Complexity of infinite sequences with zero entropy

by

Christian MAUDUIT[1] (Marseille) and
Carlos Gustavo MOREIRA[2] (Rio de Janeiro)

Abstract: The complexity function of an infinite word w on a finite alphabet A is the sequence counting, for each nonnegative n, the number of words of lenght n on the alphabet A that are factors of the infinite word w. The goal of this work is to estimate the number of words of lenght n on the alphabet A that are factors of an infinite word w with a complexity function bounded by a given function f with subexponential growth.

2000 Mathematics Subject Classification: 68R15, 37B10.

Keywords: combinatorics on words, symbolic dynamics

1 Introduction and notations

In the whole paper we denote by q a fixed integer greater or equal to 2, by A the finite alphabet $A = \{0, 1, \ldots, q-1\}$, by $A^* = \bigcup_{k \geq 0} A^k$ the set of finite words on the alphabet A and by $A^{\mathbb{N}}$ the set of infinite words (or infinite sequences of letters) on the alphabet A.

For any positive integer n we denote by π_n the projection from $A^{\mathbb{N}}$ to A^n defined by $\pi_n(w) = w_1 w_2 \ldots w_n$ if $w = w_1 w_2 \ldots w_i \ldots$ with $w_i \in A$ for any positive integer i.

If S is a finite set, we denote by $|S|$ the number of elements of S.

If $w \in A^{\mathbb{N}}$ we denote by $L(w)$ the set of finite factors of w:

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

and, for any non negative integer n, we write $L_n(w) = L(w) \cap A^n$.

[1]Research partially supported by the Brazil/France Agreement in Mathematics (Proc. CNPq 60-0014/01-5 and 69-0140/03-7).

[2]Research partially supported by CNPq.

If x is a real number, we denote

$$\lfloor x \rfloor = \sup\{n \in \mathbb{Z}, n \leq x\}$$

and

$$\lceil x \rceil = \inf\{n \in \mathbb{Z}, x \leq n\}.$$

Definition 1.1. The complexity function of $w \in A^{\mathbb{N}}$ is defined for any integer n by $p_w(n) = |L_n(w)|$.

The complexity function gives information about the statistical properties of an infinite sequence of letters. In this sense, it constitutes one possible way to measure the random behaviour of an infinite sequence (see [Que] and [PF], and see [MS1] and [MS2] for connections between measure of normality and other measures of pseudorandomness).

We have obviously $1 \leq p_w(n) \leq q^n$ for any positive integer n and it is easy to check that the sequence $(p_w(n))_{n\in\mathbb{N}}$ is bounded if and only if w is ultimately periodic. A basic result from [CH73] shows that if there exists a positive integer n such that $p_w(n) \leq n$, then the sequence $(p_w(n))_{n\in\mathbb{N}}$ is bounded. It follows from this result that non ultimately periodic sequences w with lowest complexity are such that $p_w(n) = n+1$ for any positive integer n. Such sequences, called sturmian sequences, have been extensively studied since their introduction by G. A. Hedlund and M. Morse in [HM1] and [HM2] (see [Lot, chapter 2] and [PF]).

It is interesting to notice that if w represents the q-adic expansion (resp. the continued fraction expansion) of the irrational number $\rho \in]0,1[$, then the combinatorial property of w to be a sturmian sequence implies the arithmetic property for ρ to be a transcendental number (see [FM] (resp. [ADQZ]) and see [AB2] (resp. [AB1]) for a generalization to the case where w has a sublinear complexity).

It is easy to prove the following lemma:

Lemma 1.2. *For any $w \in A^{\mathbb{N}}$ and for any $(n, n') \in \mathbb{N}^2$ we have $L_{n+n'}(w) \subset L_n(w)L_{n'}(w)$ and so $p_w(n+n') \leq p_w(n)p_w(n')$.* □

Consequence 1: It results from Lemma 1.2 that for any $w \in A^{\mathbb{N}}$, the sequence $\left(\frac{1}{n}\log_q p_w(n)\right)_{n\geq 1}$ converges. We denote $E(w) = \lim\limits_{n\to\infty} \frac{1}{n}\log_q p_w(n)$.

It can be shown (see for exemple [Kůr]) that $E(w)\log q$ is the topological entropy of the symbolic dynamical system $(X(w), T)$ where T is the one-sided shift on $A^{\mathbb{N}}$ and $X = \overline{orb_T(w)}$ is the closure of the orbit of w under the action of T in $A^{\mathbb{N}}$ ($A^{\mathbb{N}}$ is equipped with the product topology of the discrete topology on A, i.e. the topology induced for example by the distance $d(w, w') = exp(-\min\{n \in \mathbb{N} |\, w_n \neq w'_n\}))$.

Consequence 2: Another easy consequence of Lemma 1.2 is that if there exists an integer n_0 such that $p_w(n_0) < q^{n_0}$, then $p_w(n) = o(q^n)$

This simple remark shows that there are necessary conditions to verify for a non decreasing sequence of integers $(p(n))_{n\in\mathbb{N}}$ to be the complexity function of some $w \in A^{\mathbb{N}}$ (see for instance [Fer]). But the characterization of all complexity functions (i.e. necessary and sufficient conditions for a non decreasing sequence of integers $(p(n))_{n\in\mathbb{N}}$ to be the complexity function of some $w \in A^{\mathbb{N}}$) remains an open problem.

Nethertheless, let us mention that J. Cassaigne gave a complete answer to this question in the special case where p is linear ([Cas2]) and that some partial results concerning the case where p is sublinear can be found in [Ale] and [Cas1].

If we weaken the question by asking only which are the possible orders of magnitude for complexity functions, the problem remains still open, but it follows from an unpublished result due to J. Goyon [Goy] that for any $k \geq 1$ and any $(\alpha_1, \alpha_2, \cdots, \alpha_k)$ in $(1, +\infty)$ x $\mathbb{R}^{k-1}$, there exists $w \in A^{\mathbb{N}}$ such that $p_w(n)$ has order of magnitude $n^{\alpha_1}(\log\, n)^{\alpha_2} \cdots (\log \cdots \log\, n)^{\alpha_k}$ (see also [Cas2] for the case $1 < \alpha_1 < 2$).

There are many references concerning the construction of infinite sequences w with low complexity, i.e. such that $p_w(n) = O(n^k)$ for some $k \geq 1$ (see [All] or [Fer] for a survey concerning these constructions). But, as it is pointed in [Cas3], "not many examples are known which have intermediate complexity, i.e. for which $E(w) = 0$ but $\frac{\log p_w(n)}{\log n}$ is unbounded". In [Cas3] J. Cassaigne constructed a large family of infinite sequences with intermediate complexity and proved the following result:

Theorem 1.3. *Let $\tau : \mathbb{R}^+ \to \mathbb{R}^+$ be a function such that :*
(i) $\lim\limits_{t\to+\infty} \frac{\tau(t)}{\log t} = +\infty$,
(ii) τ is differentiable, except possibly at 0,
(iii) $\lim_{t\to+\infty} \tau'(t)t^a = 0$ *for some* $a > 0$,
(iv) τ' is decreasing.

Then there exists $w \in \{0,1\}^{\mathbb{N}}$ such that $\log p_w(n) = \tau(n)$. Moreover w can be taken to be uniformly recurrent.

This construction is rich enough to include examples such that $\tau(n) = n^{\alpha}$ $(0 < \alpha < 1)$, $\tau(n) = (\log(n+1))^{\alpha}$ $(\alpha > 1)$ or $\tau(n) = n^{\alpha+\beta\cos(\log(n+1)^{\gamma})}$ $(\alpha > 0,\ |\beta| < \alpha$ and $\gamma \in \mathbb{R})$.
In the same spirit, the first step of our work consists (sections 3 and 4), for any given function f verifying some reasonable conditions, to construct a huge set of infinite words w such that p_w is close to f (proposition 4.8).

2 Results

Definition 2.1. We say that a function f from $\mathbb{N}$ to $\mathbb{R}^+$ verifies the conditions $(\mathcal{C}_0)$ if

(i) $f(n+1) > f(n) \geq n+1$ for any $n \in \mathbb{N}$,
(ii) $\exists\ n_0 \in \mathbb{N},\ n \geq n_0 \Rightarrow f(2n) \leq f(n)^2$ and $f(n+1) \leq f(1)f(n)$,
(iii) the sequence $\left(\frac{1}{n}\log_q f(n)\right)_{n\geq 1}$ converges to zero.

Examples 2.2. Let us give two typical examples of functions satisfying the conditions $(\mathcal{C}_0)$. In the rest of our paper, we will apply our results to these two examples in order to help the reader to understand them and to get a precise idea about the order of magnitude of our estimates.

Example A: For each $\alpha \geq 1$, the function f is defined by $f(0) = 1$, $f(n) = n + q - 1$ for $1 \leq n < n_0$ and $f(n) = n^{\alpha}$ for $n \geq n_0$, with $n_0 = sup\left(2, \frac{1}{q^{1/\alpha}-1}\right)$.

Example B: For each $0 < \alpha < 1$, the function f is defined by $f(0) = 1$, $f(n) = n+q-1$ for $1 \leq n < n_0$ and $f(n) = q^{n^{\alpha}}$ for $n \geq n_0$, with $n_0 = inf\{n \in \mathbb{N}, q^{(n+1)^{\alpha}} - q^{n^{\alpha}} \geq 1$ and $q^{n^{\alpha}} \geq n + q\}$.

Our work concerns the study of infinite sequences w the complexity function of which is bounded by a given function f verifying the conditions $(\mathcal{C}_0)$.

More precisely, our goal is to estimate the number of words of length n on the alphabet A that are factors of an infinite word with a complexity function less than f. The sturmian case ($f(n) = n+1$) was studied by F. Mignosi in [Mig], who proved an explicit formula conjectured by S. Dulucq and D. Gouyou-Beauchamps in [DG]: the number of words of length n on the alphabet $\{0,1\}$ that are factors of a sturmian infinite word is exactly $1 + \sum_{i=1}^{n}(n-i+1)\Phi(i)$, where Φ is the Euler function (this is asymptotically equivalent to n^3/π^2). This formula can be found also in [KLB], but it seems that the first proof of this formula appears in an earlier paper by E. Lipatov ([Lip]). A geometric proof of it is due to J. Berstel and Pocchiola in [BP] and a combinatorial proof was given by A. de Luca and F. Mignosi in [LM] (see [Lot]). Some partial generalizations concerning the case $f(n) = kn+1$ (for $k \geq 2$) were done by F. Mignosi and L. Zamboni in [MZ]. In the case of positive entropy (i.e. $\lim\limits_{n\to\infty} \frac{1}{n}\log_q f(n) > 0$), some sharp estimates can be obtained by using a different method. This will be the scope of a future work.

In all this paper f is a function from $\mathbb{N}$ to $\mathbb{R}^+$ verifying the conditions $(\mathcal{C}_0)$.

Let us denote $W(f) = \{w \in A^{\mathbb{N}}, p_w(n) \leq f(n), \forall n \in \mathbb{N}\}$ and $\mathcal{L}_n(f) = \bigcup\limits_{w\in W(f)} L_n(w)$.

The aim of sections 3 and 4 is to give upper bounds and lower bounds for $|\mathcal{L}_n(f)|$.

We will exhibit (theorems 3.1 and 4.1) for any given function f satisfying the conditions $(\mathcal{C}_0)$ functions φ and ψ of approximately the same order of magnitude such that for n big enough, we have

$$q^{\psi(n)} \leq |\mathcal{L}_n(f)| \leq q^{\varphi(n)}.$$

In particular, these functions φ and ψ will satisfy

$$\lim_{n\to+\infty} \frac{1}{n}\psi(n) = \lim_{n\to+\infty} \frac{1}{n}\varphi(n) = 0.$$

3 Upper bounds for $|\mathcal{L}_n(f)|$

For any integers k and N we have

$$\begin{aligned} \mathcal{L}_{kN}(f) &= \bigcup_{w\in W(f)} L_{kN}(w) \\ &\subset \bigcup_{w\in W(f)} L_N^k(w) \quad \text{by Lemma 1.2.} \end{aligned}$$

But

$$\begin{aligned}
\bigcup_{w\in W(f)} L_N^k(w) &= \bigcup_{\substack{w\in A^{\mathbb{N}}\\ |L_n(w)|\le f(n),\forall\, n\in\mathbb{N}}} L_N^k(w)\\
&\subset \bigcup_{\substack{w\in A^{\mathbb{N}}\\ |L_N(w)|\le f(N)}} L_N^k(w)\\
&\subset \bigcup_{\substack{S\subset A^N\\ |S|\le f(N)}} S^k = \bigcup_{\substack{S\subset A^N\\ |S|= f(N)}} S^k,
\end{aligned}$$

so that

$$\begin{aligned}
|\mathcal{L}_{kN}(f)| &\le \sum_{\substack{S\subset A^N\\ |S|= f(N)}} f(N)^k = f(N)^k\binom{q^N}{f(N)}\\
&\le f(N)^k q^{Nf(N)}.
\end{aligned}$$

We will now choose the parameter k in order to optimize this majoration.

Let us suppose that $N\ge N_0$, where N_0 satisfies $f(N_0)>q$ and take $k=\lfloor\frac{Nf(N)}{\log_q f(N)}\rfloor$ in order to obtain that $|\mathcal{L}_{kN}(f)|\le q^{2Nf(N)}$. It is easy to verify that if f satisfies $(\mathcal{C}_0)$ then the sequence $\left(\lfloor\frac{Nf(N)}{\log_q f(N)}\rfloor\right)_{N\ge N_0}$ is non decreasing, so the sequence $\left(N\lfloor\frac{Nf(N)}{\log_q f(N)}\rfloor\right)_{N\ge N_0}$ is strictly increasing.

Let $F(N)=N\lfloor\frac{Nf(N)}{\log_q f(N)}\rfloor$ for any integer N, and $F^*(n)=\min\{m\in\mathbb{N}|F(m)\ge n\}$, for any $n\in\mathbb{N}$.

If we still denote by F a (arbitrary) continuous and strictly increasing extension of F from $\mathbb{R}^+$ to $\mathbb{R}^+$, it follows that $F^*(n)\le F^{-1}(n)+1$ for any $n\in\mathbb{N}$.

Given an integer n, let $N=F^*(n)$. We have $F(N-1)<n\le F(N)$.

If follows from the previous estimate that

$$\begin{aligned}
|\mathcal{L}_n(f)| &\le |\mathcal{L}_{F(N)}(f)|\\
&\le q^{2Nf(N)}\\
&= q^{\varphi(n)}
\end{aligned}$$

with

$$\varphi(n)=2F^*(n)f(F^*(n)). \tag{1}$$

As

$$\lim_{N\to\infty} \frac{1}{N} \log_q f(N) = 0,$$

we remark that, for any integer n such that $F^*(n) \geq n_0 + 1$, we have

$$\begin{aligned}
\frac{\varphi(n)}{n} &\leq \frac{\varphi(F(N))}{F(N-1)} \\
&= \frac{2Nf(N)}{F(N-1)} \\
&\leq \frac{2qNf(N-1)}{(N-1)\lfloor \frac{(N-1)f(N-1)}{\log_q f(N-1)} \rfloor} = O(\frac{\log_q f(N-1)}{N-1}) = o(1).
\end{aligned}$$

Finaly, we have proved the following theorem:

Theorem 3.1. $|\mathcal{L}_n(f)| \leq q^{\varphi(n)}$ *where* φ *is defined by (1).*

Examples 3.2.

- For f defined in Example A, we have

$$F(N) = N \lfloor \frac{N^{\alpha+1}}{\alpha \log_q N} \rfloor = \frac{N^{\alpha+2}}{\alpha \log_q N} + O(N),$$

so that

$$F^{-1}(n) = \left(\frac{\alpha}{\alpha+2}\right)^{1/(\alpha+2)} n^{1/(\alpha+2)} (\log_q n)^{1/(\alpha+2)} + O(n^{1/(\alpha+2)})$$

and

$$f(F^*(n)) = \left(\frac{\alpha}{\alpha+2}\right)^{\alpha/(\alpha+2)} n^{\alpha/(\alpha+2)} (\log_q n)^{\alpha/(\alpha+2)} + O(n^{\alpha/(\alpha+2)} (\log_q n)^{(\alpha-1)/(\alpha+2)})$$

(since $F^*(n) = F^{-1}(n) + O(1)$), so that

$$\varphi(n) = 2\left(\frac{\alpha}{\alpha+2}\right)^{(\alpha+1)/(\alpha+2)} n^{(\alpha+1)/(\alpha+2)} (\log_q n)^{(\alpha+1)/(\alpha+2)} + O(n^{(\alpha+1)/(\alpha+2)} (\log_q n)^{\alpha/(\alpha+2)}).$$

- For f defined in Example B, we have

$$F(N) = N \lfloor N^{1-\alpha} q^{N^\alpha} \rfloor = N^{2-\alpha} q^{N^\alpha} + O(N),$$

so that

$$F^{-1}(n) = (\log_q n)^{1/\alpha} - \frac{2-\alpha}{\alpha^2}(\log_q n)^{\frac{1}{\alpha}-1}\log_q\log_q n + O\left((\log_q n)^{\frac{1}{\alpha}-2}(\log_q\log_q n)^2\right),$$

and

$$f(F^*(n)) = n(\log_q n)^{-(2-\alpha)/\alpha} + O(n(\log_q n)^{-2/\alpha}(\log_q\log_q n)^2) + O(n(\log_q n)^{2-3/\alpha}) =$$

$$= (1+o(1))n(\log_q n)^{-(2-\alpha)/\alpha}$$

(since $F^*(n) = F^{-1}(n) + O(1)$), so that

$$\varphi(n) = \frac{2n}{(\log_q n)^{(1-\alpha)/\alpha}} + O\left(\frac{n(\log_q\log_q n)^2}{(\log_q n)^{1/\alpha}}\right) = \frac{(2+o(1))n}{(\log_q n)^{(1-\alpha)/\alpha}}.$$

4 Lower bounds for $|\mathcal{L}_n(f)|$

The main goal of this section is to give lower bounds for $|\mathcal{L}_n(f)|$ when f satisfies the conditions $(\mathcal{C}_0)$. To do this, we will construct, for any fixed $\eta_0 > 0$, a large family W of infinite words w with a complexity function p_w close to f and then minorate $|\bigcup_{w\in W} L_n(w)|$. We will end with the following theorem:

Theorem 4.1. *For any fixed $\eta_0 > 0$ there exists an integer N_0 such that for any $n \geq N_0$ we have*

$$|\mathcal{L}_n(f)| > \exp\left(\left(\frac{1}{8} - \eta_0\right)\frac{n}{G^{-1}(4n)}\log\frac{n}{G^{-1}(4n)}\right),$$

where $G(x) = 2xg(x)$ and g is a function verifying the conditions $(\mathcal{C}_0)$ such that for any integer $n \geq N_0$

$$\min(G((2+\eta_0)n\log^2 n), G((2+\eta_0)n)^2) \leq f(n).$$

4.1 Construction of a large family W of infinite words

Let $(a_k)_{k\geq 1}$ be the sequence of integers defined by $a_1 = 1$, $a_2 = 3$ and for $k \geq 2$

$$a_{k+1} = \left\lceil \left(2 + \frac{1}{\log^2 a_k}\right) a_k \right\rceil$$

and $(b_k)_{k\geq 1}$ be the sequence of integers defined by $b_k = a_{k+1} - 2a_k$.

Lemma 4.2. *For any $k \geq 3$ we have $2^k < a_k < 2^{k+1}$.*

Proof. An easy computation shows that $a_1 = 1$, $a_2 = 3$, $a_3 = 9$, $a_4 = 20$, $a_5 = 43$ and $a_6 = 90$.

As we have $a_{k+1} \geq 2a_k$ for any $k \geq 1$, it follows that $a_k > 2^k$ for any $k \geq 3$.

For the majoration, we can proceed as follow:

For any $k \geq 3$ we have $a_{k+1} < 2a_k + \frac{a_k}{k^2 \log^2 2} + 1$ so that for any $k \geq 3$

$$\frac{a_{k+1}}{a_k} < 2 + \frac{1}{k^2 \log^2 2} + \frac{1}{2^k}.$$

It follows that for any $k \geq 5$ we have

$$\frac{a_{k+1}/2^{k+1}}{a_k/2^k} < 1 + \frac{1}{2k^2 \log^2 2} + \frac{1}{2^{k+1}} < 1 + \frac{3}{2k^2} < \frac{1 - \frac{2}{k+2}}{1 - \frac{2}{k+1}}$$

so that for $k \geq 6$ we have

$$\frac{a_{k+1}}{2^{k+1}} < \frac{90}{64} \prod_{i=6}^{k} \frac{1 - \frac{2}{i+2}}{1 - \frac{2}{i+1}} < \frac{90}{64} \prod_{i=6}^{\infty} \frac{1 - \frac{2}{i+2}}{1 - \frac{2}{i+1}} = \frac{90}{64} \cdot \frac{7}{5} = \frac{63}{32},$$

proving that $a_k < 2^{k+1}$ for any $k \geq 7$.

$\square$

Remark 4.3. The sequence $(a_k/2^k)_{k \geq 1}$ is increasing, so that it follows from lemma 4.2 that $\lim_{n \to \infty} a_k/2^k = a$, with $a \in]1, 2[$.

Remark 4.4. For any $k \geq 1$ we have $2a_k < a_{k+1} \leq 3a_k$ and for any fixed $\eta_1 > 0$ we can easily compute explicitely $k_1 \in \mathbb{N}$ such that for any $k \geq k_1$ we have $a_{k+1} < \lceil a 2^{k+1} \rceil \leq (2 + \eta_1) a_k$.

Let g be a function satisfying conditions $(\mathcal{C}_0)$ and K_0 be a fixed large constant which will be chosen later (depending on the parameter η_0 of the statement of Theorem 4.1). Let us define the sequence $(m_k)_{k \geq K_0}$ by $m_{K_0} = 2$ and, for $k \geq K_0$,

$$m_{k+1} = \min({m_k}^2, \lceil \frac{g(\lceil a 2^{k+1} \rceil)}{m_k} \rceil m_k).$$

Remark 4.5. The sequence $(m_k)_{k\geq 1}$ is well defined because we have $m_k \geq 2$ for any $k \geq K_0$.

Lemma 4.6. *There exists an integer $K_1 \geq K_0$ such that*

$$m_{k+1} = \lceil \frac{g(\lceil a2^{k+1} \rceil)}{m_k} \rceil m_k \text{ for any } k \geq K_1.$$

Proof. Let us first remark that, if we suppose that $m_{k+1} = {m_k}^2$ for any $k \geq K_0$, then it would follow from one side that

$$m_k = {m_{K_0}}^{2^{k-K_0}} = \lambda^{a2^{k+1}} \text{ for any } k \geq K_0,$$

with $\lambda = 2^{\frac{1}{a2^{K_0+1}}} > 1$, and from the other side that

$$m_k \leq \lceil \frac{g(\lceil a2^{k+1} \rceil)}{m_k} \rceil \text{ for any } k \geq K_0,$$

which would imply all together that

$$\begin{aligned} g(\lceil a2^{k+1} \rceil) &> m_k(m_k - 1) \\ &\geq \frac{1}{2}{m_k}^2 = \frac{1}{2}\lambda^{a2^{k+2}} \\ &> \frac{1}{2}\lambda^{\lceil a2^{k+1} \rceil}, \end{aligned}$$

which would contradict the hypothesis $\lim_{n\to\infty} \frac{1}{n} \log_q g(n) = 0$.

This prove the existence of an integer K_1 such that $m_{K_1+1} = \lceil \frac{g(\lceil a2^{K_1+1} \rceil)}{m_{K_1}} \rceil m_{K_1}$, i.e. such that $\lceil \frac{g(\lceil a2^{K_1+1} \rceil)}{m_{K_1}} \rceil \leq m_{K_1}$.

It is now easy to prove by induction over k that

$$\lceil \frac{g(\lceil a2^{k+1} \rceil)}{m_k} \rceil \leq m_k \text{ for any } k \geq K_1.$$

As for any $(x, n) \in \mathbb{R} \times \mathbb{Z}$ the inequality $\lceil x \rceil \leq n$ is equivalent to the inequality $x \leq n$, it is equivalent to prove that

$$g(\lceil a2^{k+1} \rceil) \leq {m_k}^2 \text{ for any } k \geq K_1.$$

Indeed, this is true for $k = K_1$ and if we suppose that $g(\lceil a2^{k+1}\rceil) \leq {m_k}^2$, i.e. that $m_{k+1} = \lceil \frac{g(\lceil a2^{k+1}\rceil)}{m_k}\rceil m_k$, then we have

$$\begin{aligned} g(\lceil a2^{k+2}\rceil) &\leq g(2\lceil a2^{k+1}\rceil) \\ &\leq (g(2\lceil a2^{k+1}\rceil))^2 \text{ by condition } (\mathcal{C}_0, ii) \\ &\leq (\lceil \frac{g(\lceil a2^{k+1}\rceil)}{m_k}\rceil m_k)^2 = {m_{k+1}}^2. \end{aligned}$$

□

The following lemma shows that the sequences $(m_k)_{k\geq K_0}$ and $(g(\lceil a2^k\rceil))_{k\geq K_0}$ have the same order of magnitude:

Lemma 4.7. *(i) For any integer $k \geq K_0$ we have $m_k \leq 2g(\lceil a2^k\rceil)$.*

(ii) For any integer $k \geq K_1 + 1$ we have $m_k \geq g(\lceil a2^k\rceil)$.

Proof. (i) Let us prove this inequality by induction over k. It is true for $k = K_0$ and, if we suppose that $m_k \leq 2g(\lceil a2^k\rceil)$, it follows that

$$\begin{aligned} m_{k+1} &\leq \lceil \frac{g(\lceil a2^{k+1}\rceil)}{m_k}\rceil m_k \\ &\leq 2g(\lceil a2^{k+1}\rceil), \end{aligned}$$

because $\frac{g(\lceil a2^{k+1}\rceil)}{m_k} \geq \frac{g(\lceil a2^k\rceil)}{m_k} \geq \frac{1}{2}$ (we recall that if $x \geq \frac{1}{2}$, then we have $\lceil x\rceil \leq 2x$).

(ii) If $k \geq K_1$, we have

$$m_{k+1} = \lceil \frac{g(\lceil a2^{k+1}\rceil)}{m_k}\rceil m_k \geq g(\lceil a2^{k+1}\rceil).$$

□

Starting from $M(K_0) = \{0^{a_{K_0}}, 0^{a_{K_0}-1}1\}$ we define by induction for each $k \geq K_0$ a set $M(k)$ of m_k words of length a_k as follows:

If $M(k)$ has been already constructed, we choose for each $\alpha \in M(k)$ a set $X(\alpha) \subset M(k)$ with $|X(\alpha)| = \frac{m_{k+1}}{m_k}$. Then we construct

$$M(k+1) = \{\alpha 0^{b_k}\beta, \quad \alpha \in M(k), \quad \beta \in X(\alpha)\}.$$

We denote by $\mathcal{M}(k)$ the union, for all the possible choices of the sets $X(\alpha)$, of the sets $M(k)$ and by W the set of infinite words w on the alphabet A such that $\pi_{a_k}(w) \in \mathcal{M}(k)$ for any integer $k \geq K_0$.

4.2 Complexity of the elements of W

The goal of this paragraph is to show the following proposition:

Proposition 4.8. *For any fixed $\eta_0 > 0$ there exists an integer n_0 such that for any $n \geq n_0$ we have, for any $w \in W$:*

$$\frac{1}{2}g((\frac{1}{2} - \eta_0)n) < p_w(n) < \min(G((2+\eta_0)n\log^2 n), G(((2+\eta_0)n)^2).$$

Proof. It is easy to bound p_w from below:
If $a_k \leq n < a_{k+1}$, we have

$$\begin{aligned} p_w(n) &\geq m_k \quad \text{by construction} \\ &\geq g(\lceil a2^k \rceil) \quad \text{by Lemma 4.7 ii} \\ &\geq g(a_k). \end{aligned}$$

It follows from Remark 4.4 that, if $n \geq a_{k_1}$ we have $p_w(n) \geq \frac{1}{2}g((\frac{1}{2} - \eta_1)n)$.

We have now to give upper bounds for p_w.

Lemma 4.9. *Let τ be the function defined on the interval $[e^2, +\infty)$ by $\tau(x) = \frac{x}{(\log x)^2}$. The function τ is strictly increasing and, for any fixed $\eta_2 > 0$, we can compute explicitely $n_2 \in \mathbb{N}$ such that for any $n \geq n_2$ we have*

$$\tau^{-1}(n) \leq (1+\eta_2)n\log^2 n. \tag{2}$$

Proof. The study of the derivative of the function τ shows easily that τ is strictly increasing on the interval $[e^2, +\infty)$. The inequality (2) is then equivalent to

$$n \leq \tau((1+\eta_2)n\log^2 n),$$

that is equivalent to

$$\left(1 + \frac{\log(1+\eta_2)}{\log n} + 2\frac{\log\log n}{\log n}\right)^2 \leq 1 + \eta_2,$$

which clearly holds for n large enough. $\square$

For any fixed $\eta_3 > 0$, let us fix η_1 and η_2 respectively in Remark 4.4 and Lemma 4.9 such that $(2+\eta_1)(1+\eta_2) \leq 2+\eta_3$. We denote $n_3 = \max(b_{k_1+1}, n_2)$.

To bound p_w from above let us consider, for any integer $n \geq n_3$, $k_0(n)$ the smallest integer such that $b_{k_0(n)} \geq n$.

Lemma 4.10. *For any $n \geq n_3$, we have $\lceil a2^{k_0(n)} \rceil \leq (2+\eta_3) n \log^2 n$.*

Proof. By definition of $k_0(n)$ we have

$$b_{k_0(n)-1} < n \leq b_{k_0(n)}$$

and by definition of $(b_k)_{k\geq 1}$ we have

$$\tau(a_{k_0(n)-1}) \leq b_{k_0(n)-1} < \tau(a_{k_0(n)-1}) + 1.$$

It follows from Lemma 4.9 that

$$a_{k_0(n)-1} < \tau^{-1}(n) \leq (1+\eta_2) n \log^2 n$$

and from Remark 4.4 that

$$\lceil a2^{k_0(n)} \rceil \leq (2+\eta_1) a_{k_0(n)-1} < (2+\eta_3) n \log^2 n.$$

□

Let us now use the fact that every factor of length $n \geq n_3$ in w must be a factor of some element of $M(k_0(n))$ preceeded or followed by a sequence of zeros.

This means that for $n \geq n_3$ we have

$$\begin{aligned} p_w(n) &\leq (n - 1 + a_{k_0(n)}) m_{k_0(n)} + 1 \\ &\leq (n + a_{k_0(n)}) m_{k_0(n)} \\ &< 2(n + (2+\eta_3) n \log^2 n) g((2+\eta_3) n \log^2 n). \end{aligned}$$

If we fix now $\eta_4 > 0$ such that $\eta_4 > \eta_3$, there exist an integer $n_4 \geq n_3$ such that for $n \geq n_4$ we have

$$p_w(n) < G((2+\eta_4) n \log^2 n).$$

Let us now give another upper bound for p_w that will give a better result when g is growing very fast.

Every factor of length n in w must be a factor of some element of $M(k+1)$ (where $a_k \leq n < a_{k+1}$) preceded or followed by a sequence of zeros, or a factor of $M(k+1)$ followed by b_r zeros (for some $k+1 \leq r \leq k_0(n)$) followed by another factor of $M(k+1)$.

This gives the estimate valid for $n \geq n_3$:

$$\begin{aligned} p_w(n) &\leq (n + a_{k+1})m_{k+1} + (k_0(n) - k)nm_{k+1}^2 \\ &\leq 4ng(\lceil a2^{k+1} \rceil) + 4(k_0(n) - k)n \cdot g(\lceil a2^{k+1} \rceil)^2 \\ &\leq 4ng((2 + \eta_1)n) + 4\log_2((2 + \eta_3)n \log^2 n)n \cdot g((2 + \eta_1)n)^2. \end{aligned}$$

This shows that there exists an integer $n_5 \geq n_3$ such that for $n \geq n_5$ we have

$$p_w(n) < G((2 + \eta_1)n)^2.$$

To finish the proof of Proposition 4.8 it is enough, for any fixed η_0, to take in the previous arguments $\eta_1 < \eta_0$, $\eta_4 < \eta_0$ and $n_0 = \max(a_{k_1}, n_4, n_5)$.

□

Remark 4.11. The above majoration of $k_0(n) - k$ is a simple application of Lemmas 4.2 and 4.10. It is easy to improve it by showing that

$$k_0(n) - k = 2\frac{\log \log n}{\log 2} + O(1).$$

Corollary 4.12. *If g verifies the conditions $(\mathcal{C}_0)$, η_0 and n_0 are as in the statement of Proposition 4.8 and K_0 satisfies $b_{K_0} > n_0$ then, for any $w \in W$ and any $n \geq 1$, we have $p_w(n) \leq f(n)$.*

Proof. We have two cases:

i) If $n \leq b_{K_0}$, by construction a factor of size n of a word $w \in W$ has at most one letter equal to 1 and all other letters equal to 0, so $p_w(n) \leq n + 1 \leq f(n)$.

ii) If $n > b_{K_0}$, we have $k_0(n) > K_0$ and, since $b_{K_0} > n_0$, we have $p_w(n) < \min(G((2 + \eta_0)n \log^2 n), G((2 + \eta_0)n)^2) \leq f(n)$.

□

4.3 Minoration of $\left|\bigcup_{w\in W} L_n(w)\right|$

For any $k \geq K_0$, let

$$r(k) = \lceil \log_2 m_k \rceil. \tag{3}$$

For every integer $n \geq a_{K_0+r(K_0)}$, let k be the unique integer verifying

$$a_{k-1+r(k-1)} \leq n < a_{k+r(k)}$$

and let s defined by

$$a_{k+s} \leq n < a_{k+s+1}$$

(we have $r(k-1) - 1 \leq s \leq r(k) - 1$).

We will now construct subsets of W as follows. Let us enumerate the set $M(k)$ obtained in the construction described in section 4.1 as follows: $M(k) = \{\alpha_1(k), \alpha_2(k), \ldots, \alpha_{m_k}(k)\}$. We can decide that for $k' \geq k$ we have $\alpha_{j+1}(k') \in X(\alpha_j(k'))$ for each $1 \leq j \leq m_{k'}$ (we put $\alpha_{m_{k'}+1} := \alpha_1$) and

$$M(k'+1) = \{\alpha_1(k'+1), \alpha_2(k'+1), \ldots, \alpha_{m_{k'+1}}(k'+1)\}$$

where we enumerate the elements of $M(k'+1)$ in such a way that

$$\begin{aligned}
\alpha_1(k'+1) &= \alpha_1(k')0^{b_{k'}}\alpha_2(k') \\
\alpha_2(k'+1) &= \alpha_3(k')0^{b_{k'}}\alpha_4(k') \\
&\vdots \\
\alpha_{\lfloor \frac{m_{k'}+1}{2} \rfloor}(k'+1) &= \alpha_{m_{k'}-1}(k')0^{b_{k'}}\alpha_{m_{k'}}(k') \quad \text{ for } m_{k'} \text{ even} \\
&= \alpha_{m_{k'}}(k')0^{b_{k'}}\alpha_1(k') \quad \text{ for } m_{k'} \text{ odd.}
\end{aligned}$$

This construction gives

$$\alpha_1(k+s) = \alpha_1(k)0^{b_k}\alpha_2(k)0^{b_{k+1}} \ldots 0^{b_{k+1}}\alpha_{2^s-1}(k)0^{b_k}\alpha_{2^s}(k)$$

where $\alpha_1(k), \alpha_2(k), \ldots, \alpha_{2^s}(k)$ appear in this order as factors of length a_k.

Lemma 4.13. *i) For every integer* $n \geq a_{K_0+r(K_0)}$ *we have* $n < 4a_k 2^s < 4a_k m_k$.

ii) For every integer $n \geq \max(a_{k_1+r(k_1)}, a_{K_1+1+r(K_1+1)})$ *we have*

$$n > \frac{1}{4} G\left(\frac{a_k}{2+\eta_1}\right).$$

Proof. i) We have

$$\begin{aligned} n &< a_{k+s+1} && \text{by construction,} \\ &< 2^{k+s+2} && \text{by Lemma 4.2,} \\ &< 2^{s+2} a_k && \text{by Lemma 4.2.} \end{aligned}$$

The second inequality results from the fact that

$$2^s \leq 2^{r(k)-1} = 2^{\lceil \log_2 m_k \rceil - 1} < m_k.$$

ii) We have, for any $k \geq K_1 + 2$,

$$\begin{aligned} n &\geq a_{k-1+r(k-1)} && \text{from the definition of } k, \\ &> 2^{k-1+r(k-1)} && \text{by Lemma 4.2,} \\ &> \frac{1}{2} a_{k-1} 2^{r(k-1)} && \text{by Lemma 4.2,} \\ &\geq \frac{1}{2} a_{k-1} m_{k-1} && \text{from the definition of } r, \\ &\geq \frac{1}{2} a_{k-1} g(\lceil a 2^{k-1} \rceil) && \text{by Lemma 4.7 ii,} \\ &\geq \frac{1}{2} a_{k-1} g(a_{k-1}). \end{aligned}$$

It follows from Remark 4.4 that if $k \geq \max(k_1 + 1, K_1 + 2)$, we have

$$n > \frac{1}{2} a_{k-1} g(a_{k-1}) > \frac{1}{4} \cdot \frac{2a_k}{2+\eta_1} g\left(\frac{a_k}{2+\eta_1}\right) = \frac{1}{4} G\left(\frac{a_k}{2+\eta_1}\right).$$

□

We have $2^s \leq 2^{r(k)-1} < m_k$ and, if we denote by W_0 the set of all infinite words obtained by this construction, we have

$$\left| \bigcup_{w \in W_0} L_n(w) \right| \geq A_{m_k}^{2^s} = \frac{m_k!}{(m_k - 2^s)!}.$$

For any fixed $\eta_5 > 0$ there is k_5 such that for any $k \geq k_5$ we have

$$\frac{m_k!}{(m_k - 2^s)!} \geq ((m_k!)^{1/m_k})^{2^s} \geq (m_k/e)^{2^s} \geq m_k^{(1-\eta_5)2^s}.$$

Then, for any $k \geq \max(k_1 + 1, K_1 + 2, k_5)$ we have

$$\begin{aligned} \frac{m_k!}{(m_k - 2^s)!} &\geq \exp((1 - \eta_5)2^s \log m_k) \\ &> \exp\left((1 - \eta_5)\frac{n}{4a_k} \log \frac{n}{4a_k}\right) \quad \text{by Lemma 4.13.i,} \\ &> \exp\left(\frac{1 - \eta_5}{4(2 + \eta_1)} \cdot \frac{n}{G^{-1}(4n)} \log \frac{n}{4(2 + \eta_1)G^{-1}(4n)}\right) \quad \text{by Lemma 4.13.ii.} \end{aligned}$$

Now for any fixed $\eta_0 > 0$ and any $\eta_1 > 0$ fixed as in part 4.2 (in particular $\eta_1 < \eta_0$), let us choose η_5 such that $\eta_5 < 4\eta_0(2 + \eta_1) - \frac{\eta_1}{2}$. Then, we have $\frac{1}{8} - \eta_0 < \frac{1-\eta_5}{4(2+\eta_1)}$ and we conclude that there exists an integer $N_0 = \max(a_{k_1+r(k_1)}, a_{K_1+1+r(K_1+1)}, a_{k_5-1+r(k_5-1)}, n_0)$ such that, for any $n \geq N_0$ we have

$$\left|\bigcup_{w\in W} L_n(w)\right| \geq \left|\bigcup_{w\in W_0} L_n(w)\right| > \exp\left((\frac{1}{8} - \eta_0)\frac{n}{G^{-1}(4n)} \log \frac{n}{G^{-1}(4n)}\right).$$

□

Examples 4.14.

- For f defined in Example A, we can take for $N \geq e^{\frac{2\alpha}{\alpha-1}}$

$$G(N) = \frac{N^\alpha}{(2 + \eta_0)^\alpha \log^{2\alpha} N},$$

so that

$$G^{-1}(4n) = \frac{4^{1/\alpha}(2 + \eta_0)}{\alpha^2} n^{1/\alpha} \log^2 n + O(n^{1/\alpha} \log n \log\log n).$$

If we combine this with the result obtained in section 3, we conclude that there are positive constants $c_1(\alpha)$ and $c_2(\alpha)$ such that, for n big enough, we have

$$\exp\left(c_1(\alpha)\frac{n^{(\alpha-1)/\alpha}}{\log n}\right) < |\mathcal{L}_n(f)| < \exp\left(c_2(\alpha)n^{(\alpha+1)/(\alpha+2)}(\log n)^{(\alpha+1)/(\alpha+2)}\right)$$

(indeed we can take any $c_1(\alpha) < 4^{-1/\alpha}\alpha(\alpha - 1)/16$ and any $c_2(\alpha) > 2(\log q)^{\frac{1}{\alpha+2}}(\frac{\alpha}{\alpha+2})^{\frac{\alpha+1}{\alpha+2}}$).

- For f defined in Example B, we can take for $N \geq (2+\eta_0)(\frac{2}{\alpha \log q})^{1/\alpha}$

$$G(N) = q^{\frac{N^\alpha}{2(2+\eta_0)^\alpha}}$$

so that

$$G^{-1}(4n) = \left(\frac{2}{\log q}\right)^{1/\alpha} (2+\eta_0)(\log n)^{1/\alpha} + O((\log n)^{(1-\alpha)/\alpha}).$$

Combining this with the result obtained in section 3, we conclude that there are constants $c_1(\alpha)$ and $c_2(\alpha)$, $0 < c_1(\alpha) < c_2(\alpha)$ such that, for n big enough, we have

$$\exp\left(c_1(\alpha)\frac{n}{(\log n)^{(1-\alpha)/\alpha}}\right) < |\mathcal{L}_n(f)| < \exp\left(c_2(\alpha)\frac{n}{(\log n)^{(1-\alpha)/\alpha}}\right)$$

(indeed we can take any $c_1(\alpha) < \frac{1}{16}(\frac{\log q}{2})^{\frac{1}{\alpha}}$ and any $c_2(\alpha) > 2(\log q)^{\frac{1}{\alpha}}$).

4.4 An open question

Our method does not work for sequences with sublinear complexity. A natural open problem is to give sharp estimates for $|\mathcal{L}_n(f)|$ when f is a linear function.

In order to state more precise questions, let us give some definitions. Let $g_0(x) = x, g_1(x) = x+1$, and, for $k > 0$ and $x > 0$ large, $g_{k+1}(x) = \exp(g_k(\log(x))$ and $g_{-k}(x) = g_k^{-1}(x)$. We say that an increasing function f from $\mathbb{R}^+$ to $\mathbb{R}^+$ is *morally polynomial* if there is $k \geq 0$ such that $g_{-k}(x) \leq f(x) \leq g_k(f(x))$ for every x sufficiently large, and that f is *morally exponential* if $\log f$ is morally polynomial. We have the following questions:

i) Is it true that $\ell(n) = |\mathcal{L}_n(f)|$ is morally polynomial for any linear function f ?

ii) Does there exist some $A > 0$ such that $\ell(n) = |\mathcal{L}_n(f)|$ is morally exponential for $f(n) = An$?

Clearly we cannot have positive answers to both of these questions. On the other hand, it is not clear whether we will have a positive answer to one of them, since there are functions which are neither morally polynomial nor morally exponential, (e.g. increasing functions f such that $f \circ f = exp$). However, any logarithmico-exponential function f (in the sense of Hardy) satisfying $x \leq f(x) \leq q^x$ for every large x is morally polynomial or morally exponential (see section 4.1 of [Har]).

References

[AB1] B. Adamczewski and Y. Bugeaud, On the complexity of algebraic numbers II. Continued fractions, *Acta Math.*, 195 (2005), 1-20.

[AB2] B. Adamczewski and Y. Bugeaud, On the complexity of algebraic numbers I. Expansions in integer bases, *Annals of Math.*, 165 (2) (2007), 547-565.

[Ale] P. Alessandri, Codages de rotations et basses complexités, *Ph. D. thesis. Université Aix-Marseille 2*, 1996.

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

[ADQZ] J.-P. Allouche, J.-L. Davison, M. Queffélec, L. Zamboni, Transcendence of Sturmian or morphic continued fractions, *J. Number Theory*, 91(1) (2001), 39-66.

[BP] J. Berstel and M. Pocchiola, A geometric proof of the enumeration formula for Sturmian words, *Inter. J. Algebra Comput.*, 3 (1993), 349-355.

[Cas1] J. Cassaigne, Special factors of sequences with linear subword complexity, In *Developments in language theory, II (Magdeburg, 1995)*, 25-34. Word Sci. Publishing, 1996.

[Cas2] J. Cassaigne, Complexité et facteurs spéciaux, *Bull. Belg. Math. Soc. Simon Stevin*, 4(1) (1997), 67-88.

[Cas3] J. Cassaigne. Constructing infinite words with intermediate complexity. In *Developments in language theory, II (DLT 2002, Kyoto), Lecture Notes in Computer Sciences 2450*, 173-184. Springer, 2003.

[CH] E.M. Coven and G.A. Hedlund, Sequences with minimal block growth, *Math. Systems Theory* 7 (1973), 138-153.

[DG] S. Dulucq and D. Gouyou-Beauchamps, Sur les facteurs des suites de Sturm, *Theoret. Comput. Sci.* 71 (1990), 381-400.

[Fer] S. Ferenczi, Complexity of sequences and dynamical systems, *Discrete Math.*, 206(1-3) (1999), 145-154.

[FM] S. Ferenczi and C. Mauduit, Transcendence of numbers with a low complexity expansion, *J. Number Theory*, 67 (1997), 146-161.

[Goy] J. Goyon, Construction de suites dont la complexité est donnée, Preprint, *Institut des Mathématiques de Luminy*, 1997.

[Har] G. H. Hardy, Orders of Infinity, *Cambridge Tracts in Mahematics*, no. 12, Cambridge University Press, second edition, 1924.

[HM1] G. A. Hedlund and M. Morse, Symbolic dynamics. *Amer. J. Math.*, 60 (1938), 815-866.

[HM2] G. A. Hedlund and M. Morse, Symbolic dynamics II. Sturmian trajectories. *Amer. J. Math.*,62 (1940), 1-42.

[KLB] J. Koplowitz, M. Lindenbaum and A. Bruckstein, The number of digital straight lines on an $n \times n$ grid, *IEEE Trans. Inform. Th.*, 36 (1990), 192-197.

[Kůr] P. Kůrka, Topological and symbolic dynamics, *Cours spécialisés SMF*,volume 11, 2003.

[Lip] E. P. Lipatov, On some classification of binary words and some properties of uniformity classes, *The Problems of Cybernetics* 39 (1982), 67-84 (in Russian).

[LM] A. de Luca and F. Mignosi, Some combinatorial properties of Sturmian words, *Theoret. Comput. Sci.* 136 (1994), 361-385.

[Lot] M. Lothaire, Algebraic combinatorics on words, volume 90 of *Encyclopedia of Mathematics and Its Applications.* Cambridge University Press, 2002.

[MS1] C. Mauduit and A. Sárközy, On finite binary pseudorandom sequences. I. Measure of pseudorandomness, the Legendre symbol. *Acta Arithmetica* 82 (1997), 365-377.

[MS2] C. Mauduit and A. Sárközy, On finite binary pseudorandom sequences. II. The Champernowne, Rudin-Shapiro and Thue-Morse sequence, a further construction. *J. Number Theory* 73 (2) (1998), 256-276.

[Mig] F. Mignosi, On the number of factors of sturmian words, *Theoret. Comput. Sci.* 82 (1991), 71-84.

[MZ] F. Mignosi and L. Zamboni, On the number of Arnoux-Rauzy words, *Acta Arithmetica* 101 (2002), 121-129.

[PF] N. Pytheas Fogg. Substitutions in dynamics, arithmetics and combinatorics, *Lecture Notes in Mathematics 1794*, Springer, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.

[Que] M. Queffélec, Substitution dynamical systems — spectral analysis, *Lecture Notes in Mathematics 1294*, Springer, 1987.

CHRISTIAN MAUDUIT
Institut de Mathématiques de Luminy
163, avenue de Luminy
13288 Marseille Cedex 9, France
E-mail: mauduit@iml.univ-mrs.fr

CARLOS GUSTAVO MOREIRA
Instituto de Matemática Pura e Aplicada
Estrada Dona Castorina 110
22460-320 Rio de Janeiro, RJ, Brasil
E-mail: gugu@impa.br