Theoretical Informatics and Applications
Will be set by the publisher
DOI: (will be inserted later)

SEQUENCES OF LOW ARITHMETICAL COMPLEXITY

S.V. Avgustinovich[1], J. Cassaigne[2] and A.E. Frid[1]

Abstract. Arithmetical complexity of a sequence is the number of words of length n that can be extracted from it according to arithmetic progressions. We study uniformly recurrent words of low arithmetical complexity and describe the family of such words having lowest complexity.

Mathematics Subject Classification. 68R15.

Introduction

The classical function of subword complexity of an infinite word counts the number of its factors of a given length n. According to this definition of complexity, periodic words are the simplest since they have ultimately constant complexity, and random words have maximal possible complexity. It is well-known that complexity of a not ultimately periodic word is at least $n + 1$, and the family of *Sturmian* words having complexity $n+1$ is extensively studied and has non-trivial properties [4]. In general, the function of subword complexity has intriguing properties, and generates a very interesting classification of infinite words which is still far from being complete [6, 10].

Recently, several modifications of the notion of complexity, also based on counting words related to the infinite word, have been introduced. They include palindrome complexity [2], modified complexity by Nakashima *et al.* [18], pattern complexity by Restivo and Salemi [19], maximal pattern complexity by Kamae and

Keywords and phrases. Arithmetical complexity, infinite words, Toeplitz words, special factors, period doubling word, Legendre symbol, paperfolding word.

[1] Sobolev Institute of Mathematics SB RAS, Koptyug Av. 4, Novosibirsk, Russia;
avgust@math.nsc.ru & frid@math.nsc.ru
Supported in part by RFBR grants 03-01-00796 and 05-01-00364 and by the Program of high school scientific development of Russian Ministry of Education (grant 8287).

[2] Institut de Mathématiques de Luminy, case 907, 163 Av. de Luminy, 13288 Marseille Cedex 9, France; cassaigne@iml.univ-mrs.fr

Zamboni [16] and others. *Arithmetical complexity*, defined in 2000 by Avgustinovich, Fon-Der-Flaass and Frid [3], also belongs to this family and counts not only factors of a word but all words occurring in it in arithmetic progressions.

For each of these complexity functions, usual questions arise, including classification of infinite words according to their complexity, possible rates of complexity growth, etc. In particular, it is always interesting to understand how low the complexity of a non-periodic word can be and study words of lowest complexity, analogous to Sturmian words. For example, palindrome complexity can be ultimately zero, and minimal maximal pattern complexity of a non-periodic word is $2n$ [16].

In this paper, we study uniformly recurrent non-periodic words of lowest arithmetical complexity. Contrary to the situation with Sturmian words, here we are not able to write down a word of minimal complexity, but can find a family of words with decreasing lower limits of arithmetical complexity divided by n, which tend to be minimal. We also prove that the words found are essentially the only uniformly recurrent words of such low arithmetical complexity.

The technique we use here also allows to characterize all uniformly recurrent words of linear arithmetical complexity [14]. It is interesting that all such words are Toeplitz words not Sturmian words. In their turn, Sturmian words have arithmetical complexity $\Theta(n^3)$ [7,11] which depends on the *slope* of the word; in [7], we found it explicitly for many cases including the Fibonacci word. For other results on arithmetical complexity, see [12,13].

After introducing basic required notions in Sections 1 and 2, in Section 3 we describe the main tool for the proof, namely infinite special branches, and state the main theorem which is a description of all uniformly recurrent words which could be considered to have lowest arithmetical complexity. We prove it in Sections 4 and 5 and study complexity of the listed words in Section 6, extracting the needed family of words having lowest complexity. Section 7 is the conclusion, and properties of lowest possible growth of arithmetical complexity of uniformly recurrent words are listed in it.

1. Preliminaries

Let Σ be a finite alphabet; then the set of all finite (infinite) words on Σ is denoted by Σ^* (Σ^ω). If t is a non-empty finite word, then t^ω denotes the infinite word $tt\cdots t\cdots$. In what follows, we use the terms "infinite word" and "sequence" as synonyms.

A *factor* of a finite or infinite word w on Σ is a finite word u such that $w = s_1 u s_2$ for some (possibly empty) words s_1 and s_2; the set of factors of a word w is denoted by $F(w)$. The length of a finite word w is denoted by $|w|$.

A finite or infinite word $w_1 w_2 \cdots w_n \cdots$ where $w_i \in \Sigma$, is called *(q-)periodic* if for all $i > 0$ such that $i + q \le |w|$ we have $w_i = w_{i+q}$. The minimal possible length q is called the *period* of w. An infinite word is periodic if and only if it is equal

t^ω for some t. An infinite word is called *ultimately periodic* if it is equal to st^ω for some t and (possibly empty) s.

An *arithmetical subsequence* of an infinite word $w = w_1w_2\cdots w_n\cdots$, where $w_i \in \Sigma$, is a word $w_d^k = w_kw_{k+d}w_{k+2d}\cdots w_{k+nd}\cdots$, where $k, d > 0$. If the word w is finite, $w = w_1w_2\cdots w_n$, then we also denote by w_d^k the (finite) word $w_kw_{k+d}w_{k+2d}\cdots w_{k+md}$, where $k + md \leq n < k + (m + 1)d$. A factor of some w_d^k is called an *arithmetical subword* of w.

The *arithmetical closure* of w is the set of its arithmetical subwords, *i.e.*, $A(w) = \cup_{d,k>0}F(w_d^k)$.

One of the most famous results about the arithmetical closure of an infinite word is the Szemerédi theorem, which can be stated as follows. Let $w(a, n)$ denote the number of occurrences of a letter $a \in \Sigma$ to the prefix of length n of w.

Theorem 1.1 (Szemerédi [20])**.** *If* $\limsup_N w(a, N)/N > 0$*, then* $a^n \in A(w)$ *for all* n*.*

The *(subword) complexity* of a finite or infinite word w is the function $f_w(n)$ counting the number of its factors of length n. Similarly, the *arithmetical complexity* $a_w(n)$ counts the number of words of $A(w)$ of length n. This paper is devoted to sequences having extremely low arithmetical complexity.

An infinite word $w \in \Sigma^\omega$ is called *uniformly recurrent* if all its factors occur in it an infinite number of times with bounded gaps, *i.e.*, if each sufficiently long factor of w contains all factors of w of a given length. Equivalently, an infinite word is uniformly recurrent if and only if each of its prefixes occurs in it an infinite number of times with bounded gaps. Clearly, an ultimately periodic word is uniformly recurrent if and only if it is periodic. In this paper, we consider only uniformly recurrent words.

The Szemerédi theorem is clearly valid for any symbol occurring in a uniformly recurrent word. We also know the following two lemmas about uniformly recurrent words. Both of them seem to be folklore, although the proof of the first one can be found in [3], and the second one is proved *e.g.* in [15] (Prop. 6).

Lemma 1.2. *An arithmetical subsequence of a uniformly recurrent word is uniformly recurrent.*

Lemma 1.3. *If the languages of factors of two uniformly recurrent words have an infinite intersection, then these languages coincide.*

A language $F \subseteq \Sigma^*$ is called *factorial* if it is closed under taking factors. A factorial language is called *prolongable* if for each of its elements u we have $aub \in F$ for some $a, b \in \Sigma$. Clearly, for each infinite word w, the languages $F(w)$ and $A(w)$ are factorial; it w is uniformly recurrent, they are also prolongable. Analogously to the definition for infinite words, we define the arithmetical closure $A(F)$ and the subword complexity $f_F(n)$ of a factorial language F. In particular, for each infinite word w we by definitions have $A(F(w)) = A(w)$ and $a_w(n) = f_{A(w)}(n)$. Since $A(A(F)) = A(F)$ for any factorial language F, it is indeed a closure.

An infinite word y is said to belong to the *orbit* of an infinite word w if $F(y) \subseteq F(w)$. If w is uniformly recurrent, this implies to $F(y) = F(w)$ (and

this implication can be considered as an equivalent definition of a uniformly recurrent word). Since the set of factors does not change for all elements of the orbit of a uniformly recurrent word, it is reasonable to consider orbits when dealing with subword and arithmetical complexities.

2. Toeplitz words

Let ? be a symbol called *gap* which does not belong to Σ. A *pattern* is a word $P \in (\Sigma \cup \{?\})^*$. A Toeplitz transform $T_P : (\Sigma \cup \{?\})^\omega \to (\Sigma \cup \{?\})^\omega$ maps an infinite word w, probably containing gaps, to the infinite word obtained from P^ω by filling gaps with letters of w, *i.e.*, if $P = u_1?u_2? \cdots u_q?$ and $w = w_1 w_2 \cdots w_n \cdots$, where $u_i \in \Sigma^*$ and $w_i \in \Sigma \cup \{?\}$ for all i, then $T_P(w) = u_1 w_1 u_2 w_2 \cdots u_q w_q u_1 w_{q+1} \cdots u_q w_{2q} u_1 \cdots$. The mapping T_P can be defined also for finite words whose length is divided by the number of gaps in the pattern P as the result of substituting w to the gaps of the appropriate power of P. For example, $T_{1?(-1)?}((-1)(-1)??11) = 1(-1)(-1)(-1)1?(-1)?11(-1)1$.

More generally, for each $u, w \in (\Sigma \cup \{?\})^\omega$, we define $T_u(w)$ as the result of substituting w to the gaps of u, so, $T_P(w) = T_{P^\omega}(w)$.

From now on, we consider $\Sigma = \{1, -1\}$. Let w be a finite or infinite word on $\Sigma \cup \{?\}$; in what follows we define the word w' as the result of interchanging 1's and -1's in w. For example, $(111?(-1))' = (-1)(-1)(-1)?1$.

Let a pattern P start with a symbol of Σ. Then clearly the equation $x = T_P(x)$ has a unique solution in Σ^ω. It can be built by the following iterating process: let $U_0 = ?^\omega$; for each $i > 0$, we define $U_i = T_{U_{i-1}}(P^\omega)$; then $x = \lim_{i \to \infty} U_i$. The word x is called a *Toeplitz word* generated by the pattern P.

Analogously, we can consider the equation $x = T_P(x')$: if P starts with a symbol of Σ, then this equation also has a unique solution on Σ. To build it, we start with $U_0 = ?^\omega$ and define $U_i = T_{U_{i-1}}(P^\omega)$ for odd i and $U_i = T_{U_{i-1}}(P'^\omega)$ for even i. For example, for $P = 1?(-1)?$ we have $U_1 = P^\omega$, $U_2 = (T_P(P'))^\omega = (1(-1)(-1)?11(-1)?)^\omega$, and

$$x = 1(-1)(-1)111(-1)(-1)1(-1)(-1)(-1)11(-1)1 \ldots.$$

In both cases, we say that a symbol of x and its position are of nth order if it occurs instead of a gap not earlier than in U_{n+1}. In particular, each symbol of x is of order 0, although its maximal order can be arbitrarily high.

A pattern is called *regular* if it looks like $P = u_1?u_2? \cdots u_q?$, where $|u_1| = |u_2| = \cdots = |u_q|$. In this paper, we shall need only regular patterns. Toeplitz words generated by them fall in both classes considered in [8] and [17]; in particular, their subword complexity grows linearly. Clearly, if the pattern P is regular, then the symbols of nth order are exactly those at positions divided by $|u_1|^n$.

At last, note that the infinite word $x \in \Sigma^\omega$ satisfying $x = T_P(x')$ is also a Toeplitz word generated by one pattern. For example, if the number of gaps in P divides its length (in particular, if P is regular), then x is generated by the pattern $T_P(P')$.

3. Infinite special branches

A word u is called *(left) special* in a factorial language F on the binary alphabet $\Sigma = \{1, -1\}$ if both $1u$ and $(-1)u$ belong to F.

Clearly, a prefix of a special word is special, so, special words of F constitute a prefixial tree. In what follows, we identify an infinite branch of this tree, *i.e.*, a family of words u_1, u_1u_2, ..., $u_1 \cdots u_n$, ... such that all $u_i \in \Sigma$ and all $u_1 \cdots u_n$ are special in F, with the infinite word $u = u_1 \cdots u_n \cdots$. Such word u will be called an *infinite special branch* of F. An infinite special branch of F can be defined also as a limit of a sequence of special words.

The following easy statement explains our interest to infinite special branches.

Lemma 3.1. *Let F be a prolongable factorial language. If for all c we have $f_F(n) < kn - c$ for some n, then F has at most $k - 1$ infinite special branches.*

Proof. It is well-known that for each prolongable factorial language F the equality $f_F(n + 1) - f_F(n) = s_F(n)$ holds, where $s_F(n)$ is the number of special words of length n in F. If F has at least k infinite special branches, then starting from some length N we have $s_F(n) \geq k$. Thus, for all $n > N$ we have $f_F(n) \geq f_F(N) + k(n - N) = kn - c$, where $c = kN - f_F(N)$. □

In what follows we minimize the number of infinite special branches of $A(w)$.

Lemma 3.2. *Let u be an infinite special branch of the arithmetical closure $A(F)$, where F is a factorial language. Then so are u_k^k for all $k > 0$.*

Proof. By the definition of an infinite special branch of $A(F)$, for all m we have $1u_1u_2 \cdots u_{mk} \in A(F)$ and $-1u_1u_2 \cdots u_{mk} \in A(F)$. Taking arithmetical factors, we see that $1u_ku_{2k} \cdots u_{mk}, -1u_ku_{2k} \cdots u_{mk} \in A(F)$ for all m, so, $u_ku_{2k} \cdots u_{mk} \cdots = u_k^k$ is also an infinite special branch of $A(F)$. □

Lemma 3.3. *For any non-periodic uniformly recurrent infinite word w, the language $A(w)$ has at least 2 infinite special branches starting from different letters.*

Proof. Let us consider the set of words special in $F(w)$. Since w is not periodic, this set is infinite and in particular its prefixial tree has an infinite branch v. Suppose that it is constant, say, equal to 1^ω. This means that w contains arbitrarily long powers of 1, but since w is uniformly recurrent, this implies $w = 1^\omega$, which is periodic. A contradiction. So, v contains both symbols. Without loss of generality, let v start with 1 and its mth symbol be equal to -1. Then v_m^m is an infinite special branch of $A(w)$ due to Lemma 3.2; it starts with the other symbol than v. So, v and v_m^m are the two branches required. □

Corollary 3.4. *If w is a non-periodic uniformly recurrent word, then $a_w(n) \geq 2n$ for all n.*

From now on we consider the case when w is uniformly recurrent and $A(w)$ has exactly two infinite special branches. Note that it is not possible on alphabets of cardinality more than two because in fact there is an infinite special branch in

w starting with each symbol of the alphabet. This justifies our restriction to the binary alphabet $\Sigma = \{1, -1\}$.

Theorem 3.5. *Up to renaming symbols, all uniformly recurrent infinite words whose arithmetical closure has only two infinite special branches are those belonging to orbits of Toeplitz words defined by the following equations:*

(1) $w = w(p) = T_{1^{p-1}?}(w')$, *where* p *is a prime number;*
(2) $w = w_L(p) = T_{u?}(w)$ *or* $w = w_{L'}(p) = T_{u?}(w')$ *for some prime* $p \geq 3$, *where the word* u *is the sequence of Legendre symbols modulo* p: $u = \left(\frac{1}{p}\right)\left(\frac{2}{p}\right)\cdots\left(\frac{p-1}{p}\right)$;
(3) $w = w_{pf} = T_{1?(-1)?}(w)$ *and* $w = w_{pf'} = T_{1?(-1)?}(w')$ *(paperfolding words);*
(4) $w = T_P(w)$ *and* $w = T_P(w')$, *where* $P = 1?(-1)?(-1)?1?$ *or* $P = 1?1?(-1)?(-1)?$.

The next two sections are devoted to the proof of Theorem 3.5.

4. If infinite special branches are two

Let us start proving the theorem and consider a uniformly recurrent infinite word w whose arithmetical closure $A(w)$ has exactly two infinite special branches. Note that at least one of them is also an infinite special branch of $F(w)$ (which exists since w is not ultimately periodic). This branch belongs to the orbit of w, and since w is uniformly recurrent, has the same set of factors as w itself. Without loss of generality, we identify it with w and assume that it starts with 1. In what follows, we prove that w is one of the sequences listed in the statement of Theorem 3.5.

Recall that the sequence obtained from w by renaming 1 to -1 and *vice versa* is denoted by w'.

Claim 4.1. The infinite special branch of $A(w)$ not equal to w is equal to w'.

Proof. Let us denote the special branch of $A(w)$ not equal to w by v. By Lemma 3.2, $w_k = 1$ implies $w_k^k = w$ and $w_k = -1$ implies $w_k^k = v$ for all k. Let us consider i and j such that $w_i = 1$ and $w_j = -1$. Then w_{ij} is the jth symbol of $w_i^i = w$, so it is equal to -1. On the other hand, it is the ith symbol of $w_j^j = v$, so we obtain that $v_i = -1$ for all i such that $w_i = 1$. Symmetrically, if we consider v instead of w, we see that $v_i = -1$ implies $w_i = 1$, *i.e.*, these conditions are equivalent. So, $v = w'$. □

The following claim explains the choice of 1 and -1 as symbols of Σ: here they are interpreted as integers.

Claim 4.2. For all $i, j > 0$ we have $w_{ij} = w_i.w_j$.

Proof. As it was mentioned in Claim 4.1, the equalities $w_i = 1$ and $w_j = -1$ imply $w_{ij} = -1$. By the analogous argument, $w_i = w_j = 1$ implies $w_{ij} = 1$ (since it is the ith symbol of $w_j^j = w$). At last, if $w_i = w_j = -1$, then due to Claim 4.1 w_{ij} is

the ith symbol of $w_j^j = w'$; it is not equal to the ith symbol of w and thus is equal to 1. □

Corollary 4.3. *For all $m > 0$ we have $w_{m^2} = 1$.*

Claim 4.4. For all $d, k > 0$ either the arithmetical subsequence w_d^k is periodic, or its set of factors is equal to that of w or w'.

Proof. Note that w_d^k is uniformly recurrent due to Lemma 1.2. So, if it is ultimately periodic, then it is periodic. On the other hand, if it is not ultimately periodic, then its language of factors has an infinite special branch. Due to Claim 4.1, it must coincide with w or w', and its set of factors coincide with that of w_d^k. □

Claim 4.5. There exist some prime p and some $k < p$ such that the word w_p^k is periodic.

Proof. By the Szemerédi theorem, there are arbitrarily long arithmetical subsequences of the form 1^n and $(-1)^n$ in $A(w)$. Suppose that there is no periodic subsequence among the subsequences w_d^k. Due to Claim 4.4 it means that 1^n belongs to $F(w)$ or to $F(w')$ for all n, but this is impossible since w and w' are uniformly recurrent and contain (-1)s. So, the words 1^n for all n belong to the language of factors of some periodic arithmetical subsequence w_p^k of w. Here we choose p and then k as minimal possible values. It remains to prove that p is prime and $k < p$.

Suppose that p is not prime: $p = qr$ for some $1 < q < p$. Since p is chosen to be minimal, w_q^k is not periodic and $w_p^k = (w_q^k)_r^1$ constitutes a periodic arithmetical subsequence of difference $r < p$ in it. But this contradicts to the minimality of p since due to Claim 4.4, $F(w_q^k) = F(w)$ or $F(w_q^k) = F(w')$ and the minimal difference of a periodic subsequence in w and w' coincide. A contradiction.

Now suppose that $k \geq p$. If $k = p$, then $w_p^k = w_p^p$ is an infinite special branch of $A(w)$; it is equal to w or w' and thus is not periodic, a contradiction. If $k > p$, consider the arithmetical subsequence w_p^{k-p}. By minimality of k, it is ultimately periodic but not strictly periodic. Thus, it is not uniformly recurrent. This contradicts to Lemma 1.2.

The claim is proved. □

Claim 4.6. There exist a regular pattern $P = u_1?u_2? \cdots u_q?$, where $u_1, \ldots, u_q \in \Sigma^{p-1}$, for p taken from the statement of Claim 4.5, such that $w = T_P(w)$ or $w = T_P(w')$.

Proof. Let k be the number of the first symbol of the periodic arithmetical subsequence from Claim 4.5 and q be the minimal period of w_p^k.

First, let us mention that $w_p^p = w$ or $w_p^p = w'$ due to Lemma 3.2 and Claim 4.1. So, it remains to prove that for all $i \in \{1, \ldots, qp-1\}$, $i \not\equiv 0 \pmod p$, and for all $n > 0$ we have $w_i = w_{i+nqp}$.

For each $i \in \{1, \ldots, qp-1\}$, $i \not\equiv 0 \pmod p$, let us choose j such that $ij \equiv k \pmod p$. Consider the infinite words w_i^i and w_{i+nqp}^{i+nqp} for some n. Due to Lemma 3.2,

they both are infinite special branches in $A(w)$ and thus due to Claim 4.1 are equal to w or w'.

Their jth symbols are equal respectively to $w_{i+(j-1)i} = w_{ij}$ and $w_{j(i+nqp)} = w_{ij+(jn)qp}$. We see that their numbers are congruent to $ij \equiv k$ modulo qp, so, they are symbols of w_p^k whose positions in it are congruent modulo q. They are equal since w_p^k is q-periodic. So, w_i^i and w_{i+nqp}^{i+nqp} cannot be inverse and thus are equal; in particular, $w_i = w_{i+nqp}$. The claim is proved. □

Claim 4.7. The minimal period of each of the infinite words w_p^i, $i \in \{1, \dots, p-1\}$ (denoted by q) is the same.

Proof. We know from Claim 4.6 that not only one but all subsequences w_p^k, $k \in \{1, \dots, p-1\}$, are periodic. So, we can repeat the arguments of the proof of Claim 4.6 starting from each of them. In particular, we can choose the position k so that the minimal period q of w_p^k is the least. But repeating the arguments of Claim 4.6, we see that all subsequences w_p^i, $i \in \{1, \dots, p-1\}$, are also q-periodic. □

Claim 4.8. Under notations of claim 4.6, the minimal period q is a power of p.

Proof. Suppose by contrary that q is divided by some $d > 1$ such that p does not divide d. The word w_d^d must be equal to w or w', but the minimal period of its periodic arithmetical subsequence $(w_d^d)_p^1 = w_{pd}^d = (w_p^d)_d^1$ divides q/d, contradicting Claim 4.7. □

Claim 4.9. Under notations of Claim 4.6, if $p > 2$, then the minimal period q is equal to 1.

Proof. Suppose the contrary, then due to Claim 4.8 we have $q = p^m$, where $m > 0$. Recall that the first symbol of w (and thus of w_p^1) is equal to 1. Claim 4.7 says that q is the minimal period of w_p^1, so, for some $i \in \{1, \dots, q-1\}$ we have $w_{ip+1+np^{m+1}} = -1$ for all n. Due to Corollary 4.3, it is possible only if $x^2 \not\equiv ip+1 \pmod{p^{m+1}}$ for all x. But by a classical result of number theory [5] this is not the case since the Legendre symbol is $\left(\dfrac{ip+1}{p}\right) = \left(\dfrac{1}{p}\right) = 1$. □

Proof of Theorem 3.5. Let p be the prime difference of a periodic arithmetical subsequence from Claim 4.5. If $p > 2$, then due to Claims 4.6 and 4.9 we have $w = T_{u?}(w)$ or $w = T_{u?}(w')$ for some word $u = w_1 \cdots w_{p-1}$ of length $p-1$; by our assertion, $w_1 = 1$. Let us consider a primitive root $\overline{r}$ modulo p (it always exists). Cases 1 and 2 of the statement of the theorem correspond respectively to $w_r = 1$ and $w_r = -1$.

Indeed, if $w_r = 1$, then due to Claim 4.2 we have $w_{r^2} = w_{r^3} = \cdots = w_{r^{p-1}} = 1$. Since $\overline{r}$ is a primitive root modulo p, for each $j \in \{1, \dots, p-1\}$ we have $j \equiv r^i \pmod{p}$ for some i, and thus $w_j = 1$. So, $u = 1^{p-1}$. The equation $w = T_{1^{p-1}?}(w)$ gives the word 1^ω which is periodic, and $w = T_{1^{p-1}?}(w')$ corresponds to Case 1 (for $p \geq 3$).

Analogously, if $w_r = -1$, then due to Claim 4.2 we have $w_{r^2} = 1$, $w_{r^3} = -1$ etc., so, $w_{r^k} = 1$ for even k and $w_{r^k} = -1$ for odd k. But $j \equiv r^{2k}$ for some k if and only if j is a quadratic residue modulo p. We see that the word u is equal to the Legendre sequence, corresponding to Case 2.

Now let $p = 2$; due to Claims 4.6 and 4.8, then $w = T_P(w)$ or $w = T_P(w')$, where $P = w_1?w_3?\cdots w_{2^m-1}?$ for some m. Here m is not necessarily minimal, because P in the equation can always be substituted with any of its powers; in particular, we can assume that $m \geq 3$. It is known [5] that each odd number j is congruent modulo 2^m to $(-1)^a 5^b$ for some $a \in \{0, 1\}$ and $b \in \{0, \ldots, 2^{m-2} - 1\}$. So, due to Claim 4.2, the pattern P is uniquely determined by w_5 and w_{2^m-1}. Each of them can be equal to 1 or -1, which gives the four patterns, namely,

- If $w_5 = w_{2^m-1} = 1$, then $P = (1?)^{2^m}$;
- If $w_5 = 1$ and $w_{2^m-1} = -1$, then $P = (1?(-1)?)^{2^{m-1}}$;
- If $w_5 = -1$ and $w_{2^m-1} = 1$, then $P = (1?(-1)?(-1)?1?)^{2^{m-2}}$;
- If $w_5 = w_{2^m-1} = -1$, then $P = (1?1?(-1)?(-1)?)^{2^{m-2}}$.

It is not difficult to see that the exponents 2^m, 2^{m-1}, 2^{m-2} can be omitted and the word w is in fact generated by a pattern P' of length 2, 4 or 8 respectively.

In the first situation we have $P' = 1?$. The equation $w = T_{P'}(w)$ gives 1^ω, which is periodic, and $w = T_{P'}(w')$ gives the *period doubling word* 1(–1)111(–1)1 (–1)1(–1)111(–1)$\cdots$ (see, *e.g.* [9]), completing Case 1 by the sequence for $p = 2$.

In the second situation, $P' = 1?(-1)?$, and we obtain two *paperfolding words* [1] corresponding to Case 3. The remaining two patterns 1?(–1)?(–1)?1? and 1?1? (–1)?(–1)? correspond to Case 4.

We have proved that the only uniformly recurrent sequences which *can* have only two infinite special branches in the arithmetical closure are listed in Theorem 3.5. In the next section, we prove that they *do* have only two infinite special branches.

5. End of the proof of Theorem 3.5

In this section, we study the sequences corresponding to cases 1–4. We prove that they satisfy the conditions of the theorem, *i.e.*, that they are uniformly recurrent and infinite special branches in their arithmetical closures are indeed only 2.

First, we mention that all Toeplitz words generated by one pattern are uniformly recurrent as it is discussed in [8].

Lemma 5.1. *Let w be one of the sequences listed in Theorem 3.5. If some of its arithmetical subsequences w_d^k is not periodic, then $F(w_d^k) = F(w)$ or $F(w_d^k) = F(w')$.*

Proof. First, let us consider together the cases 1 and 2. Clearly, for each of the sequences $w(p)$, $w_L(p)$ and $w_{L'}(p)$, we have

$$w_i = w_{i+p^m} \text{ for } p^m \not| i. \tag{1}$$

Besides, let $(d,p)=1$. If $w=w(p)$, for all i we have

$$w_i = w_{di}. \tag{2}$$

Analogously, if $w=w_L(p)$ or $w=w_{L'}(p)$, then for all i we have

$$w_i = \left(\frac{d}{p}\right) w_{di}. \tag{3}$$

Note that it is sufficient to consider the case of w_d^k with d coprime with p. Indeed, suppose that $p|d$ (say, $d=d'p$). If $p \not| k$, then w_d^k is periodic (and equal to 1^ω or $(-1)^\omega$). If $p|k$ (say, $k=k'p$), then w_d^k is an arithmetical subsequence of w_p^p. In turn, w_p^p is equal to w' if $w=w(p)$ or $w=w_{L'}(p)$; it is equal to w if $w=w_L(p)$. These cases correspond to $w_d^k=(w_{d'}^{k'})'$ or $w_d^k=w_{d'}^{k'}$, so, to prove that $F(w_d^k)=F(w)$ or $F(w_d^k)=F(w')$ it is sufficient to prove the same for $w_{d'}^{k'}$.

So, suppose that $(p,d)=1$ and consider a subsequence $u=w_d^k$ for an arbitrary k. Note that for each $m>0$, we can find n_m such that $k+(n_m-1)d \equiv 0 \pmod{p^m}$. Let us consider the word $u(m)=u_{n_m+1}u_{n_m+2}\cdots u_{n_m+p^m-1}$. For each $i=1,\ldots,p^m-1$, the ith symbol of $u(m)$ is $u_{n_m+i}=w_{k+(n_m+i-1)d}=w_{id}$, where the latter equality is due to (1) and the definition of n_m. Then, by (2) or (3) we obtain $u_{n_m+i}=w_{id}=w_i$ if $w=w(p)$ or $u_{n_m+i}=\left(\frac{d}{p}\right)w_i$ if $w=w_L(p)$ or $w=w_{L'}(p)$. So, $u(m)$ is equal to the prefix of length p^m-1 of w or w'. Due to Lemma 1.2, w_d^k is uniformly recurrent; its factors $u(m)$ of increasing lengths coincide with factors of w or w'. So, due to Lemma 1.3, $F(w_d^k)=F(w)$ or $F(w_d^k)=F(w')$, which was to be proved.

The cases 3 and 4 can be considered analogously. For all of the sequences corresponding to them and for all $m>2$ we have

$$w_i = w_{i+2^m} \text{ if } 2^{m-2} \not| i \tag{4}$$

for all m; the equalities analogous to (2) or (3) differ for each of the sequences. For example, for $w=T_{1?1?(-1)?(-1)?}(w)$ we for all i have

$$w_{2i}=w_i, w_{3i}=w_i, w_{5i}=-w_i, w_{7i}=-w_i. \tag{5}$$

Consider an arithmetical subsequence $u=w_d^k$ of w. Like in the previous cases, we prove that it is sufficient to consider odd differences d, because a non-periodic subsequence of even difference is equal or inverse to a subsequence of twice smaller difference. So, let d be odd; then for each $m>0$ there exists a number n_m such that $k+(n_m-1)d \equiv 0 \pmod{2^m}$, so that u_{n_m} is a symbol of mth order in w. Consider the subword $u(m)=u_{n_m+1}u_{n_m+2}\cdots u_{n_m+2^{m-2}-1}$ of u; for all $i=1,\ldots,2^{m-2}-1$ we have $u_{n_m+i}=w_{k+(n_m+i-1)d}=w_{id}$ because of (4) and the definition of n_m (note that $2^{m-2} \not| id$ because $2^{m-2} \not| i$, $2 \not| d$). Suppose that $i=2^h(2l+1)$ (here $h<m-2$). Then, for each of the sequences considered, we

have 4 cases corresponding to the remainders 1, 3, 5, 7 modulo 8. For example, for $w = T_{1?1?(-1)?(-1)?}(w)$ and $d \equiv 5 \pmod 8$ (say, $d = 8j + 5$ for an integer j) we have $u_{n_m+i} = w_{id} = w_{2^h(2l+1)(8j+5)} = w_{2^{h+3}j(2l+1)+2^h5(2l+1)} = w_{2^h5(2l+1)}$, where the latter equality holds because of (4). But $2^h5(2l+1) = 5i$, so, $u_{n_m+i} = w_{5i} = -w_i$ because of (5). This means that $u(m)$ is the prefix of w' of length $2^{m-2} - 1$, and since u and w' are uniformly recurrent, $F(u) = F(w')$. All other sequences and remainders modulo 8 can be considered analogously: we always obtain $F(u) = F(w)$ or $F(u) = F(w')$. □

To complete the proof of the theorem, it remains to mention that each of the languages $F(w)$ (and symmetrically $F(w')$) has only one infinite special branch. We shall prove it for a larger family of Toeplitz words $w = T_P(w)$ generated by regular patterns. All sequences from the statement of the theorem belong to this class.

Lemma 5.2. *Let P be a regular pattern, $P = u_1?u_2? \cdots u_q?$, where $|u_1| = |u_2| = \cdots = |u_q| = h - 1$, and $q = h^m$ for some integer $m \geq 0$. Let $w = T_P(w)$ be a non-periodic Toeplitz word generated by P. Then $F(w)$ has exactly one infinite special branch coinciding with w.*

Proof. Let u be a factor of w. Note that only one of its h arithmetical subwords u_h^1, u_h^2, …, u_h^h of difference h (namely, the one consisting of symbols of 1st order in w) can be not q-periodic. Moreover, if u is sufficiently long (say, $|u| \geq N$), one of these words *is* not q-periodic because $w_h^h = w$ is uniformly recurrent and non-periodic. Now suppose that a factor u of w of length at least N is special. The considerations above must hold for both $1u$ and $-1u$, so, this is u_h^h which is not q-periodic. So, all occurrences of u in w start at positions equal to 1 modulo h. If the length of u is not less than hN, then u_h^h is also a special factor of $w = w_h^h$ of length at least N, so, $(u_h^h)_h^h = u_{h^2}^{h^2}$ is not q-periodic and consists of symbols of w having 2nd order. So, all occurrences of u in w start at positions equal to 1 modulo h^2. Continuing these arguments, we see that for all $k \geq 0$ a special factor of w of length not less than $h^{m+k}N$ always occurs in w starting with positions equal to 1 modulo h^{m+k+1}. Since $q = h^m$, we see that the prefix of length $h^{k+1} - 1$ of each special factor of w having length at least $h^{m+k}N$ consists of symbols of order less than $k + 1$ and coincides with the prefix of w of length $h^{k+1} - 1$.

So, a sequence of words special in $F(w)$ can converge only to w itself, which is the unique infinite special branch required. □

Lemma 5.2 is valid for all sequences listed in Theorem 3.5. It can be applied directly to the sequences defined by $w = T_P(w)$; for those defined by $w = T_P(w')$, we have $w = T_{T_P(P')}(w)$, where the pattern $T_P(P')$ in all cases satisfies the conditions of the lemma.

So, the infinite special branches of the arithmetical closure of each of the listed sequences are those of $F(w)$ and $F(w')$. They are two, and the theorem is proved.

6. Lowest arithmetical complexity

Now our goal is to investigate the arithmetical complexity of the listed sequences and to find the lowest one. We have proved in the previous section that if w is one of the sequences listed in Theorem 3.5, then

$$A(w) = F(w) \cup F(w') \cup P,$$

where P is the set of factors of periodic arithmetical subsequences of w. Due to Lemma 5.2 the set of factors of w and w' contain only one infinite special branch each, whereas the infinite special branches of $A(w)$ are two. So, $F(w) \neq F(w')$. Since w and w' are uniformly recurrent, and due to Lemma 1.3, the set $F(w) \cap F(w')$ is finite. In its turn, P is a union of the languages of factors of a finite number of periodic sequences, each of which is uniformly recurrent; so, $F(w) \cap P$ and $F(w') \cap P$ are also finite. The set P contains an ultimately constant number of words of each length, we denote it by c. Since w and w' have the same subword complexity, we see that for sufficiently large n

$$a_w(n) = 2f_w(n) + c.$$

Now let us discuss the subword and thus arithmetical complexities of each of the cases 1–4. Note that computing the subword complexity of each individual word from this list is not a problem due to the techniques described in [8] or [17].

The family of words from Case 1 can be uniformly treated: for each prime p we have

$$f_{w(p)}(n) = \begin{cases} n + p^a & \text{for } 2p^a - p^{a-1} < n \le p^{a+1}, \\ 2n - p^{a+1} + p^a & \text{for } p^{a+1} < n \le 2p^{a+1} - p^a \end{cases}$$

for all $a \ge 1$. Since the longest power of 1 occurring in $F(w)$ is 1^{2p-1}, and the periodic infinite words adding elements to $A(w)$ are 1^ω and $(-1)^\omega$, we have $a_{w(p)}(n) = 2f_{w(p)}(n) + 2$ for all $n \ge 2p$. So, as a whole we have

$$a_{w(p)}(n) = \begin{cases} 2n & \text{for } n \le 2, \\ 2n + 2 & \text{for } 2 < n \le p, \\ 4n - 2p + 2 & \text{for } p \le n \le 2p - 1, \\ 2n + 2p^a + 2 & \text{for } 2p^a - p^{a-1} < n \le p^{a+1}, a \ge 1, \\ 4n - 2p^{a+1} + 2p^a + 2 & \text{for } p^{a+1} < n \le 2p^{a+1} - p^a, a \ge 1. \end{cases}$$

In particular, $(a_{w(p)}(n) - 2)/n = 2$ for $2 < n \le p$, and for $n \ge 2p$ we have

$$\frac{2p+2}{p} \le \frac{a_{w(p)}(n) - 2}{n} \le \frac{6p-2}{2p-1};$$

here both limits are attained at an infinite number of ns, respectively $n = p^{a+1}$ and $n = 2p^{a+1} - p^a$, where a is a positive integer.

In Case 2, we similarly have $a_w(n) = 2f_w(n) + 2$ for both $w_L(p)$ and $w_{L'}(p)$ for all p and for all sufficiently large n (in particular, for all $n \geq p$). It can be proved also that for sufficiently large n it holds $f_{w_L(p)}(n) = f_{w_{L'}(p)}(n) > f_{w(p)}(n)$.

Conjecture 6.1. It seems that $f_{w_L(p)}(n) = f_{w_{L'}(p)}(n) \geq 2n$ for all p and n. Here the equality is attained at an infinite number of points of the form p^a.

In Case 3, we have $f_w(n) = 4n$ for all $n \geq 7$ [1] and $a_w(n) = 8n+4$ for all $n \geq 14$. In Case 4, we by similar technique see that $f_w(n) = 8n$ for all sufficiently large n. So, the arithmetical complexity of these words cannot pretend to be minimal.

Thus, we see that we are not able to write down a uniformly recurrent non-periodic words of "minimal" arithmetical complexity function. Words whose sets of factors coincide with those of $w(p)$, $p \to \infty$, constitute a family of words having decreasing upper and lower limits of arithmetical complexity, tending to $3n$ and $2n$ respectively. It is interesting to mention that although these words are not Sturmian and are obtained by a completely different construction, their maximal pattern complexity is minimal [16].

7. Conclusion

Denote by $\mathcal{R}$ is the set of all non-periodic uniformly recurrent infinite words. From the arguments above, we can conclude the following:

- $\inf\limits_{w\in\mathcal{R}} \overline{\lim\limits_{n\to\infty}} \dfrac{a_w(n)}{n} = \lim\limits_{p\to\infty} \overline{\lim\limits_{n\to\infty}} \dfrac{a_{w(p)}(n)}{n} = 3$,
  and $\overline{\lim\limits_{n\to\infty}} \dfrac{a_w(n)}{n} \geq 3$ for all $w \in \mathcal{R}$
  (we have not proved the strict inequality here because the case of 3 infinite special branches is to be considered for it).
- $\inf\limits_{w\in\mathcal{R}} \underline{\lim\limits_{n\to\infty}} \dfrac{a_w(n)}{n} = \lim\limits_{p\to\infty} \underline{\lim\limits_{n\to\infty}} \dfrac{a_{w(p)}(n)}{n} = 2$,
  but $\underline{\lim\limits_{n\to\infty}} \dfrac{a_w(n)}{n} > 2$ for all $w \in \mathcal{R}$.
- Let us define the function $a(n) = \min\limits_{w\in\mathcal{R}} a_w(n)$. Then
  $a(n) = \min\limits_{p \text{ prime}} a_{w(p)}(n) = 2n + 2$ for all $n \geq 2$, since we can always choose $p > n$.
- If $\underline{\lim\limits_{n\to\infty}} \dfrac{a_w(n)}{n} < 3$ for some $w \in \mathcal{R}$, then $F(w) = F(w(p))$, $F(w) = F(w_L(p))$, or $F(w) = F(w_{L'}(p))$ for some prime $p \geq 3$.

If Conjecture 6.1 holds, the latter statement can be strengthened to:

- If $\underline{\lim\limits_{n\to\infty}} \dfrac{a_w(n)}{n} < 3$ for some $w \in \mathcal{R}$, then $F(w) = F(w(p))$.

At last, we would like to emphasize that all obtained results are valid only for uniformly recurrent words. Not uniformly recurrent words of lower arithmetical complexity may exist, although we cannot predict their possible form.

References

[1] J.-P. Allouche, The number of factors in a paperfolding sequence. *Bull. Austral. Math. Soc.* **46** (1992) 23–32.

[2] J.-P. Allouche, M. Baake, J. Cassaigne, D. Damanik, Palindrome complexity. *Theoret. Comput. Sci.* **292** (2003) 9–31.

[3] S.V. Avgustinovich, D.G. Fon-Der-Flaass, A.E. Frid, *Arithmetical complexity of infinite words, in Words*, Languages & Combinatorics III, edited by M. Ito and T. Imaoka, Singapore, World Scientific Publishing. *ICWLC 2000, Kyoto, Japan, March 14–18* (2003) 51–62.

[4] J. Berstel, P. Séébold, Sturmian words, in *Algebraic combinatorics on words*, edited by M. Lothaire, Cambridge University Press (2002).

[5] A.A. Bukhshtab, *Number Theory*, Uchpedgiz, Moscow (1960) (in Russian).

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

[7] J. Cassaigne, A. Frid, On arithmetical complexity of Sturmian words, in *Proc. WORDS 2005*, Montreal (2005) 197–208.

[8] J. Cassaigne, J. Karhumäki, Toeplitz words, generalized periodicity and periodically iterated morphisms. *European J. Combin.* **18** (1997) 497–510.

[9] D. Damanik, Local symmetries in the period doubling sequence. *Discrete Appl. Math.* **100** (2000) 115–121.

[10] S. Ferenczi, Complexity of sequences and dynamical systems. *Discrete Math.* **206** (1999) 145–154.

[11] A. Frid, *A lower bound for the arithmetical complexity of Sturmian words*, Siberian Electronic Mathematical Reports 2, 14–22 [Russian, English abstract].

[12] A. Frid, Arithmetical complexity of symmetric D0L words, *Theoret. Comput. Sci.* **306** (2003) 535–542.

[13] A. Frid, On Possible Growth of Arithmetical Complexity. *Theor. Inform. Appl.* accepted

[14] A. Frid, Sequences of linear arithmetical complexity. *Theoret. Comput. Sci.* **339** (2005) 68–87.

[15] J. Justin, G. Pirillo, Decimations and Sturmian words. *Theor. Inform. Appl.* **31** (1997) 271–290.

[16] T. Kamae, L. Zamboni, Maximal pattern complexity for discrete systems. *Ergodic Theory Dynam. Systems* **22** (2002) 1201–1214.

[17] M. Koskas, Complexités de suites de Toeplitz. *Discrete Math.* **183** (1998) 161–183.

[18] I. Nakashima, J. Tamura, S. Yasutomi, I. Nakashima, J.-I. Tamura, S.-I. Yasutomi, *-Sturmian words and complexity. *J. Théor. Nombres Bordeaux* **15** (2003) 767–804.

[19] A. Restivo, S. Salemi, Binary patterns in infinite binary words, in Formal and Natural Computing, Springer Berlin, edited by W. Brauer et al. *Lect. Notes Comput. Sci.* **2300**, (2002) 107–116.

[20] E. Szemerédi, On sets of integers containing no k elements in arithmetic progression. *Acta Arith.* **27** (1975) 199–245.

Communicated by le nom de l'éditeur.
Received October 1, 2003. Accepted October 30, 2003.