ELSEVIER

Theoretical Computer Science 218 (1999) 177–195

Theoretical Computer Science

Complexity for finite factors of infinite sequences

Sébastien Ferenczi[a,*], Zoltán Kása[b]

[a] *Institut de Mathématiques de Luminy, CNRS – UPR 9016, Case 930 – 163 avenue de Luminy, F13288 Marseille Cedex 9, France*

[b] *Faculty of Mathematics and Informatics, Babeş-Bolyai University, str. Kogălniceanu 1, RO-3400 Cluj, Romania*

Abstract

We define several notions of language complexity for finite words, and use them to define and compute some new complexity functions for infinite sequences. In particular, they give a new characterization of Sturmian sequences and discriminate between Sturmian sequences with bounded or unbounded partial quotients.

Keywords: Infinite words; Complexity; Finite factors

The language complexity $p(n)$ of an infinite sequence U is the number of its factors of length n; this notion, which measures the randomness of the sequence U, has been extensively studied in these last years, see [1] and [13] for surveys; the rate of growth of $p(n)$ is a fundamental characteristic of the sequence U, and of the symbolic dynamical system associated to U. However, if the sequence U is given to us through its finite initial segments $u_0...u_N$, it is not clear which N we have to take to compute $p(n)$; and, if we have a long finite sequence and would like to know from which infinite sequence it may come, it is impossible to use $p(n)$ for that. So we need a notion of complexity for finite words, and such a notion has been proposed independently by Iványi [16] and Shallit [20]; we use this complexity $K(w)$, and a related one $C(w)$ (first suggested by Rauzy), to devise new complexity functions for an infinite sequence, denoted by $K^+(n)$, $K^-(n)$, $C^+(n)$, $C^-(n)$; these functions can be computed, or at least approximated, considering only words of length n. They give informations on the language complexity $p(n)$, while they have interesting properties of their own. The main aim of this paper is to compare the K and C complexities with the more usual p.

We first devise some methods to estimate the finite-word complexities of infinite sequences; these involve the *Rauzy graphs*, and some of their properties which have

* Corresponding author. E-mail: ferenczi@gargan.math.univ-tours.fr.

0304-3975/99/$ - see front matter
PII: S0304-3975(98)00258-8

not yet been studied; in particular, it will be very useful to compute the *dispersion function* of the sequence, which associates to n the maximum length of an allowed path without repetition of a vertex in the nth Rauzy graph. We then proceed to estimate our new complexity functions for "usual" sequences, the C-complexities being generally more accessible than the K-complexities. A particularly satisfying case is the case of *Sturmian sequences* ($p(n) = n + 1$), which can be fully characterized by their $C^+(n)$, and are the non-ultimately periodic sequences with the smallest $C^+(n)$, while $C^-(n)$ discriminates between Sturmian sequences with bounded or unbounded partial quotients. $C^+(n)$ can also be used to discriminate between some known sequences with $p(n) = 2n + 1$ or $p(n) = 2n$, namely the *Arnoux–Rauzy sequences* and *rotation sequences*. We give also example of finite-word complexities for sequences where $p(n)$ is polynomial or exponential.

1. Definitions and notations

Let U be an infinite (one-sided) sequence, $U = u_0u_1\ldots$, on a finite alphabet A.

Definition 1. The **language** of length n of u, denoted by $L_n(U)$, is the set of all factors of length n of U.

The **language complexity** of U is the function from $\mathbb{N}^*$ to $\mathbb{N}^*$ defined by $p_U(n) = \#L_n(U)$.

If w is a finite word, we define $p_w(n)$ to be the number of different factors of length n of w for $n \leqslant |w|$, and put $p_w(n) = 0$ for $n > |w|$.

We say that the word w **occurs** at place i in U if $w = u_iu_{i+1}\ldots u_{i+|w|-1}$.

The sequence U is **periodic** if $u_{n+k} = u_n$ for every $n \geqslant 0$. The sequence U is **ultimately periodic** if $u_{n+k} = u_n$ for every $n \geqslant n_0$.The sequence U is **recurrent** if every word occurring in U occurs at infinitely many places. The sequence U is **uniformly recurrent** or **minimal** if every word occurring in U occurs at infinitely many places with bounded gaps.

We define the K-complexities in the following way

Definition 2. For a finite word, the **total complexity** is

$$K(w) = \sum_{j=1}^{|w|} p_w(j),$$

while for an infinite sequence, we can define the **upper** and **lower total finite-word complexity function** by

$$K_U^+(n) = \max_i K(u_iu_{i+1}\ldots u_{i+n-1}),$$

$$K_U^-(n) = \min_i K(u_iu_{i+1}\ldots u_{i+n-1}).$$

We define the C-complexities in the following way.

Definition 3. For a finite word, the **maximal complexity** is

$$C(w) = \max_{j=1}^{|w|} p_w(j),$$

while for an infinite sequence, we can define the **upper** and **lower maximal finite-word complexity function** by

$$C_U^+(n) = \max_i C(u_i u_{i+1} \dots u_{i+n-1}),$$

$$C_U^-(n) = \min_i C(u_i u_{i+1} \dots u_{i+n-1}).$$

In the sequel, we generally omit the subscript U when the sequence is unambiguous.

We recall that for a real number x, $[x]$ is the largest integer smaller or equal to x, and we denote by $[x]'$ the smallest integer greater or equal to x.

2. General results

2.1. First bounds for finite-word complexities

Proposition 1. *If U is uniformly recurrent and non-ultimately periodic,*

$$C_U^-(n) \to +\infty, \quad K_U^-(n) - n \to +\infty,$$

when $n \to +\infty$.

Proof. If U is uniformly recurrent, for every m there exists r such that every factor of length m of U occurs in every factor of length r of U. If U is not ultimately periodic, $p_U(m) \uparrow +\infty$ if $m \to +\infty$. So, for any fixed k, we choose n such that $p_U(n) \geqslant k$, r as above and $n = 2r$; then, for every factor w of length m of U, $p_w(m/2) \geqslant k$; by stating simply that $p_w(l) \geqslant 1$ for $1 \leqslant l \leqslant n$, we get $C_U^-(n) \geqslant k$ and $K_U^-(n) \geqslant n + k$. □

Note that uniform recurrence cannot be replaced by recurrence in the above proposition, see Section 5 for counter-examples.

Lemma 1. *Suppose that $p_U(n) \leqslant f(n)$ for some continuous increasing real-valued function with $f(1) \geqslant 1$, and let $L(n)$ be the unique $1 \leqslant L \leqslant n$ such that $f(L) = n - L + 1$; then*

$$C_U^+(n) \leqslant f([L(n)]) \wedge n - [L(n)] + 1$$

and

$$K_U^+(n) \leqslant \sum_{k=1}^{[L(n)]} f(k) + \tfrac{1}{2}(n - [L(n)])(n - [L(n)] + 1).$$

Proof. Let $|w|$ be a finite factor of U, of length n, and $k \leqslant n$; then

$$p_w(k) \leqslant p_U(k) \wedge n - k + 1,$$

and then we apply the definitions. □

Corollary 1. *For any sequence on a finite alphabet*

$$K_U^+(n) \leqslant \frac{n(n+1)}{2}, \text{ [17]} \quad C_U^+(n) \leqslant n.$$

Corollary 2. *If U is ultimately periodic, $C_U^+(n)$ and $K_U^+(n)/n$ are bounded.*

The lemma is also easy to apply when $f(n) = an + b$ as $L(n) = (n+1-b)/(a+1)$; the precise result for $f(n) = n+1$ will be stated below.

Corollary 3. *If $p_U(n) \leqslant an^r$,*

$$n - C_U^+(n) = \Omega\left(\left(\frac{n}{a}\right)^{1/r}\right), \qquad \frac{n^2}{2} - K_U^+(n) = n\Omega\left(\left(\frac{n}{a}\right)^{1/r}\right).$$

Corollary 4. *For any sequence on s letters,*

$$C_U^+(n) \leqslant s^k \vee n - k,$$

$$K_U^+(n) \leqslant \frac{(n-k)(n-k+1)}{2} + s^{k+1} - 1$$

for the unique k such that $s^k + k - 1 \leqslant n < s^{k+1} + k$. In particular, $n - C_U^+(n) = \Omega(\log n / \log s)$ and $n^2/2 - K_U^+(n) = n\Omega(\log n / \log s)$.

This last result is proved in [20] (for $s = 2$, but the generalization is straightforward). It is a consequence of Lemma 1, with $p_U(n) \leqslant s^n$, and the fact that the chosen k satisfies $s^k \leqslant n - k + 1$ and $s^{k+1} > n - k$.

2.2. The shape of the complexity function of a word

The following proposition was proved in [10], see also [11]. It will help us to estimate the total complexities.

Proposition 2. *Let w be a finite word, $p_w(n)$ its complexity function. There exist $1 \leqslant n_1 \leqslant n_2 \leqslant |w|$ such that*

- $p_w(n+1) > p_w(n)$ *for* $1 \leqslant n < n_1$,
- $p_w(n+1) = p_w(n)$ *for* $n_1 \leqslant n < n_2$,
- $p_w(n+1) = p_w(n) - 1$ *for* $n_2 \leqslant n \leqslant |w|$.

2.3. Rauzy graphs and dispersion functions

We define the **Rauzy graph** Γ_n of a sequence U in the following way: the vertices are the points of $L_n(U)$, with an edge from w to w' if w and w' occur successively in U, that is if $w = av$ and $w' = vb$ for letters a and b and a word v of $L_{n-1}(U)$; we label this edge by $avb \in L_{n+1}$, and the set of edges is $L_{n+1}(U)$.

If we enumerate successively the words $u_i u_{i+1} \dots u_{i+n-1}$, $u_{i+1} \dots u_{i+n}$, up to $u_{i+k-1} \dots u_{i+k+n-2}$, we get a finite path in Γ_n; such a path in Γ_n is called an **allowed path of length** k.

If $w = w_1...w_n$ occurs in U, it defines a path in Γ_m for any $m \leqslant n$, namely $w_1...w_m$, ... $w_{n-m+1}...w_n$; it is immediate that $p_w(m)$ *is the number of different vertices on that path.*

Definition 4. A path in Γ_n is called **without repetition** if all its vertices are different. We call **dispersion function** of U and denote by $d(n)$ the function which associates to n the maximum number of vertices of an allowed path in Γ_n without repetition.

Lemma 2. *If U is not ultimately periodic, for all n, $d(n) \geqslant n+1$.*

Proof. Suppose first U is recurrent; then $d(n+1) \geqslant d(n)$ (if $u_i \dots u_{i+n-1}$, ... $u_{i+d-1} \dots u_{i+d+n-2}$ is allowed and without repetition, then $u_i u_{i+1} \dots u_{i+n}$, ... $u_{i+d-1} \dots u_{i+d+n-1}$ is allowed and without repetition); if $d(1) = 1$, u must be constant; so if $d(n) < n+1$, there exists m such that $d(m+1) = d(m)$.

Let $u_i u_{i+1} \dots u_{i+n-1}$, ... $u_{i+d-1} \dots u_{i+d+n-2}$ be an allowed path without repetition in Γ_n, of maximal number of vertices; we can take $i > 0$ as U is recurrent. We look at the $d+1$ edges $u_{i-1} \dots u_{i+n-1}$, ... $u_{i+d-1} \dots u_{i+d+n-1}$; the first d of them are different, the last d of them are different, so if the first and the last are different, they form an allowed path with $d+1$ vertices without repetition in Γ_{n+1}. Hence we must have $u_{i+d+n-1} = u_{i+n-1}$; $u_{i+d} \dots u_{i+d+n}$ must be equal to some $u_j \dots u_{j+n}$ for some $i \leqslant j \leqslant d-1$, and we must have $j = i$ as all others $u_j \dots u_{j+n}$ have a different n-prefix; so $u_{i+d+n} = u_{i+n}$. We can then apply the same reasoning to the path $u_{i+1} \dots u_{i+n}$, ... $u_{i+d} \dots u_{i+d+n-1}$, and get $u_{i+d+n+1} = u_{i+n+1}$; by iteration, we get that U is ultimately periodic, of period d.

If U is not recurrent, we apply the same reasoning after replacing Γ_n by Γ'_n, the graph whose vertices are the words of $L_n(u)$ which occur infinitely often in U, and $d(n)$ by $d'(n)$, the maximum number of vertices of an allowed path in Γ'_n without repetition of a vertex; we have still $d'(n+1) \geqslant d'(n)$ (if a word occurs infinitely often, so does its prefix, so the same remark as for d applies), and again $d'(m) = d'(m+1)$ implies ultimate periodicity. Hence we must have $d(n) \geqslant d'(n) \geqslant n+1$. □

Our Lemma 2 implies in particular the famous result [15] that whenever U is not ultimately periodic, $p_U(n) \geqslant n+1$ for all n.

2.4. Bounds using the Rauzy graph

Lemma 3. *Suppose U is not ultimately periodic, and that $d(n) \geqslant g(n)$ for all n, for some increasing function $g(n) \leqslant n+1$ (this last condition is always realised because of Lemma 2); let $M(n)$ be the unique $1 \leqslant M \leqslant n$ such that $g(M) = n - M + 1$; then for all n*

$$C_U^+(n) \geqslant g([M(n)])$$

and

$$K_U^+(n) \geqslant (g([M(n)]))^2 - 1.$$

In particular, if there exists $\frac{1}{2} \leqslant c < 1$ *such that for all* n $d(n)/(n+d(n)) \geqslant c$, *then for all* n

$$C_U^+(n) \geqslant cn + c - 1, \quad K_U^+(n) \geqslant (cn + c - 1)^2 - 1.$$

Proof. Let $m = [M(n)]$; we can find an allowed path in Γ_m, with at least $g(m)$ vertices, without repetition; let $u_i u_{i+1} \dots u_{i+m-1}$ be its first vertex; then $u_i u_{i+1} \dots u_{i+g(m)+m-2}$ contains at least $g(m)$ different factors of length m, and is a subword of $u_i \dots u_{i+n-1}$. Hence $C_U^+(n) \geqslant g(m)$. Now, as $g(k) \geqslant k+1$, and because of the shape of $p_{u_i u_{i+1} \dots u_{i+k-1}}$ dictated by Proposition 2, the smallest possible value for $K_U^+(n)$ is reached when $p_{u_i u_{i+1} \dots u_{i+k-1}}$ increases by one at a time at the left of m, and decreases by one at a time at the right of m, which gives the claimed bound for $K_U^+(n)$.

In the particular case where $g(n) = cn/(1-c)$, we have $m = [(1-c)(n+1)]$, hence the last assertions. □

Proposition 3. *For every nonultimately periodic sequence and every* n,

$$C_U^+(n) \geqslant \left[\frac{n}{2}\right] + 1, \qquad K_U^+(n) \geqslant \left[\frac{n^2}{4} + n\right].$$

Proof. This is implied immediately from Lemmas 2 and 3. □

The use of the dispersion function allows us to give lower bounds to the finite-word complexities; it can also help to precise upper bounds, but we need then to examine another quantity in the Rauzy graphs.

Lemma 4. *Let* $d(n)$ *be defined as above; we denote by* $e(n)$ *the largest integer* e *such that any allowed path of length* $d(n)+e$ *in* γ_n *has at most* $d(n)$ *different vertices. Suppose that, for infinitely many values of* n, *we have simultaneously*

$$\frac{d(n)}{n + d(n) - 1} \leqslant c \quad \text{and} \quad \frac{p(n - e(n))}{n + d(n) - 1} \leqslant c,$$

then, for infinitely many n, $C_U^+(n) \leqslant cn$.

Proof. Let w be any word of length $n + d(n) - 1$. The subwords of length k of w define a path $u_a \dots u_{a+k-1}$, ..., $u_{a+n-k+d(n)-1} \dots u_{a+n+d(n)-2}$, of length $d(n) + n - k$ in Γ_k, so of course $p_w(n) \leqslant d(n)$ for $k \geqslant n$, and this is also true for $n - e(n) \leqslant k < n$ by definition of $e(n)$. Hence $p_w(k) \leqslant d(n)$ for any $k \geqslant n - e(n)$. □

3. Sturmian sequences

3.1. Description of the Rauzy graphs

A **Sturmian** sequence is a sequence U such that $p_U(n) = n + 1$, see [3] for a recent survey about them. A Sturmian sequence is not ultimately periodic (other-

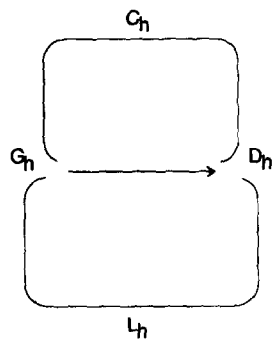

Fig. 1. Sturmian graph.

wise $p_U(n)$ would be bounded) and is uniformly recurrent ([15], and hence recurrent (this can also be deduced from Lemma 2). The Rauzy graphs for Sturmian sequences and their evolution with n are described in [2], and more precisely in [4, 7].

For a Sturmian sequence, the graph Γ_n contains one vertex D_n which is **right special**: it has two outgoing edges D_n0 and D_n1; and one vertex G_n is **left special**: it has two incoming edges $0G_n$ and $1G_n$; D_n and G_n may be the same vertex; every vertex except G_n has one incoming edge, every vertex except D_n has one outgoing edge.

We say that the **central branch** contains the vertex G_n, its successors as far as D_n (included) and (if they exist) the edges between them. The other vertices and edges form two branches, beginning with one of the two outgoing edges of D_n, ending with one of the two incoming edges of G_n, and containing the edges and (if they exist) vertices between them. The lengths of these last two branches are always different, and we call them, respectively, **short branch** and **long branch**. The **short circuit** C_n (resp. **long circuit** L_n) begins with G_n and is made with the central branch followed by the short (resp. long) branch (Fig. 1).

The vertices of Γ_{n+1} are the edges of Γ_n; if $D_n \neq G_n$, the vertices of a branch of Γ_{n+1} are the edges of the same branch of Γ_n (there is a **split** of an edge). If $G_n = D_n$, the central branch of Γ_n is reduced to one vertex, and there is a **burst**: for a **reversing** burst, the vertices of the central branch of Γ_{n+1} are the edges of the long branch of Γ_n, while for a **non-reversing** burst the vertices of the central branch of Γ_{n+1} are the edges of the short branch of Γ_n; in both cases, the short branch of Γ_{n+1} is reduced to one edge.

We denote by γ_n the infinite path going through $u_0 \dots u_{n-1}$, $u_1 \dots u_n$, ... It is made with a succession of short and long circuits (except that the first circuit may be truncated at the beginning). If for n there is a split, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by C_n, L_{n+1} by L_n; if there is a non-reversing burst, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by C_n, L_{n+1} by C_nL_n ; if there is a reversing burst, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by L_n, L_{n+1} by L_nC_n.

From this description, we can re-prove that Lemma 2 applies to Sturmian sequences; for the required path in Γ_m, we can either start from the first vertex of the long branch and follow the long then the short circuit, or start from the first vertex of the short branch and follow the short then the long circuit; this proof appears in [6]: the Sturmian sequences have the property of **grouped factors**, all their factors occur successively from some place in the sequence.

Every Sturmian sequence is associated to an irrational rotation of the torus $\mathbb{T}_1$, $Tx = x + \alpha \bmod 1$ in the following way: if $P_0 = [0, 1-\alpha[$ and $P_1 = [1-\alpha, 1[$, we associate to each point x the sequence $PN(x)$ defined by $PN(x)_n = i$ if $T^n x \in P_i$; and if U is Sturmian, $U = PN(x)$ for some (unique) irrational α and some x ([15]). In all what follows, $0 < \alpha < 1$ is the irrational number defined by U: let $\alpha = [0; a_1, \ldots, a_n, \ldots]$ be its simple continued fraction expansion, and $q_{n+1} = a_{n+1}q_n + q_{n-1}$, $p_{n+1} = a_{n+1}p_n + p_{n-1}$, $p_{-1} = 1$, $p_0 = 0$, $q_{-1} = 0$, $q_0 = 1$. Then, the full description of the Rauzy graphs can be found in [4] or [7]: the reversing bursts take place for $n = q_p + q_{p-1} - 2$, $p \in \mathbb{N}$; for $n = q_p + q_{p-1} - 2 + s$, $1 \leqslant s \leqslant q_p - 1$, the central branch has length $q_p - s + 1$, the long branch has length $q_{p-1} + s - 1$, the short branch has length $s - 1$. The non-reversing bursts take place, if $a_{p+1} > 1$, for $2 \leqslant k \leqslant a_{p+1}$, $n = kq_p + q_{p-1} - 2$; for $n = kq_p + q_{p-1} - 2 + s$, $1 \leqslant s \leqslant q_p - 1$, the central branch has length $q_p - s + 1$, the long branch has length $(k-1)q_p + q_{p-1} + s - 1$, the short branch has length $s - 1$. If $n = kq_p + q_{p-1} - 2 + s$, $1 \leqslant s \leqslant q_p - 1$, $1 \leqslant k \leqslant a_{p+1}$, the short circuits appear in the infinite path γ_n in groups of $a_{p+1} - k$ or $a_{p+1} - k + 1$ and the long circuits are isolated in γ_n if $k < a_{p+1}$, the short circuits are isolated in γ_n if $k = a_{p+1}$.

Also, whenever $x = 0$, we have $G_n = u_0 \ldots u_{n-1}$ for all n.

3.2. *Finite-word complexities*

Proposition 4. *If U is Sturmian,*

$$C_U^+(n) = \left\lceil \frac{n}{2} \right\rceil + 1, \qquad K_U^+(n) = \left\lfloor \frac{n^2}{4} + n \right\rfloor.$$

Proof. We apply Lemma 1 to get the upper bounds, and Proposition 3 to get the lower bounds; as $L(n) = n/2$, the bounds coincide and we get exactly the above expressions. □

Proposition 5. *If U is a Sturmian sequence associated to an irrational α which has an infinite number of partial quotients at least equal to G*

$$\liminf_{n\to+\infty} \frac{C_U^-(n)}{n} \leqslant \frac{2}{G-2}, \qquad \liminf_{n\to+\infty} \frac{K_U^-(n)}{n^2} \leqslant \frac{2}{G-2},$$

$$\limsup_{n\to+\infty} \frac{C_U^-(n)}{n} \geqslant \frac{G}{2+2G}, \qquad \limsup_{n\to+\infty} \frac{K_U^-(n)}{n^2} \geqslant \frac{G}{4+4G}.$$

In particular, if α has unbounded partial quotients,

$$\liminf_{n\to+\infty} \frac{C_U^-(n)}{n} = \liminf_{n\to+\infty} \frac{K_U^-(n)}{n^2} = 0,$$

$$\limsup_{n\to+\infty} \frac{C_U^-(n)}{n} = \frac{1}{2}, \qquad \limsup_{n\to+\infty} \frac{K_U^-(n)}{n^2} = \frac{1}{4}.$$

For any strictly increasing function $f(n)$, we can find a Sturmian sequence U for which $C_U^-(n) \leqslant f(n)$ and $K_U^-(n) \leqslant nf(n)$ for infinitely many n.

Proof. We choose a p such that $a_{p+1} = G$, let $k = [G/2]$, choose $m = q_p + q_{p-1} - 1$, $n = km$, and a word $w = w_1 ... w_n$ such that the path $w_1 ... w_n$, ..., $w_n ... w_{2n-1}$ in Γ_n is included in $k+1$ consecutive short circuits; because there are only non-reversing bursts between m and n, the path $w_1 ... w_l$, ..., $w_n ... w_{n+l-1}$ in Γ_l is included in $k+1$ consecutive short circuits for any $m \leqslant l \leqslant n$, and hence the number of different vertices on this path is at most q_p. The number of different subwords of length l in w is the number of different vertices on the path $w_1 ... w_l$, ..., $w_{n-l+1} ... w_n$ in Γ_l; hence it is at most q_p for $m \leqslant l \leqslant n$; for $l < m$, $p_w(l) \leqslant p_U(l) = l + 1 \leqslant m$. Hence $C(w) \leqslant n/k$ and $K(w) \leqslant n^2/k$, and this situation happens for infinitely many n.

For the same value of m, let $n = 2m$. Any path of length $m+1$ in Γ_m is included in one, two, or three consecutive circuits, and, in the worst case, contains q_p different vertices; hence, for any word w of length n, $p_w(n/2) \geqslant q_p$; but

$$\frac{q_p}{q_p + q_{p-1}} \geqslant \frac{a_p}{1 + a_p}.$$

Hence the upper limit for C^-, taking Proposition 4 into account. For K^-, we apply Proposition 2 to the function $p_w(k)$, the case giving the smallest $K(w)$ being the case where $p_w(k)$ increases or decreases only by one at a time.

And, by choosing a_{p+1} much larger than q_p in the first part of the proof, we prove the last assertion of the proposition. □

Proposition 6. *If U is a Sturmian sequence associated to an irrational α with partial quotients not greater than M,*

$$\frac{n}{2M+2} \leqslant C_U^-(n) \leqslant \left[\frac{n}{2} + 1\right] - \frac{n}{32M+34},$$

$$\frac{n^2}{(2M+2)^2} \leqslant K_U^-(n) \leqslant \left[\frac{n^2}{4} + n\right] - \frac{n^2}{(32M+34)^2}.$$

Proof. Let $m = [n/2]$. We check that if $q_p + q_{p-1} - 1 \leqslant m \leqslant q_{p+1} - 2$, there are always at least q_p different vertices in any path of length $m+1$ in Γ_m; for $q_{p+1} - 1 \leqslant m \leqslant q_p + q_{p+1} - 2$, the short circuit is isolated, and there are always at least $q_{p+1} - q_p$ different vertices in any path of length $m+1$ in Γ_m; the worst possibility for $C_U^-(n)/n$ happens for $n/2 = Mq_p + q_{p-1}$, hence our lower bound; the one for K_U is computed as in the end of the previous proof.

For the upper bound, we shall have to look at different cases.

We begin by

$$q_p + \frac{q_{p-1}}{4} \leqslant \frac{n}{2} \leqslant q_p + \frac{3q_{p-1}}{4},$$

that is the case when $n/2$ falls before a reversing burst, but after the previous burst. We take w such that the path it defines in $\Gamma_{q_p+q_{p-1}-2}$ begins by the last $(q_{p-1})/16$ words of the short circuit C (the length of C being q_{p-1}), the two following circuits being long circuits L (this is possible as the following burst is reversing). If

$$\frac{n}{2} - \frac{q_{p-1}}{8} \leqslant n \leqslant \frac{n}{2} + \frac{q_{p-1}}{8},$$

then the central branch has at most $(7q_{p-1})/8$ vertices, and the first $(q_{p-1})/16$ vertices of the short branch are not in the path defined by w in Γ_m. Hence $p_w(m) \leqslant m + 1 - (q_{p-1})/16$.

We look then to what happens when $n/2$ is close to a reversing burst: suppose that

$$q_p + \frac{3q_{p-1}}{4} \leqslant \frac{n}{2} \leqslant q_{p-1} + \frac{5q_p}{4}.$$

We take w such that the path it defines in $\Gamma_{2q_p+q_{p-1}-2}$ begins by one full short circuit C, the following circuit being a long circuit L (this is always possible); then, if $q_p+q_{p-1}-1 \leqslant m \leqslant 2q_p+q_{p-1}-2$, the path defined by w in Γ_m begins by a full C followed by L, and sees twice every vertex of the central branch before having seen all the vertices; hence, if $q_p+q_{p-1}-1 \leqslant m \leqslant 3q_p/2+q_{p-1}-1$, as the central branch has at least $q_p/2$ vertices, we have $p_w(m) \leqslant m+1-q_p/2$. For $q_p+(q_{p-1})/2-1 \leqslant m \leqslant q_p+q_{p-1}-2$, let c and l be the circuits; we have $L = lc$, $C = l$; the path defined by w in Γ_m begins by llc, and, as m is smaller than twice the length of l we shall not see the vertices of the short branch; the length of c is q_{p-1}, the number of vertices of the central branch is at most $(q_{p-1})/2$, so the number of vertices of the short branch is at least $(q_{p-1})/2$. So we have, in every case,

$$p_w(m) \leqslant m + 1 - \frac{q_{p-1}}{2} \text{ for } \frac{n}{2} - \frac{q_{p-1}}{4} \leqslant m \leqslant \frac{n}{2} + \frac{q_{p-1}}{4}.$$

We suppose now that $a_{p+1} \geqslant 2$ and $q_{p-1}+5q_p/4 \leqslant m \leqslant q_{p+1}-q_p/4$ ($n/2$ falls between a reversing burst and the last non-reversing burst after it). We take w such that the path it defines in $\Gamma_{q_{p+1}-2}$ begins by one full short circuit C, the following circuit being a short circuit C (this is possible by hypothesis); the path defined by w in each Γ_m, $q_p+q_{p-1}-1 \leqslant m \leqslant q_{p+1}-2$, has the same properties. Hence, for each of these values of m, we repeat q_p vertices (q_p being the length of C) in the path defined by w in Γ_m, and $p_w(n) \leqslant m+1-q_p$ for

$$\frac{n}{2} - \frac{q_p}{4} \leqslant m \leqslant \frac{n}{2} + \frac{q_p}{4}.$$

We suppose now that

$$a_{p+1} \geqslant 2 \text{ and } q_{p+1} - \frac{q_p}{4} \leqslant m \leqslant q_{p+1} + \frac{q_p}{4}$$

($n/2$ falls near a non-reversing burst which is followed by a reversing burst). We take w such that the path it defines in $\Gamma_{q_p+q_{p+1}-2}$ begins by one full short circuit C, the following circuit being a long circuit L (this is always possible); then, if $q_{p+1} - 1 \leqslant m \leqslant q_p + q_{p+1} - 2$, the path defined by w in Γ_m begins by a full C followed by L, and sees twice every vertex of the central branch before having seen all the vertices; hence, if $q_{p+1} - 1 \leqslant m \leqslant q_p/2 + q_{p+1} - 1$, as the central branch has at least $q_p/2$ vertices, we have $p_w(m) \leqslant m + 1 - q_p/2$. For $q_{p+1} - q_p/2 - 1 \leqslant m \leqslant q_{p+1} - 2$, let c and l be the circuits; we have $L = cl$, $C = c$; the path defined by w in Γ_m begins by ccl, and repeats always q_p vertices. So we have, in every case, $p_w(m) \leqslant m + 1 - q_p/2$ for

$$\frac{n}{2} - \frac{q_p}{4} \leqslant m \leqslant \frac{n}{2} + \frac{q_p}{4}.$$

As these four cases exhaust all possibilities, we get the claimed estimates (which of course are not the best possible), first for C^-, then, by the usual method, for K^-. □

3.3. Beyond Sturmians

Lemma 5. *If U is not ultimately periodic and not Sturmian, then, for infinitely many values of n, $d(n) \geqslant n + 2$.*

Proof. Suppose first U is not ultimately periodic and recurrent. Suppose the conclusion of the lemma is not satisfied: then $d(n) = n + 1$ *for all n large enough.* We say that a path γ, without repetition, made with the vertices $S_1, \ldots, S_t$ in Γ_h is **extensible** if there exist two edges A and B such that $A \neq B$ and a path containing $A\gamma B$ is allowed. A path in γ_{h+1} following A, the edges of γ, and B is called a non-trivial extension of γ; all its edges from A to B are different (see proof of Lemma 2).

We choose an n large enough, and a path without repetition γ_n with $n + 1$ vertices in Γ_n, denoted by $S_1, \ldots, S_{n+1}$, and extensible (if this is not possible, then, by the reasoning of Lemma 2, U is ultimately periodic); let A_i be the edge S_iS_{i+1}, and let A_0 and A_{n+1} be edges extending it; as $d(n) = n + 1$, A_{n+1} goes to S_g and A_0 starts from S_d (possibly $d = g$). Let γ_{n+1} be the (allowed) path $A_0, \ldots, A_{n+1}$ in Γ_{n+1}. Let B_{-1} be an edge of Γ_{n+1} such that a path containing $B_{-1}\gamma_{n+1}$ is allowed, and suppose we have followed the edge B_{-1} and the path γ_{n+1} in Γ_{n+1}, we call B_{n+1} the next edge we see, going from A_{n+1} to A_{n+2}.

If $B_{n+1} = B_{-1}$ then we have to continue by B_0 and arrive in A_0, starting the path γ_{n+1} again; after that, we have to leave the path γ_{n+1} after a finite number of loops, otherwise U is ultimately periodic. We leave it by some edge B' leaving a vertex A_i, and a path containing $B_{i-1}A_i \ldots A_iB'$ is allowed (B_i being the edge of γ_{n+1} from A_i to A_{i+1}). If this edge B' goes to a vertex A which is an A_i, B' goes from A_i to A_j; but then A_i and A_j are adjacent vertices in Γ_{n+1}, hence adjacent edges in Γ_n, hence $S_{i+1} = S_j$, which is impossible, except if $i = n + 1$. If B' goes to a vertex A which is not an A_i, then the path $A_{i+1}, \ldots, A_i, A$ is allowed, and $d(n+1) > n+2$. Hence the only way to leave γ_{n+1} is from A_{n+1}; as we have to leave it, we can suppose (possibly after having followed the loop several times) that $B_{n+1} \neq B_{-1}$.

Now, B_{n+1} has to go to some A_j (otherwise $d(n+1) > n+2$); this A_j has the same n-prefix as A_g, hence A_{n+2} can only be A_g, or A_0 if $g = d$. And, possibly after having forgotten a number of loops, the path we follow in γ_{n+1} is extensible, and the path we follow in γ_{n+2} is a non-trivial extension of this path.

So we apply the same reasoning to this path in γ_{n+2}, to show that the next edge A_{n+3} we see (in γ_n) can be only A_{g+1}, or A_1 if $d = g$ and $A_{n+2} = A_0$, or A_0 if $d = g + 1$. And, possibly after having forgotten a number of loops, the path we follow in γ_{n+2} is extensible, and the path we follow in γ_{n+3} is a non-trivial extension of this path. By iterating again the argument $n + 1$ times, we see that any infinite path allowed in Γ_n can only go through the vertices $S_1, \ldots, S_{n+1}$, hence, as U is recurrent, there are only $n+1$ words of length n. Hence, because U is non-ultimately periodic, $p_U(m) = m + 1$ for every $m \leqslant n$ (see [14] or Section 2); as n is arbitrarily large, we get that U is Sturmian, which proves our lemma in the recurrent case, because of Lemma 2.

If U is not ultimately periodic but not recurrent, the same reasoning proves that, if $d'(n) = n + 1$ for n large enough, there are exactly $n + 1$ recurrent words of length n for all n. If the conclusion of the lemma is not satisfied, then $d(n + 1) = n + 1$ ultmately, hence $d'(n) \leqslant n+1$ ultimately, hence $d'(n) = n+1$ ultimately (otherwise, U is ultimately periodic, see proof of Lemma 2); hence, either U is Sturmian, or there exists a word w which appears only a finite number of time (and then, so do its right extensions, hence we can choose w with any prescribed large enough length). But then, the graph Γ'_n of words occurring infinitely often is a Sturmian graph (sequences with that property are studied in [18], and are called ★-*Sturmian*); we take it just after a burst such that the next burst is reversing; there is a vertex w in $\Gamma_n \setminus \Gamma'_n$, and an edge from w to some vertex w' in Γ'_n, such that an infinite path γ in Γ_n is allowed, beginning by w and w' and staying in Γ'_n after that; then the first $n + 2$ vertices on that path have to be different, except maybe if w' is on the short branch; but in that case, let m correspond to the graph just after the last reversing burst, the path γ' defined by γ in Γ_m will have its $m + 2$ first vertices different, which is a contradiction. □

Proposition 7. *If U is not ultimately periodic and not Sturmian, then, for infinitely many values of n,*

$$C_U^+(n) > \left[\frac{n}{2}\right] + 1, \qquad K_U^+(n) > \left[\frac{n^2}{4} + n\right].$$

Proof. We choose an m such that Γ_{m-1} contains a path without repetition with $m + 1$ vertices, and $n = 2m - 1$; we can find a word w of length n with $p_w(m-1) = m+1 = [n/2]+2$. Hence our proposition on C_U^+, the one on K_u^+ following by the usual method. □

As a corollary of Propositions 3 and 7, we get that if $C_U^+(n) = [n/2] + 1$, then U is Sturmian; we can then look at $C_U^-(n)$ to know to which kind of irrational number it can be associated. Also, we see that for any sequence U, whenever $C_U^+(n) = [n/2] + 1$

for at least one value of n, then $C_U^+(n) = [n/2] + 1$ for infinitely many values of n. An equivalent property is satisfied by the K-complexities.

4. Sequences with $p_U(n+1) - p_U(n) = 2$ ultimately

If $p_U(n+1) - p_U(n) = 2$ for all n large enough; then Γ_n contains either one right special factor with three outgoing edges or two right special factors with two outgoing edges, and either one left special factor with three incoming edges or two left special factors with two incoming edges.

In this section, we look at particular examples of such sequences for which the combinatorics are at least partially known, though not as thoroughly as for the Sturmian sequences; we give some rather precise bounds for C^+, but unfortunately we have not been able to give comparable estimates for K^+; so most of the times we omit the estimates for K^+, though of course Lemmas 1 and 3 may be applied.

4.1. Arnoux–Rauzy sequences

A sequence such that $p_U(n) = 2n + 1$ for all n (hence it is on three letters), such that it is uniformly recurrent and, for all n, there is one right special factor with three outgoing edges and one left special factor with three incoming edges, is called an **Arnoux–Rauzy sequence** [2, 8].

The graphs Γ_n and their evolution are then completely known, generalizing the Sturmian graphs: we denoted by D_n the right special factor, by G_n the left special factor; the central branch goes from G_n to D_n, and then the short, long and middle branch go from D_n to G_n (for the first values of n the long and middle branch may have the same length and have to be defined arbitrarily); the short, middle and long circuits are denoted by C_n, M_n, L_n (Fig. 2). There is a split if $D_n \neq G_n$, a burst if $D_n = G_n$; for a **reversing** burst, denoted by B_L, the vertices of the central branch of Γ_{n+1} are the edges of the long branch of Γ_n, for a **short** burst, denoted by B_C, the vertices of the central branch of Γ_{n+1} are the edges of the short branch of Γ_n, for a **middle** burst, denoted by B_M. the vertices of the central branch of Γ_{n+1} are the edges of the middle branch of Γ_n; when there is a burst, the short branch of Γ_{n+1} is reduced to one edge.

We denote by γ_n the infinite path going through $u_0 \dots u_{n-1}$, $u_1 \dots u_n, \dots$. It is made with a succession of short, middle and long circuits (except that the first circuit may be truncated at the beginning). If for n there is a split, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by C_n, L_{n+1} by L_n; if there is a short burst, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by C_n, M_{n+1} by C_nM_n, L_{n+1} by C_nL_n; if there is a middle burst, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by M_n, M_{n+1} by M_nC_n, L_{n+1} by M_nL_n; if there is a reversing burst, γ_n is deduced from γ_{n+1} by replacing C_{n+1} by L_n, M_{n+1} by L_nC_n, L_{n+1} by L_nM_n.

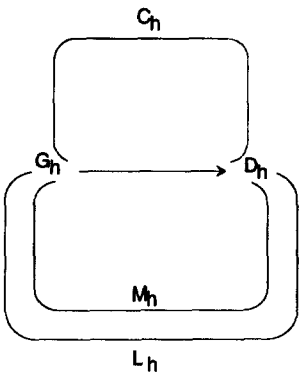

Fig. 2. Arnoux–Rauzy graph.

As a split reduces the length of the central branch, there are infinitely many bursts. It is noticed in [8] that the language of an Arnoux–Rauzy sequence is completely defined (up to a permutation of two letters) by its infinite sequence of bursts, which we consider as an infinite sequence on the symbols B_C, B_M, B_L.

Proposition 8. *If U is an Arnoux–Rauzy sequence, for infinitely many values of n*

$$C_U^+(n) = \left\lceil \frac{2n}{3} \right\rceil + 1$$

if in the sequence of bursts of U the strings of B_C have length at most M' (if $M' = 0$, this means there are no short bursts), then for all n

$$C_U^+(n) \geqslant \left(\frac{1}{2} + \frac{1}{8M' + 20} \right) n - 1$$

if in the sequence of bursts of U infinitely many strings of B_C have length at least G,

$$\liminf_{n \to +\infty} \frac{C_U^+(n)}{n} \leqslant \frac{1}{2} + \frac{15}{32G + 18}.$$

Proof. The sequence of bursts cannot be ultimately equal to B_C, as this would contradict uniform recurrence. So, for infinitely many values of p, there is a burst for $p-1$ and the next burst is reversing or middle; then the short branch has no vertex and the paths L_pM_p and M_pL_p are allowed, hence, by starting at the beginning of the long (resp. middle) branch and following L_pM_p (resp. M_pL_p), we get an allowed path with $2p+1$ vertices without repetition, and we can take $n = 3p$ and a word w corresponding to this path, hence our first formula.

For the second formula, we fix an n, and estimate $d(n)$, the maximal number of vertices of an allowed path without repetition in Γ_n; let c, m, l be the length of the short, middle, long circuit, and s the length of the central branch (the length of a

branch being its number of **vertices**); we have $c+m+l-2s = 2n+1$, $1 \leqslant s \leqslant c \leqslant m \leqslant l$, and it is easy to show by recursion that $l \leqslant m + c$ [8].

We suppose that there are at most M' consecutive short bursts. Suppose first n is such that the next burst is a B_M or B_L; then the path $M_n L_n$ is allowed: then the longest path without repetition has length $d(n) = l + m - s$; hence

$$\frac{d(n)}{n+d(n)} = \frac{2l+2m-2s}{3l+3m+c-4s} \geqslant \frac{2}{3+c/(l+m)} > \frac{4}{7}.$$

If the next burst is a B_C, then we have $d(n) = l + c - r$, and

$$\frac{d(n)}{n+d(n)} = \frac{2l+2c-2s}{3l+m+3c-4s} \geqslant \frac{2}{3+m/(l+c)}.$$

The bursts before the considered B_C are a B_L or B_M followed by $0 \leqslant a \leqslant M'-1$ bursts B_C; hence $l = ac + l_0$, $m = ac + m_0$, and either $c = m_1$, $m_0 = m_1 + c_1$, $l_0 = m_1 + l_1$, or $c = l_1$, $m_0 = l_1 + c_1$, $l_0 = m_1 + l_1$, where l_1, m_1, c_1 are the lengths of the short, middle, long circuit before this group of bursts; they satisfy $c_1 \leqslant m_1 \leqslant l_1 \leqslant 2m_1$. Hence

$$\frac{d(n)}{n+d(n)} \geqslant \frac{1}{2} + \frac{c}{(8a+6)c + 6l_0 + 2m_0} \geqslant \frac{1}{2} + \frac{1}{8M'+20}.$$

We apply then Lemma 3 to get a bound on C^+.

We look now at an upper bound for C^+, at least, in view of the first assertion, for infinitely many n. We suppose $G \geqslant 3$; let $b = [G/3] + 1$, $a = G - b - 1$. We choose an n such that $s = c$ (this means that we are just after a burst), the next bursts after n are at least $b+1$ bursts B_C, the last bursts before n are one B_M or B_L followed by a bursts B_C. Then $d(n) = l$ and $e(n) \geqslant bc$ (see Lemma 4 for the definition of $e(n)$). Then

$$\frac{d(n)}{n+d(n)-1} \vee \frac{p(n-e(n))}{n+d(n)-1} \leqslant \frac{2l}{3l+m-c} \vee \left(1 + \frac{2l+2m-(4b+2)c}{3l+m-c}\right).$$

As c, l, m have the same expressions as in the last paragraph, we have $m/c = a + m_0/c_0 \leqslant a+2 < 2b+1$, so $l + m - (2b+1)c < l$ and

$$\frac{d(n)}{n+d(n)-1} \vee \frac{p(n-e(n))}{n+d(n)-1} \leqslant \frac{2l}{3l+m-c} = \frac{2ac+2l_0}{(4a-1)c+m_0+3l_0} \leqslant \frac{1}{2} + \frac{1+2c/l_0}{14+(16a-4)c/l_0},$$

this last quantity is decreasing in c/l_0 if $a \geqslant 2$, and its smallest possible value (because of the properties of c_1, m_1, l_1) is at least $\frac{1}{3}$, which gives

$$C^+(n) \leqslant \left(\frac{1}{2} + \frac{5}{16a+38}\right) n$$

by using Lemma 4 (this bound holds also for $a = 1$, where the worst case is reached for $c/l_0 = 1$). The fact that $a \geqslant 2G/3 - 2$, and that this situation happens for infinitely

many n gives the required bound for the lower limit, which holds also trivially for $G = 0, 1, 2$. □

So, Arnoux–Rauzy sequences with unbounded strings of short bursts satisfy

$$\liminf_{n\to+\infty} \frac{C_U^+(n)}{n} = \frac{1}{2}.$$

At the opposite end, when $M' = 0$ (this is the case, for example for the Tribonacci sequence [9]), we can show that $C_U^+(n) \geqslant 3n/5$ for all n and $C_U^+(n) \leqslant 7n/11$ for infinitely many n; all these bounds can probably be improved.

For Arnoux–Rauzy sequences, the situation for the lower complexity C^- is the same as for Sturmian sequences, so we do not study it in details; C^- will oscillate between $o(n)$ and $2n/3$ if the strings of B_C have length unbounded, and between $k_1(M')/n$ and $2n/3 - k_2(M')/n$ if the strings of B_C have length bounded by M'.

4.2. Rotation sequences

Rotation sequences are defined, for an irrational rotation of the torus $\mathbb{T}_1$, $Tx = x + \alpha \bmod 1$, some $0 < \beta < 1$ such that $\beta \notin \mathbb{Z}\alpha + \mathbb{Z}$ and some $x \in \mathbb{T}_1$, in the following way: if P_0 and P_1 are the two semi-open (on the right) intervals delimited on $\mathbb{T}_1$ by the points 0 and β, $u_n = i$ if $T^n x \in P_i$; they are uniformly recurrent, and they satisfy $p_U(n) = 2n$ for n large enough, with two right special factors with two outgoing edges and two left special factor with two incoming edges [15].

We give here a short description of the Rauzy graphs for rotation sequences (they have been studied independently in [12]); for given h, a vertex of Γ_h, denoted by $w_0 \dots w_{h-1}$, corresponds to the set $\bigcap_{i=0}^{h-1} T^{-i} P_{w_i}$; the factors of length h of U correspond to the semi-open intervals of $\mathbb{T}_1$ delimited by the points $k\alpha$, $-h+1 \leqslant k \leqslant 0$, and $l\alpha + \beta$, $-h+1 \leqslant l \leqslant 0$. We denote by $J(x)$ the word corresponding to the interval whose left endpoint is x. Then there is an edge from $J(k\alpha)$ to $J((k+1)\alpha)$ and from $J(l\alpha+\beta)$ to $J((l+1)\alpha+\beta)$, $-h+1 \leqslant k \leqslant -1$, $-h+1 \leqslant k \leqslant -1$; we need four more edges to complete the graph. Because of the uniform recurrence of the sequence, there are only three possible cases for these extra edges:

Case I: $J(0) \to J(i\alpha)$, then $J(\beta) \to J(i\alpha+\beta)$ by symmetry, while $J(j\alpha) \to J((1-h)\alpha+\beta)$ and $J(k\alpha+\beta) \to J((1-h)\alpha)$.

Case I′: $J(0) \to J(k\alpha+\beta))$, $J(\beta) \to J(j\alpha)$, $J(i\alpha) \to J((1-h)\alpha)$ and $J(i\alpha+\beta) \to J((1-h)\alpha+\beta)$ by symmetry.

Case II: $J(0) \to J(i\alpha+\beta))$, $J(\beta) \to J(j\alpha)$, $J((1-h-i)\alpha) \to J((1-h)\alpha+\beta)$ and $J((1-h-j)\alpha+\beta) \to J((1-h)\alpha)$.

Proposition 9. *For rotation sequences, for every n large enough,*

$$C_U^+(n) \geqslant \left\lceil \frac{3n}{5} \right\rceil$$

for infinitely many values of n

$$C_U^+(n) = \left[\frac{2n}{3}\right]', \qquad K_U^+(n) = \left[\frac{n^2}{3} + \frac{2n}{3}\right]'.$$

Proof. For any h, suppose for example we are in Case I; in Γ_h, we start from $J((k+1)\alpha+\beta)$, go through $J(\beta)$ and $J(i\alpha+\beta)$ to $J(k\alpha+\beta)$ ($i \leqslant k$, otherwise the sequence would not be uniformly recurrent); then, again by uniform recurrence, it has to be allowed to continue to $J((1-h)\alpha)$ and hence to $J(j\alpha)$, for which we can go either as far as $J(0)$ or as far as $J((-i-1)\alpha+\beta)$; the above path is allowed, is without repetition, and, as either $i \leqslant 1-h/2$ or $j \geqslant i \geqslant -h/2$, we get that $d(h) \geqslant [3h/2]$, and our first assertion by Lemma 3. A similar reasoning works in the symmetric Case I′.

If we are in Case II, we start from $J((1-h)\alpha+\beta)$, go to $J((1-h-j)\alpha+\beta)$, then, either through $J((1-h)\alpha)$ or through $J(\beta)$, to $J(j\alpha)$, $J((1-h-i)\alpha)$, and either $J(0)$ or just before $J((1-h)\alpha+\beta)$; the fact that $p(h+1) = 2h+2$, h being large enough, and the symmetry of the graph, imply that at least three of these paths are allowed, hence we can find an allowed path without repetition with $[3h/2]$ vertices, and we have proved our first assertion.

We check that for $h = q_n - 1$ (see Section 3), if we are in Case I then $i = 1-h$, $j = k = 0$ and the above path has length $2h$; if we are in Case I′ $i = 0$, $j = k = 1-h$ and a similar path has length $2h$; if we are in Case II, then there are two central (going from a left special to a right special vertex) branches of the same length, and, after a number of splits (for each split the length of each central branch is reduced by one), there exists $h' \geqslant h$ such that we are still in Case II and $i = 1-h'-i$, and this gives a graph with an allowed path without repetition of length $2h'$. Hence we can conclude in the same way as in the first assertion of Proposition 8. □

Hence, from the point of view of C^+, rotation sequences behave like Arnoux–Rauzy sequences with unbounded strings of short bursts. Let us point out that the rotation sequences, their Rauzy graphs and their evolutions, have still to be studied in depth if we want to improve the above estimates.

5. Sequences of super-linear complexity

5.1. The power sequence

The **power sequence** U is the sequence $010011000111\ldots0^k1^k\ldots$

Proposition 10. *For the power sequence and for all n,*

$$p_U(n) = \frac{n(n+1)}{2} + 1.$$

Proof. $p_U(n+1) - p_U(n)$ is the number of different words of length n which have two different (right) extensions of length $n+1$; now, if a word is of the form $0^a1^b0^c$

(or the symmetric form by exchanging 0 and 1), it has only one possible position in the sequence and hence one extension; there remain the words 0^k1^{n-k} (this word can be followed by 1 always, and by 0 whenever $k \leqslant n-k$), and the words 1^k0^{n-k}, (this word can be followed by 0 always, and by 1 whenever $k < n-k$) ; hence $p_U(n+1) - p_U(n) = n+1$. □

Proposition 11. *For the power sequence and every n,*

$$C_U^-(n) = 1, \qquad K_U^-(n) = n,$$
$$n - 2\sqrt{n} \leqslant C_U^+(n) \leqslant n - 2\sqrt{n} + 4,$$
$$\frac{n^2}{2} - 4n\sqrt{n} \leqslant K_U^+(n) \leqslant \frac{n^2}{2} - n\sqrt{2n}.$$

Proof. The values of C^- and K^- come from the occurrence of the word 0^n for every n. For C_U^+, we observe that the only factors which can be repeated are those of the form 0^i1^j or 1^i0^j. Hence, for any integer k, starting from the beginning of the sequence, the first factor of length $2k$ which will be repeated is 0^k1^k, and this repetition occurs when we see $0^{k+1}1^{k+1}$ in U; the first factor of length $2k+1$ which will be repeated is 1^k0^{k+1}, and this repetition occurs when we see $1^{k+1}0^{k+2}$ in U; the longest path without repetition in Γ_m is the one starting with $u_0 \dots u_{m-1}$, hence we have

$$d(2k) = k^2 + k + 1, \qquad d(2k+1) = k^2 + 2k + 1.$$

Hence $d(n) \geqslant n^2/4 + n/2$, and Lemma 3 gives the lower bound for C^+. Furthermore, we do have $C^+(k^2+3k) = d(2k) = k^2+k+1$, and $C^+(k^2+4k+2) = d(2k+1) = k^2+2k+1$, and for these values our upper bound holds as $n < (k+2)^2$ implies $k \geqslant \sqrt{n} - 2$; they hold a fortiori for other values of n. The upper bound for K^+ comes from Lemma 1; for the lower bound, we choose (for example) an even k such that $k^2 + 3k \leqslant n$ and look at the word $w = u_0 \dots u_{k^2+3k-1}$; we have $p_w(2k+l) \geqslant k^2 + k + 1 - l$, which is enough to get this (rather rough) bound. □

5.2. Complete sequences

A **complete sequence** is a sequence on an alphabet with s letters such that $p_U(n) = s^n$ for all n.

Proposition 12. *If* $p_U(n) = s^n$,

$$C_U^-(n) = 1, \qquad K_U^-(n) = n,$$
$$C_U^+(n) = s^k \vee n - k,$$
$$K_U^+(n) = \frac{(n-k)(n-k+1)}{2} + s^{k+1} - 1$$

for the unique k such that $s^k + k - 1 \leqslant n < s^{k+1} + k$.

Proof. The values of C^- and K^- come from the occurrence of the word 0^n for every n. The values of C^+ and K^+ come for the results in [20]: the Rauzy graph of the sequence is then, for any n, the De Bruijn graph on s^n vertices ([5], see [20] for further references); it is used in [20] (for $s = 2$, but the generaliztion is straightforward) to prove that there exists a word of length n on s letters containing all subwords of lengths 1 to k, and $n - i + 1$ subwords of length $i > k$. But, as is noticed in [6], this word of length n must occur in U, as U contains every possible word on s letters. □

The same method is used in [6] to show that complete sequences have grouped factors.

Conversely, a sequence on s letters which satisfies $C_U^+(n) = s^k \vee n - k$ or $K_U^+(n) = ((n-k)(n-k+1))/2 + s^{k+1} - 1$ for every n is a complete sequence.

References

[1] J.-P. Allouche, Sur la complexité des suites infinies, Bull. Belg. Math. Soc. 1, 2 (1994) 133–143.

[2] P. Arnoux, G. Rauzy, Représentation géométrique de suites de complexité $2n + 1$, Bull. Soc. Math. France 119 (1991) 199–215.

[3] J. Berstel, Recent results in Sturmian words, Developments in language theory (Magedburg 1995), World Scientific, Singapore to appear.

[4] V. Berthé, Fréquences des facteurs des suites sturmiennes, Theoret. Comput. Sci. 165 (1996) 295–309.

[5] N.G. de Bruijn, A combinatorial problem, Nederl. Akad. Wetensch. Proc. 49 (1946) 758–764.

[6] J. Cassaigne, Sequences with grouped factors, preprint.

[7] N. Chekhova, Covering numbers of rotations, Theoret. Comput. Sci., accepted for publication.

[8] N. Chekhova, Nombres de recouvrement, Ph.D. Thesis, Université Aix-Marseille 2, 1997.

[9] N. Chekhova, P. Hubert, A. Messaoudi, Propriétés combinatoires, ergodiques et arithmétiques de la suite de Tribonacci, preprint.

[10] A. Colosimo, A. de Luca, Special factors in biological strings, preprint 97/42, Dipartimento di Matematica Universita' di Roma 'La Sapienza' Proc. of the workshop, Complexity of the living: a modelistic approach, Rome, February 1997, to appear.

[11] A. de Luca, On the combinatorics of finite words, Theoret. Comput. Sci. 218 (this Vol.) (1999) 13–39.

[12] G. Didier, Codages de rotations et fractions continues, J. Number Theoret, to appear.

[13] S. Ferenczi, Complexity of sequences and dynamical systems, Discrete Math., to appear.

[14] G.A. Hedlund, M. Morse, Symbolic dynamics, Amer. J. Math. 60 (1938) 815–866.

[15] G.A. Hedlund, M. Morse, Symbolic dynamics II: Sturmian trajectories, Amer. J. Math. 62 (1940) 1–42.

[16] A. Iványi, On the d-complexity of words, Ann. Univ. Sci. Budapest Sect. Comput. 8 (1987) 69–90.

[17] Z. Kása, On the d-complexity of strings, presented at the 1st Joint Conf. on Modern Applied Mathematics, Ilieni/Illyefalva, Romania, 13–17 June 1995.

[18] I. Nakashima, J.-I. Tamura, S.-I. Yasutomi, Modified complexity and ★-Sturmian words, preprint.

[19] G. Rote, Sequences with subword complexity $2n$, J. Number Theory 46 (1994) 196–213.

[20] J. Shallit, On the maximum number of distinct factors in a binary string, Graphs Comb. 9 (1993) 197–200.