-final extensible configurations finishing by  $e_i$  give birth to productive triples with two successors, which are the same as some words already counted in  $F(l)$  for the first values of  $l \geq h_n + 1$ ; then one new productive triple with two successors appears for  $l = h_n + e_i + (f_{i,1} \wedge a_{n,q_n}) + 1$  and for all values of  $l$  in an interval beginning from this value (its length is not relevant for our bound); when the value of  $l$  continues to increase, one new (different from the ones previously counted) productive triple with two successors will appear for each value  $l = h_n + e_i + (f_{i,j} \wedge a_{n,q_n}) + 1$  and for all values of  $l$  in an interval onwards. In turn, if there are extensible non-final block configurations with  $c_d = e_i$ ,  $c_{d-1} = f_{i,j}$ ,  $c_{d-2} = g_{i,j,k}$ , then one new configuration with two successors will appear for each value  $l = 2h_n + e_i + f_{i,j} + g_{i,j,k} + 1$  if  $f_{i,j} \neq a_{n,q_n}$ , or for each value  $l = 2h_n + e_i + a_{n,q_n} + (g_{i,j,k} \wedge a_{n,q_n-1}) + 1$  if

$f_{i,j} = a_{n,q_n}$ , and for all values of  $l$  in an interval onwards. Continuing in the same way, we can enumerate every new productive triple with two successors if we know all the extensible block configurations.

All the coefficients  $e, f, g$  contained in the non-final extensible block configurations are some  $a_{n,i}$  for some  $1 \leq i \leq q_n$  (if not, the configuration must be final or non-extensible) ; as the configurations must be extensible, we know also that, for each last coefficient  $e$ , some  $e + z$  must be one of the  $a_{n,i}$  for some  $z \geq 1$  (if the  $e + z$  required by the extensibility was not one of the  $a_{n,i}$ , the configuration would be final); hence we can bound  $E(l)$  by the number of points of the form  $kh_n + a_{n,i} + \dots + a_{n,i+k} + 1$  lying between  $h_n + 1$  and  $l$  satisfying the condition (of extensibility) that there exists  $J(i, k) \neq i$  with  $a_{n,i+j} = a_{n,J+j}$  if  $0 \leq j \leq k - 1$  and  $a_{n,i+k} < a_{n,J+k}$ ; for fixed  $k$ , two values  $i \neq i'$  giving the same value to the sums  $a_{n,i} + \dots + a_{n,i+k}$  are counted effectively twice except when  $a_{n,i+j} = a_{n,i'+j}$  for all  $0 \leq j \leq k$ . Or else, for  $kh_n \leq l \leq (k + 1)h_n$ ,  $E(l)$  is smaller than the sum, over all  $1 \leq k' \leq k$ , of the number of different  $k' + 1$ -tuples of the form  $a_{n,i}, \dots, a_{n,i+k'}$ , such that there exists  $J(i, k') \neq i$  with  $a_{n,i+j} = a_{n,J+j}$  if  $0 \leq j \leq k' - 1$  and  $a_{n,i+k'} < a_{n,J+k'}$ .

We fix some  $1 \leq k \leq q_n + 1$ ; let  $r$  be the number of different  $k$ -tuples satisfying the above condition; we partition them according to the values of the initial  $(k - 1)$ -tuple of coordinates; let  $t$  be the number of these values and  $u_1, \dots, u_t$  the number of possible values of the last coordinate corresponding to each of them; we have

$$r = u_1 + \dots + u_t.$$

But, because of the condition of extensibility, for the  $i$ -th value of the initial  $(k - 1)$ -tuple, the last coordinate of the  $k$ -tuple must take at least  $(u_i + 1)$  different values  $v_{i,1}, \dots, v_{i,u_i+1}$ ; and all these values  $v_{i,j}$  must be taken by different last coordinates  $a_{n,I(i,j)}$ . Hence

$$S_n \geq \sum_{i=1}^t \sum_{j=1}^{u_i+1} v_{i,j} \geq \sum_{i=1}^t \sum_{j=1}^{u_i+1} (j - 1) = \sum_{i=1}^t \frac{u_i(u_i + 1)}{2}.$$

Now it is clear from these conditions that the maximal value of  $r$  is reached when all  $u_i$  are equal to 1, and is then equal to  $S_n$ .

Hence we can bound  $E(l)$  by  $kS_n$  for  $kh_n \leq l \leq (k + 1)h_n$ . The lemma follows from  $D(l) = E(l) + F(l)$ . QED

### 4.3

Lemma 4 implies for example that if the  $S_n$ , which represent the number of spacers added at each stage, are bounded by some  $S$ , then the complexity is sub-affine, satisfying for all  $l$

$$p(l) \leq (K + S)l + b$$

for some  $b$  (as of course the  $a_{n,i}$  have to be bounded also, by some  $K$ ). This generalizes the various results in [FER], where complexities were computed for rank one systems with  $S_n = 1$ , and were smaller than  $2l - 1$ . We can thus build many examples of sub-affine complexity functions; to know whether rank one systems, or any systems, may have any sub-affine function with bounded differences as complexity function is an open question, suggested to the author by Mauduit: in a rather negative direction, let us point out that the examples considered in [FER], with  $S = 1$ , all have complexity functions larger than  $\frac{3n}{2}$ , with higher growth concentrated near the values  $h_n$ , and with growing gaps between them.

If  $S_n \leq h_n$  for all  $n$  large enough, then the complexity is at most quadratic; the systems of the end of section 3.3 prove that the estimate in Lemma 4 cannot be improved significantly when  $S_n$  is not too large.

When  $S_n$  is large, on the contrary, even within the limitations of (1) the only thing we may say is that the complexity function must be in  $o(a^l)$  for every  $a > 1$ ; this can be deduced from the proof by looking more closely, or derived from the well-known fact that the topological entropy of rank one systems is zero, and that implies a sub-exponential growth of the complexity. This, in turn, cannot be improved, even when all the  $a_{n,i}$  take only values 0 or 1:

**Proposition 3** *Let  $G$  be any function from  $N$  to  $N$  such that*

$$G(n) = o(a^n)$$

*for every  $a > 1$ . Then there exists a rank one system with*

$$p(l) > G(l)$$

*for infinitely many  $l$ . This system satisfies also*

$$p(l) \geq \frac{1}{2}(l^2 + 3l - 2)$$

for every  $l$ , with equality for infinitely many values of  $l$ .

There exists also, for any integer  $K \geq 1$ , a rank one system with  $a_{n,i} \leq K$  for every  $n$  and  $i$ , such that

$$p(l) > G(l)$$

for infinitely many values of  $l$ . This system satisfies also

$$p(l) \geq (K + 1)l - \frac{1}{2}(K^2 - K + 2)$$

for every  $l$ , with equality for infinitely many values of  $l$ .

### Proof

We build a rank one system by the following recursion formulas: at stage  $n$ ,  $h_n$  being known, we choose some integers  $k_n$  and  $r_n$ , and some

$$q_n > \frac{n^2 k_n (r_n)^{k_n+1}}{2h_n};$$

then let all the  $a_{n,i}$ ,  $1 \leq i \leq k_n (r_n)^{k_n}$  be all the possible  $k_n$ -tuples of integers ranging from 0 to  $r_n - 1$ , written successively in lexicographical order, and let all the  $a_{n,i}$ ,  $k_n (r_n)^{k_n} + 1 \leq i \leq q_n$ , be equal to 0.

The system satisfies (1); we check it is rhythmic; and, for  $k \leq k_n$ ,

$$p(kh_n + kr_n) \geq (r_n)^k h_n.$$

Now, we fix a sequence  $r_n$ ; for fixed  $r_n$  and  $h_n$ , we choose  $a$  such that

$$\log a < \frac{\log r_n}{h_n + r_n},$$

and then choose  $k_n$  large enough to have

$$G(k_n h_n + k_n r_n) < a^{k_n h_n + k_n r_n}.$$

So the first part of the proposition is proved if we take the  $r_n$  unbounded, and the second part if we take the  $r_n$  all equal to  $K + 1$ . The lower bounds come from the rhythmic property and the fact that the  $a_{n,i}$  take all possible values between 0 and infinity, or between 0 and  $K$ . QED

## 5 Systems of sub-affine complexity

### 5.1

As a partial converse to the partial results of the last section, we have the following result, which was proved for a particular case in [ARN-RAU] :

**Proposition 4** *Let  $X$  be a minimal symbolic system on a finite alphabet  $A$ , such that*

$$p_X(n) \leq an + b$$

*for some  $a \geq 1$ ; then there exist at most  $2[a]$  families of words  $B_{n,r}$ ,  $n \in \mathbb{N}$ ,  $1 \leq r \leq 2[a]$ , given by the recursion formulas*

$$B_{n,r} = B_{n-1,L(n,r,1)}B_{n-1,L(n,r,2)}\cdots B_{n-1,L(n,r,c(n,r))},$$

*and a constant  $K$  such that for any  $m > 0$  there exists  $N$  such that, for every word  $W$  of length greater than  $N$  occurring in the language of the system, there exists  $n$  such that*

$$W = SC_1B_1C_2B_2\dots C_kB_kC_{k+1}P,$$

*where each  $B_i$  is one of the  $B_{n,r}$ ,  $1 \leq r \leq 2[a]$ , each  $C_i$  is a concatenation of such words  $B_{n,r}$ , each  $B_i$  has length at least  $m$ , each  $C_i$  has length at most  $K$ ,  $S$  is a suffix of some  $B_{n,r}$ ,  $P$  is a prefix of some  $B_{n,r}$ .*

#### Proof

*Here and in all this section, we shall suppose, without loss of generality, that  $a$  is an integer.*

The hypothesis implies that there exists a sequence  $l_n \rightarrow +\infty$  such that

$$0 \leq p(l_n + 1) - p(l_n) \leq a.$$

We recall that, because of the minimality,  $X$  is the closed orbit of one sequence  $u$ , where each finite word appears infinitely often with bounded gaps. If  $u$  is ultimately periodic, the conclusion holds, so we suppose henceforth that  $u$  is not ultimately periodic, and hence  $p(m+1) - p(m) \geq 1$  for every  $m$ .

For each  $m$ , we define the *graph of words* of length  $m$ , denoted by  $G_m$ , in the following way:

$G_m$  is the graph whose vertices are the words of  $L_m$  and where there is an edge from  $v$  to  $w$  whenever  $w = ue$ ,  $v = e'u$  and  $e'ue \in L_{m+1}$  for any  $e \in A$ ,  $e' \in A$  and  $u \in L_{m-1}$ .

An edge between  $ue$  and  $e'u$ , where  $e \in A$ ,  $e' \in A$  and  $e'ue \in L_{m+1}$ , is then labelled with  $e$ .

Following Rauzy, we call **( $m$ -)segment** any path in  $G_m$

- beginning at one vertex with more than one outgoing edge,
- ending at one vertex with more than one outgoing edge, and
- passing only through vertices with one outgoing edge.

An **( $m$ -)circuit** will be any path beginning at one vertex with more than one outgoing edge and ending at the same vertex without passing through it before.

For each  $m$  there is a finite (but not necessarily bounded) number of ( $m$ -)segments, and together they contain every edge of  $G_m$ . Also, if a word  $W$  of  $L_m$  is a vertex of  $G_m$ , then its suffix of length  $m - 1$  must be a vertex of  $G_{m-1}$  with a greater number of outgoing edges; hence the vertices of  $G_m$  with more than one outgoing edge must be of the form  $eV$ , for  $e \in A$  and  $V$  a vertex of  $G_{m-1}$  with more than one outgoing edge. Given an ( $m$ -)segment from one vertex  $D$  to another vertex  $E$  in  $G_m$ , we partition it into paths going from one to the next vertex of the form  $eU$  with  $e \in A$  and  $U$  a vertex of  $G_{m-1}$  with more than one outgoing edge; the path from  $eU$  to  $e'V$  in  $G_m$  has the same label as the path from  $U$  to  $V$  in  $G_{m-1}$ ; hence the labels of ( $m$ -)segments are concatenations of labels of ( $m - 1$ -)segments (we say the labels of ( $m$ -)segments are **nested**). By the same reasoning, the labels of ( $m$ -)circuits are concatenations of labels of ( $m - 1$ -)circuits.

Hence, if the minimum length of the ( $m$ -)segments does not tend to infinity with  $m$ , it means that at least one ( $m$ -)segment will not be a concatenation of at least two ( $m - 1$ -)segments for all  $m$  big enough. Hence there exists a label  $B$  which is a label of one ( $m$ -)segment for all  $m$  big enough. So we divide the set of ( $m$ -)segments into **long ( $m$ -)segments** whose lengths tend

all to infinity with  $m$ , and **short ( $m$ -)segments** whose lengths and labels are ultimately constant when  $m$  tends to infinity (this is not a partition, one of the classes may be empty). The same division may be made for ( $m$ -)circuits.

But the ( $l_n$ -)segments have an additional property: let  $W_1(n), \dots, W_{d(n)}(n)$  be all the vertices in  $G_{l_n}$  with more than one outgoing edge, and  $f_1(n), \dots, f_{d(n)}(n)$  their respective number of outgoing edges; we have

$$p(l_n + 1) - p(l_n) = \sum_{i=1}^{d(n)} (f_i(n) - 1),$$

hence

$$d(n) \leq a$$

and

$$\sum_{i=1}^{d(n)} f_i(n) \leq 2a.$$

It is clear that in  $G_{l_n}$  there are exactly  $\sum_{i=1}^{d(n)} f_i(n)$  different ( $l_n$ -)segments, hence their number is bounded by  $2a$ . Also, as their union contains every edge of  $G_{l_n}$ , the maximum of their lengths must tend to infinity with  $n$ .

The number of ( $l_n$ -)circuits does not need to be bounded, but *there is no short ( $l_n$ -)circuit*: if there is one, as there is only a finite number of possible origins for it, we can find a word  $B$  which is the label of an ( $l_{k_n}$ -)circuit going from some  $D_n$  to  $D_n$ , for a sequence  $k_n \rightarrow +\infty$ , and a sequence of vertices  $D_n$  such that  $D_{n-1}$  is a suffix of  $D_n$  for all  $n$ ; but that implies that the word  $D_n$  is equal to a suffix of itself followed by  $B$ ; hence  $D_n$  must end with the word  $B$  for all  $n$ , with the word  $BB$  for  $n$  large enough, with the word  $B^p$  for  $n$  greater than some  $N(p)$ , and this contradicts minimality. Hence of course there is no short ( $m$ -)circuit when  $m$  tends to infinity.

Let  $m$  be one of the  $l_n$ ; any path in  $G_m$  is a concatenation of ( $m$ -)segments, possibly truncated at the beginning and the end. In this concatenation, there are short segments, with at most  $2a - 1$  different labels, hence their lengths, for all  $m = l_n$  (and thence in fact for all  $m$ ), are bounded by some  $M$ ; and, if a concatenation of adjacent short segments has a length bigger than  $aM$ , then it must pass twice through one of the vertices with more than one outgoing edge, and thus contains an ( $m$ -)circuit of length smaller than  $aM$ ; this cannot happen for  $m$  large enough. Hence, in any path in any  $G_{l_n}$ , the

concatenations of adjacent short segments have a length which is bounded by some  $K$ .

Let  $W = x_1 \dots x_p$  be a word in the language of the system; then it occurs at position  $i$  in some element  $u$  of  $X$ , hence it is the label of a path from the word  $u_{i-m} \dots u_{i-1}$  to the word  $u_{i-m+p} \dots u_{i+p-1}$  in  $G_m$  for every fixed  $m$ . Let  $p$  be much larger than  $m$ , and let  $m$  be one of the  $l_n$ : then  $W$  will be a concatenation of labels of long ( $m$ -)segments, possibly truncated at the beginning and at the end, separated by labels of paths of length smaller than  $K$ .

Let now the  $B_{n,r}$  be all the labels of the ( $l_n$ -)segments; the previous analysis shows that they satisfy the conclusions of the proposition. QED

Such systems generated by at most  $2a$  arbitrarily long words, and shorter strings of letters, naturally called strings of **spacers**, with some conditions ensuring that there are not too many spacers, are already known: they are in fact topological models of measure-theoretic **systems of rank at most  $2a$** , which generalize systems of rank one. In our case, the number of consecutive spacers is bounded by  $K$ , and Proposition 4 reads: *sub-affine complexity plus minimality imply finite rank with bounded strings of spacers*; these systems generalize the rank one systems with bounded  $a_{n,i}$ .

As may be seen in its proof, Proposition 4 is in fact true as soon as  $p(l_n + 1) - p(l_n) \leq a$  for some sequence  $l_n \rightarrow +\infty$ . The bound on the rank may not be the best possible: the primitive totally irrational *exchanges of  $r$  intervals*, coded by the partition in points of discontinuity, are in general systems of rank at most  $r$  (without spacers) and of complexity  $p(l) = (r - 1)l + 1$  (folk literature). Note also that a system of rank  $r$  with bounded strings of spacers might be also of a rank smaller than  $r$  but with unbounded strings of spacers.

## 5.2

We shall now give a stronger version of Proposition 4, using the language of substitutions (see for example [QUE] for a general theory of substitutions). We recall that a **substitution** is an application  $\sigma$  from a finite alphabet  $A$  to the set  $A^*$  of all finite words on  $A$ . It extends to a morphism of  $A^*$  by  $\sigma(ab) = \sigma(a)\sigma(b)$ .

**Proposition 5** *Let  $X$  be a minimal symbolic system on a finite alphabet  $A$ , of cardinality  $a'$ , such that  $p_X(n)$  is a sub-affine function, or, equivalently ([CAS]), such that  $p_X(n+1) - p_X(n)$  is bounded for every  $n \geq 1$ ; then there exist a finite number of substitutions  $\sigma_i$ ,  $1 \leq i \leq c$ , on an alphabet  $D = (0, \dots, d-1)$ , an application  $\alpha$  from  $D$  to  $A$ , and an infinite sequence  $(1 \leq i_n \leq c, n \geq 1)$  such that*

$$\inf_{0 \leq r \leq d-1} |\sigma_{i_1} \sigma_{i_2} \dots \sigma_{i_n} r| \rightarrow +\infty$$

*when  $n \rightarrow +\infty$ , and any word of the language of the system is a subword of  $\alpha \sigma_{i_1} \sigma_{i_2} \dots \sigma_{i_n} 0$  for some  $n$ .*

**Proof**

We remark first that because of [CAS], the sub-affinity of  $p_X(n)$  implies the boundedness of its differences; hence we can apply the reasoning of Proposition 4 to our system; but here, if  $p(n+1) - p(n) \leq M$ , then for any  $m$  there are at most  $2M$  ( $m$ -)segments, and we can put  $l_m = m$  for any  $m \geq 1$ ; so we take as recursion formulas the formulas giving the ( $m$ -)segments as concatenation of ( $m-1$ -)segments.

Moreover, the number of ( $m-1$ -)segments in one ( $m$ -)segment is equal to one plus the number of vertices  $eU$ ,  $e \in A$ , on this ( $m$ -)segment such that  $U \in G_{m-1}$  has more than one outgoing edge but  $eU \in G_m$  has not; the segment cannot cross each of these points more than once, hence this number is smaller than  $M(p(1) - 1) + 1$ . So we get the conclusion of Proposition 4 with  $d_0 \leq 2M$  families of words  $B_{n,r}$  where the length  $c(n,r)$  of the recursion formulas giving the  $B_{n,r}$  as concatenations of  $B_{n-1,s}$  is always smaller than  $M(p(1) - 1) + 1$ . The length of the strings of spacers is still bounded by  $K$ .

Let now, for every fixed  $n$ , the  $W_{n,r}$  be all the words of the form  $UV$ , where  $U$  is any one of those  $B_{n,r}$  which are labels of long" segments in the above construction and  $V$  is any word of length 0 to  $K$  on the alphabet  $A$ ; for convenience, we number them from 0 to  $d-1$ ; their number  $d$  is bounded by  $2M(a')^{K+1}$ , and all labels of paths in any  $G_m$  are concatenations of the  $W_{n,r}$ , possibly truncated at the beginning and at the end, with the length of each  $W_{n,r}$ ,  $0 \leq r \leq d-1$ , tending to infinity with  $n$ .

Let  $W = x_1 \dots x_p$  be a given word of the language of the system; as the language is minimal, there exists  $p'$  such that  $W$  is a subword of any word of

length  $p'$  of the language. As the lengths of the  $W_{n,r}$  tend to infinity, for each  $n$  large enough  $W$  must be included in  $W_{n,0}$ , which is a word of the system.

Now, the  $W_{n,r}$  are also given by recursion formulas

$$W_{n,r} = W_{n-1,M(n,r,1)}W_{n-1,M(n,r,2)}\cdots W_{n-1,M(n,r,t(n,r))},$$

and the length  $t(n,r)$  of this formula is always smaller than  $a'M$ . Hence we can write

$$W_{n,r} = \alpha_0\sigma_1\sigma_2\cdots\sigma_n r$$

for  $0 \leq r \leq d-1$ , where  $\sigma_n$  is defined on the alphabet  $(0, \dots, d-1)$  by

$$\sigma_n r = M(n,r,1)M(n,r,2)\cdots M(n,r,t(n,r))$$

for  $0 \leq r \leq d-1$  and  $n \geq 1$ ,  $\alpha_0(r) = W_{0,r}$ , and  $\alpha_0$  extends naturally to a morphism from  $(0, \dots, d-1)^*$  to  $A^*$ .

The number of different substitutions  $\sigma_n$  is bounded by  $d^{da'M}$ ;  $\alpha_0$  may be written as a product of  $\alpha$  by one substitution on  $D$ , where  $\alpha$  is a map from  $D$  to  $A$ . Hence our final expression. QED

The above proposition can be read as: minimal systems with sub-affine complexity are generated by a finite number of substitutions. Using a terminology initiated by Vershik, we propose to call such systems **S-adic systems**, and we have proved one half of a conjecture of Host, showing that *minimal systems of sub-affine complexity are S-adic*. This implies that systems with a reasonably low" language complexity can be built with a reasonably low" algorithmic complexity.

We give now a measure-theoretic consequence of this result; we recall that a topological system is **uniquely ergodic** if it has only one invariant probability measure.

**Corollary 3** *A minimal and uniquely ergodic system of sub-affine complexity cannot be strongly mixing.*

**Proof**

Such a system satisfies the conclusion of Proposition 5; by making the geometric construction of *Rokhlin towers* (see for example [KAL]), we can then

find a set  $E$  such that  $\mu(E \cap T^{h_n} E) \geq \frac{1}{L}\mu(E)$ , where  $\mu$  is the invariant measure,  $h_n$  is some length of words  $W_{n,r}$ , and  $L$  is the maximum of the length of the  $\sigma_n$ . QED

Note that the fact that the lengths of the words tend to infinity, which generalizes the notion of *primitivity* of a substitution, is necessary to make the proposition nonempty. This prevents us from getting, if we know that  $p_X(n) \leq an + b$  or that  $p_X(n+1) - p_X(n) \leq M$ , a universal bound on the number of letters, and hence on the *exact rank* of the system (the analogue of the rank, but without allowing any spacers). However, in an important particular case generalizing [ARN-RAU], we do have a universal upper bound:

**Proposition 6** *For minimal systems on an alphabet with three letters such that*

$$p_X(n+1) - p_X(n) \leq 2$$

*for all  $n$  large enough, Proposition 5 is satisfied with  $d \leq 3$  and  $c \leq 3^{27}$ .*

**Proof**

Here we know that for all  $n$  large enough, there are at most two vertices in  $G_n$  with more than one outgoing edge ; we call them  $D_n$  and (if it exists)  $E_n$ . This time, the words we use to generate the language will be the labels of the ( $n$ -)circuits; they will have all the required properties if we show their number is suitably bounded.

We suppose now that, from  $D_n$  to  $D_n$ , there are at least four different ( $n$ -)circuits. This can happen only if  $E_n$  does exist, and

- either there are two ( $n$ -)segments from  $D_n$  to  $E_n$ , two ( $n$ -)segments from  $E_n$  to  $D_n$ , and the four possible paths  $D_n E_n D_n$  are allowed, that is: their labels are in the language of the system;
- or there is an ( $n$ -)segment from  $E_n$  to  $E_n$ , and a path  $D_n E_n \dots E_n D_n$  using this segment more than once is allowed.

In both cases this means that four paths of the type  $(B \text{ or } C)FE_n(G \text{ or } H)$  are allowed, with the common section  $FE_n$ . Hence, for some  $m > n$ , there exist two words of the form  $uUE_n$  and  $vUE_n$ ,  $u \in A$ ,  $v \in A$ ,  $U \in L_{m-n-1}$ , which have more than one outgoing edge in  $G_m$ . Hence these must be the only two points in  $G_m$  with two outgoing edges; hence, for any  $p \geq m$ , every

point with more than one outgoing edge in  $G_p$  must have  $E_n$  as its suffix, and every (p-)circuit is a concatenation of ( $n$ -)circuits going from  $E_n$  to  $E_n$ .

Hence this cannot happen for  $D_n$  and  $E_n$  simultaneously; around at least one of them there are at most three ( $n$ -)circuits, and we can use the labels of these three circuits to generate the language of the system. And each of these three circuits is, because of the above reasoning, a concatenation of at most three ( $n$ -)segments, hence of at most nine ( $n-1$ -)segments, hence of at most nine ( $n-1$ -)circuits. And we know the lengths of all the ( $n$ -)circuits tend to infinity. QED

### 5.3

As for the converse of the converse, that is to get an upper bound of the complexity of finite rank systems, it is a rather difficult problem. For systems generated by one substitution with some reasonable conditions (primitivity or weaker conditions), the complexity is sub-affine ([COB]). The same method should work also for many S-adic systems: the full S-adic conjecture of Host is that there exists a (possibly stronger) version of S-adicity which is equivalent to sub-affine complexity for minimal systems.

For general finite rank systems, where the number of spacers satisfies conditions similar to (1), the best we may hope for, in view of what happens for rank one systems, is that  $\liminf_{l \rightarrow +\infty} \frac{p(l)}{l^k}$  is finite for some  $k$  ( $k = 1$  if the number of consecutive spacers is bounded); but even this seems to be true only when the different words have comparable lengths. Let us remark however that, except for interval exchanges and substitutions, there are very few known systems of finite rank strictly greater than one: the composition powers  $T^k$  of mixing rank one systems are the only examples which come to mind.

## References

- [ADA] T.M. ADAMS: Smorodinsky's conjecture, *submitted*, (1993).
- [ADA-FRI] T.M. ADAMS, N.A. FRIEDMAN: Staircase mixing, *submitted*, (1992).

- [ARN-RAU] P. ARNOUX, G. RAUZY: Représentation géométrique de suites de complexité  $2n + 1$ , *Bull. Soc. Math. France* 119 (1991), p. 199-215.
- [CAS] J. CASSAIGNE: Facteurs spéciaux des suites de complexité sous-affine, *preprint*.
- [CHA] R.V. CHACON : A geometric construction of measure-preserving transformations, *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* , Univ. of California Press (1965), p. 335-360.
- [COB] A. COBHAM: Uniform tag sequences, *Math. Systems Theory* 6 (1972), p. 164-192.
- [deJ] A. del JUNCO: Transformations with discrete spectrum are stacking transformations, *Canadian J. Math.* 24 (1976), p. 836-839.
- [FER] S. FERENCZI: Les transformations de Chacon: combinatoire, structure géométrique, liens avec les systèmes de complexité  $2n + 1$ , *Bull. Soc. Math. France* 123 (1995), p. 271-292.
- [HED-MOR1] G.A HEDLUND, M. MORSE: Symbolic dynamics, *Amer. J. Math.* 60 (1938), p. 815-866.
- [HED-MOR2] G.A HEDLUND, M. MORSE: Symbolic dynamics II. Sturmian trajectories, *Amer. J. Math.* 62 (1940), p. 1-42.
- [KAL] S. KALIKOW: Twofold mixing implies threefold mixing for rank-1 transformations, *Ergodic Th. Dyn. Syst.* 4 (1984), p. 237-259.
- [ORN] D.S. ORNSTEIN: On the root problem in ergodic theory, *Proc. of the Sixth Berkeley Symposium in Mathematical Statistics and Probability*, Univ. of California Press (1970), p. 347-356.
- [ORN-RUD-WEI] D.S. ORNSTEIN, D.J. RUDOLPH, B. WEISS: Equivalence of measure-preserving transformations, *Memoirs Amer. Math. Soc.* 262 (1982).
- [QUE] M. QUEFFELEC: Substitution dynamical systems - Spectral analysis, *Lecture Notes in Math.* vol. 1294 (1987), Springer-Verlag.