ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Theoretical Computer Science 339 (2005) 68–87

Theoretical Computer Science

www.elsevier.com/locate/tcs

Sequences of linear arithmetical complexity☆

A.E. Frid*

Theoretical Computer Science Department, Sobolev Institute of Mathematics, SB RAS Koptyug av., 4, 630090 Novosibirsk, Russia

Abstract

Arithmetical complexity of infinite sequences is the number of all words of a given length whose symbols occur in the sequence at positions which constitute arithmetical progressions. We show that uniformly recurrent sequences whose arithmetical complexity grows linearly are precisely Toeplitz words of a specific form.

Keywords: Arithmetical complexity; Infinite word; Subword complexity; Toeplitz word; Uniformly recurrent word; Special words; *S*-adic conjecture

1. Introduction

Subword complexity $f_w(n)$ of an infinite word w is a classical function defined in 1975 [7] as the number of factors of w of length n. Later, several modifications of this notion have been introduced. Most of them are functions counting factors of the infinite word *and* some other words of a given length reflecting structure of the word, i.e., functions which are not less than subword complexity. These are *d-complexity* introduced in 1987 by Iványi [11], *pattern complexity* introduced in 2002 by Restivo and Salemi [14], *maximal pattern complexity* by Kamae and Zamboni [12] which is also dated 2002. *Arithmetical complexity*, defined by Avgustinovich, Fon-Der-Flaass and the author in 2000 [3], also belongs to this family. It counts words from the *arithmetical closure* of an infinite word w, i.e., words built by symbols whose numbers in w constitute arithmetical progressions.

☆ Supported in part by RFBR Grants 02-01-00939 and 03-01-00796 and by Russian Science Support Foundation.

* Tel.: +7 3832 332594; fax: +7 3832 332598.
E-mail address: frid@math.nsc.ru.

0304-3975/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2005.01.009

The research of arithmetical closure was inspired by the famous Van der Waerden theorem which states that the arithmetical closure of an infinite word always contains a word a^n for an arbitrary power n and some symbol a. In [3], it was shown that if subword complexity grows linearly, then arithmetical complexity can grow both linearly and exponentially. Standard questions to be answered arise: what is arithmetical complexity of known classes of infinite words? What is the lowest possible complexity? What are possible growth rates of arithmetical complexity? Which words have linear arithmetical complexity? Partial results on the first of these questions were obtained in [3,9]. The second question was investigated in [2] for the case of uniformly recurrent words. A family of words with various sub-polynomial growths of arithmetical complexity was constructed in [10], making a contribution to the third problem. This paper is devoted to the answer to the fourth question for the important case of uniformly recurrent words: we characterize uniformly recurrent words whose arithmetical complexity grows linearly. Up to the set of factors, they are exactly Toeplitz words [5,13] of a special form.

Note that sequences of linear subword complexity are not yet classified, and their characterization is an important unsolved problem [8]. The existing but never clearly stated *S-adic conjecture* offers to somehow describe such sequences as generated by a finite number of substitution-like mappings. Our characterization of sequences of linear arithmetical complexity is of the same kind since it involves a finite number of Toeplitz transforms which generate each of such sequences.

Sequences of linear pattern or maximal pattern complexity also are not yet classified. It seems that the question about arithmetical complexity is easiest in the family, although the case of non-recurrent words is still open.

The paper is organized as follows. Main notions and statement (Theorem 1) are given in Section 2 and discussed in Section 4 where the notion of Toeplitz words is defined. Lemma 5 in Section 4 demonstrates several equivalent conditions each of those could be used for the statement of Theorem 1. The "if" part of the proof of Theorem 1 is contained in Section 6 which is relatively independent from others. The technique of special infinite words used for the "only if" proof is introduced in Section 7; in the end of that section, a short sketch of the "only if" proof is given. Main part of the "only if" proof is given in Sections 9 and 10. All other sections contain auxiliary notions and statements.

2. Main definitions and theorem

In what follows we consider right infinite words on a finite alphabet Σ; the set of such words is, as usual, denoted by Σ^ω. The terms "infinite word" and "sequence" are used below as synonyms.

The set of factors of a word w is denoted by $F(w)$. Let w_i denote the ith symbol of an infinite word w: $w = w_1 w_2 \cdots w_n \cdots$. An infinite word $w_d^k = w_k w_{k+d} w_{k+2d} \cdots w_{k+nd} \cdots$, where $d, k > 0$, is called an *arithmetical subsequence* of w, and d is called its *difference*. In this paper we consider only arithmetical subsequences and sometimes omit the word "arithmetical" before "subsequences".

A factor of an arithmetical subsequence of w is called an *arithmetical factor* of w, and the set of arithmetical factors of w is its *arithmetical closure*, denoted by $A(w)$:

$$A(w) = \bigcup_{d,k>0} F(w_d^k).$$

The number of words of length n in the arithmetical closure of w, denoted by $a_w(n)$, is called *arithmetical complexity* of w. Clearly, the arithmetical complexity of a word is greater than or equal to its *subword complexity* $f_w(n)$, which is the number of factors of w of length n.

An infinite word of the form $uu\cdots u\cdots = u^\omega$ is called $|u|$-*periodic*; the word u is called a *prefix period* of u^ω.

The *orbit* $\mathcal{O}(w)$ of an infinite word w is the set of infinite words whose set of factors is included in $F(w)$. A word w is called *uniformly recurrent* if each of its factors occurs in it infinitely many times with bounded gaps, or, equivalently, if $\mathcal{O}(w)$ coincides with the set of words having the same set of factors that w. Since arithmetical complexity is a function of set of factors, it is the same for all words from the orbit of a uniformly recurrent word.

Let us say that a sequence u is *canonically p-regular* if for all $k > 0$, $i \in \{1, \ldots, p^k - 1\}$, the sequence $u^i_{p^k}$ is periodic.

Example 1. Let us define the function $u(i)$ as the largest exponent of 2 dividing i, modulo 2. The *period doubling word* $u_{\mathrm{pd}} = u(1)u(2)\cdots u(n)\cdots = 01000101010001000100\cdots$ [6] is canonically 2-regular since for each $k > 0$ and $i \in \{1, \ldots, 2^k - 1\}$ we have $(u_{\mathrm{pd}})^i_{2^k} = (u(i))^\omega$.

Theorem 1. *A non-periodic uniformly recurrent infinite word has linearly growing arithmetical complexity if and only if it belongs to the orbit of some canonically p-regular word w, where p is prime and $w^{p^a}_{p^a} = w^{p^b}_{p^b}$ for some $a \neq b$.*

Example 2. For the period doubling word $u = u_{\mathrm{pd}}$, we have $u = u^1_1 = u^4_4$. So, arithmetical complexity of any word from its orbit is linear (and does not depend on the choice of the word); in fact, it lies between $3n + 2$ and $10n/3 + 2$ for all $n \geqslant 4$ [2].

The condition that p is prime is crucial: in Section 7 we shall give an example of a word with non-linear arithmetical complexity which fits all conditions of the theorem except that $p = 6$.

Our technique of the proof of Theorem 1 cannot be generalized to words which are not uniformly recurrent. On the other hand, we do not know a non-trivial example of a non-uniformly recurrent word of linear arithmetical complexity. (There is a family of trivial examples of the form vw, where w is a uniformly recurrent infinite word and v is an arbitrary finite prefix.) This allows us to state the following

Conjecture 1. *If an infinite word is not uniformly recurrent but has linear arithmetical complexity, then it is obtained from a uniformly recurrent word by adding a finite prefix.*

3. Properties of arithmetical subsequences

In this section, we give some more technical notations concerning arithmetical subsequences and state several easy, folklore, or classical results which will be useful below.

Recall that w_d^k denotes the arithmetical subsequence of w having difference d and starting from the symbol of w numbered k. We shall often use the following equality which holds for all a, b, c, and d:

$$(w_b^a)_d^c = w_{bd}^{a+(c-1)b}.$$

In particular, this gives

$$(w_a^a)_d^c = w_{ad}^{ac}.$$

The *shuffle* of sequences $a, b, \ldots, x \in \Sigma^\omega$, denoted by $\sqcup$, is the sequence consisting of alternated symbols of $a, b, \ldots, x$, i.e.,

$$a \sqcup b \sqcup \cdots \sqcup x = a_1 b_1 \cdots x_1\ a_2 b_2 \cdots x_2\ a_3 b_3 \cdots x_3 \cdots.$$

In particular, we by definitions have for all d that

$$w = w_d^1 \sqcup w_d^2 \sqcup \cdots \sqcup w_d^d.$$

The following lemmas are obvious and are stated here just to simplify reading of the text below:

Lemma 1. *Each arithmetical subsequence of a periodic sequence is periodic.*

Lemma 2. *The shuffle of several sequences is periodic if and only if all shuffled sequences are periodic.*

The next lemma is also easy.

Lemma 3. *Let u be an arithmetical subsequence of an infinite word v. Then each $v' \in \mathcal{O}(v)$ contains an arithmetical subsequence u' of the same difference such that $u' \in \mathcal{O}(u)$.*

Proof. Let us suppose that $u = v_d^k$ and color all symbols of v on positions $k + id$, $i \geqslant 0$, red. Then each finite word consisting of successive red symbols is a factor of u. For all $n > 0$, let us consider the prefix $v'(n)$ of v' of length n. Since $v' \in \mathcal{O}(v)$, $v'(n)$ occurs somewhere in v, and if $n \geqslant K = \max(k, d)$, then this occurrence contains red symbols which constitute an arithmetic progression of difference d starting from a symbol numbered k_n and going to the end of $v'(n)$; the word constituted by symbols of this progression will be denoted by $u'(n)$. Clearly, $u'(n) \in F(u)$ for all n, and the length of $u'(n)$ tend to infinity with $n \to \infty$. Here k_n can be chosen to be not greater than K. Thus, some number k' will occur in the sequence $\{k_n\}_{n=K}^\infty$ an infinite number of times: suppose that $k_{n_i} = k'$ for all $i = 1, 2, \ldots$. Then for all i, $u'(n_i)$ is a prefix of $u'(n_{i+1})$ and they both are factors of u. Thus, the required word u' is $u' = (v')_d^{k'} = \lim_{i\to\infty} u'(n_i)$. □

The next result is folklore, but for the sake of completeness, its proof is contained in [3].

Lemma 4. *An arithmetical subsequence of a uniformly recurrent word is uniformly recurrent.*

The remaining two results are classical theorems of number theory.

Theorem 2 (*Van der Waerden, 1927*). *The arithmetical closure of each word on a finite alphabet Σ contains arbitrarily large powers of symbols of the form a^n, where n is an arbitrary positive integer and $a \in \Sigma$.*

Theorem 3 (*Dirichlet, 1837*). *Let* $\gcd(l, k) = 1$; *then the arithmetical progression* $l, l + k, \ldots, l + nk, \ldots$ *contains an infinite number of primes.*

4. Discussion of the main result in terms of Toeplitz words

Uniformly recurrent words of linear arithmetical complexity, characterized by Theorem 1, admit several other characterizations in terms of Toeplitz transforms. In this section, we discuss and prove them.

Let ? be a new symbol called *gap*, not belonging to Σ. A finite word on $\Sigma \cup \{?\}$ is called a *pattern*. In what follows patterns on $\Sigma \cup \{?\}$, unlike words on Σ, are denoted by capitals.

Let P be a pattern and $w \in (\Sigma \cup \{?\})^\omega$ be an infinite word. In what follows we denote by $P \cdot w$ the result of substituting the gaps in P^ω by successive symbols of w, starting from the first symbol. If $w = P \cdot w$ for some $w \in \Sigma^\omega$, then w is called the *Toeplitz word* generated by P and denoted by $T(P)$. Clearly, if the first symbol of P is not a gap, then the equation $w = P \cdot w$ has a unique solution.

More generally, let $P_1, P_2, \ldots, P_n, \ldots$ be a sequence of patterns. Consider the sequence $\{U_i\}_{i=0}^\infty$ of infinite words defined by $U_0 = ?^\omega$, $U_i = P_1 \cdot P_2 \cdot \ldots \cdot P_i \cdot ?^\omega$ for all $i > 0$. Clearly, each of the words U_i is periodic, which allows us to define the product of patterns $P_1 \cdot P_2$ as the minimal prefix period of U_2. So, $(\cdot)$ is a non-commutative associative operation on the set of patterns.

If infinitely many of patterns P_i start with a symbol of Σ, then the sequence $\{U_i\}_{i=0}^\infty$ converges to an infinite word on Σ naturally denoted by $P_1 \cdot P_2 \cdot \ldots \cdot P_n \cdot \ldots = \prod_{i=1}^\infty P_i$. It is called the *Toeplitz word* generated by the sequence $\{P_i\}_{i=1}^\infty$; if all P_i are equal to the same pattern P, this word is equal to $T(P)$.

A pattern is called (d-)regular if it belongs to $(\Sigma^{d-1}?)^q$ for some q. The set of all regular (d-regular) patterns is denoted by $\mathcal{P}$ (respectively, $\mathcal{P}_d$). The family of regular patterns from $\mathcal{P}_d$ containing l gaps is denoted by $\mathcal{P}_d^l$, i.e., $\mathcal{P}_d^l = (\Sigma^{d-1}?)^l$.

Clearly, the product of a p-regular and a q-regular patterns is pq-regular.

Example 3. The pattern $P_{\mathrm{pf}} = 0?1?$ is 2-regular. We have $P_{\mathrm{pf}} \cdot ?^\omega = (0?1?)^\omega$, $P_{\mathrm{pf}} \cdot P_{\mathrm{pf}} \cdot ?^\omega = (001?011?)^\omega$, and thus $P_{\mathrm{pf}} \cdot P_{\mathrm{pf}} = 001?011?$, which is 4-regular, etc. As a limit, we obtain the famous *paperfolding word* $u_{\mathrm{pf}} = \prod_{i=1}^\infty P_{\mathrm{pf}} = T(P_{\mathrm{pf}}) = 0010011000110110 \cdots$. It can be checked that it is canonically 2-regular.

Example 4. The period doubling word u_{pd} can be obtained as $T(010?)$, i.e., is a Toeplitz word generated by a 4-regular pattern $010? = 0? \cdot 1?$.

These examples hint that the classes of canonically p-regular words and Toeplitz words generated by regular patterns are close to each other. This is indeed the case, and the explicit relations are given by the following:

Lemma 5. *Let p be a prime number and w be an infinite word. The following conditions are equivalent*:

(1) *w is canonically p-regular and $w_{p^a}^{p^a} = w_{p^b}^{p^b}$ for some $a \neq b$;*
(2) *$w = R_1 \cdot R_2 \cdot \ldots \cdot R_n \cdot \ldots \cdot$, where all patterns R_i are p-regular and the sequence $\{R_i\}_{i=1}^{\infty}$ is ultimately periodic;*
(3) *$w \in \mathcal{P}_{p^m} \cdot T(\mathcal{P}_{p^k})$ for some k and m;*
(4) *$w \in \mathcal{P} \cdot T(\mathcal{P}_{p^k})$ for some k.*

Proof. (1) $\Rightarrow$ (2). If w is canonically p-regular, then the sequence $w' = w_1 \cdots w_{p-1}?w_{p+1} \cdots w_{2p-1}?w_{2p+1} \cdots$ is pL-periodic, where L is the lcm of periods of $w_p^1, \ldots, w_p^{p-1}$. So, if we define R_1 as the prefix of length pL of w', then R_1 is a p-regular pattern, and we have $w = R_1 \cdot w_p^p$. The sequence w_p^p and all of $w_{p^a}^{p^a}$ can be treated analogously, and thus we see that $w = R_1 \cdot R_2 \cdot \ldots \cdot R_n \cdot \ldots \cdot$. In its turn, $w_{p^a}^{p^a} = w_{p^b}^{p^b}$ implies $R_a = R_b$ and $R_{a+t} = R_{b+t}$ for all t, so, the sequence of patterns $\{R_i\}_{i=a}^{\infty}$ is $(b-a)$-periodic, and $\{R_i\}_{i=1}^{\infty}$ is ultimately periodic.

(2) $\Rightarrow$ (3). Using notations of the previous paragraph, we can define $P = R_1 \cdot \ldots \cdot R_{a-1}$ and $Q = R_a \cdot \ldots \cdot R_{b-1}$ to have $w = P \cdot T(Q)$. So, m can be defined as $a-1$ and k as $b-a$.

(3) $\Rightarrow$ (4). This implication is obvious.

(4) $\Rightarrow$ (1). Let $w = R \cdot u$, where $R = w_1 \cdots w_{q-1}?w_{q+1} \cdots w_{2q-1}?w_{(L-1)q+1} \cdots w_{Lq-1}? \in \mathcal{P}_q^L$ is a q-regular pattern for some q and $u \in T(\mathcal{P}_{p^k})$. Then u is clearly p-regular. Let us prove that so is w. Suppose first that $q = lp$ for some l, then for all $i \in \{1, \ldots, p-1\}$ we have $w_p^i = w_q^i \sqcup w_q^{i+p} \sqcup \cdots \sqcup w_q^{i+(l-1)p}$. Each of the shuffled sequences is periodic, and thus w_p^i is periodic. It remains to prove that w_p^p is canonically p-regular, but it is obtained from u by applying an l-regular pattern $R' = w_p \cdots w_{(l-1)p}? \cdots ?w_{(k-1)q+p} \cdots w_{(k-1)q+(l-1)p}?$. If $p|l$, we can continue the process and thus see that the main case is that of p not dividing q (since p is prime, this implies that p and q are coprime).

In this case, for all $i \in \{1, \ldots, p\}$ we have $w_p^i = w_{pq}^i \sqcup w_{pq}^{i+p} \sqcup \cdots \sqcup w_{pq}^{i+(q-1)p}$. Positions $i, i+p, \ldots, i+(q-1)p$ of starting symbols of the shuffled subsequences take all possible values modulo q one time each. So, all of them except one are not equivalent to 0 modulo q; their respective subsequences are periodic. The position equal to 0 modulo q, let us denote it by ql, gives the subsequence $w_{pq}^{ql} = (w_q^q)_p^l = u_p^l$.

If $l < p$, then this subsequence is also periodic since u is canonically p-regular. Thus, due to Lemma 2, so is w_p^i.

If $l = p$, then the subsequence w_{pq}^{lq} is equal to u_p^p which is canonically p-regular since u is canonically p-regular. But $l = p$ means that $i + jp = qp$ for some $i \in \{1, \ldots, p-1\}$

and $j \in \{0, \ldots, q-1\}$, which is possible only when $i = p$ and $j = q-1$. Thus, the subsequences $w^1_p, w^2_p, \ldots, w^{p-1}_p$ are periodic, whereas

$$w^p_p = w^p_{pq} \sqcup w^{2p}_{pq} \sqcup \cdots \sqcup w^{p(q-1)}_{pq} \sqcup u^p_p.$$

Here the first $q-1$ sequences in this shuffle are L-periodic, so, the equality means that $w^p_p = R' \cdot u^p_p$, where $R' \in \mathcal{P}^L_q$ and u^p_p is canonically p-regular. We see that w^p_p falls into the same class than w and can be treated analogously: we can show that $(w^p_p)^i_p = w^{ip}_{p^2}$ are periodic for all $i \in \{1, \ldots, p-1\}$, and $(w^p_p)_p = w^{p^2}_{p^2} = R'' \cdot u^{p^2}_{p^2}$. Continuing the process by induction we see that all subsequences $w^i_{p^k}$, where $i \in \{1, \ldots, p^k - 1\}$, are periodic, and thus w is canonically p-regular.

It remains to prove that $w^{p^a}_{p^a} = w^{p^b}_{p^b}$ for some $a \neq b$. Indeed, at each step when we pass from $w = R \cdot u$ to $w^p_p = R' \cdot u^p_p$, the new regular pattern R' is completely defined by R, and its length qL is the same that the length of R. So, the sequence of such patterns is ultimately periodic with some period r. On the other hand, since $u \in T(\mathcal{P}_{p^k})$, the sequence of sequences $u^{p^a}_{p^a}$ is k-periodic. So, the sequence of sequences $w^{p^a}_{p^a}$ is ultimately lcm (k, r)-periodic, which proves the implication and thus the lemma. □

So, we could state Theorem 1 using any of the equivalent conditions of Lemma 5.

Example 5. Let us consider the word $w = 230230231230230231 \cdots = (23?) \cdot T(0?1?)$. It is canonically 2-regular with $w = w^1_1 = w^4_4$ and is equal to $T(R_1 \cdot R_2) = R_1 \cdot R_2 \cdot R_1 \cdot R_2 \cdot \ldots \cdot$, where $R_1 = 2?0?3?2?1?3?$ and $R_2 = 3?0?2?3?1?2?$.

5. Properties of regular words

Let us say that an infinite word w is *d-regular* for some positive integer d if for each $k > 0$ there exists $i_k \in \{1, \ldots, d^k\}$ such that all subsequences $w^i_{d^k}$ with $i \in \{1, \ldots, d^k\} \setminus \{i_k\}$ are periodic. [1] Symbols occurring in w at positions congruent to i_k modulo d^k are called kth *order* symbols of w. In particular, all symbols of w are of order 0. The *maximal order* of a symbol is defined naturally and can be finite or infinite.

By definitions, a word is canonically d-regular if and only if it is d-regular with $i_k = d^k$ for all k. Symbols of kth order in a canonically d-regular word $w = \prod_{i=1}^{\infty} P_i$, where P_i are d-regular patterns, are exactly the symbols substituted from gaps not earlier than in $W_k = P_1 \cdot \ldots \cdot P_k \cdot ?^\omega$.

In this section, we state some results on d-regular words which will be needed later. The first several lemmas are easy.

Lemma 6. *Each sequence from the orbit of a d-regular infinite word is d-regular.*

[1] This definition does not coincide with that of k-regular sequences by Allouche and Shallit [1].

Lemma 7. *The word obtained from a* (*canonically*) *d-regular word by applying a symbol-to-symbol morphism is* (*canonically*) *d-regular.*

Lemma 8. *The maximal order of all symbols except perhaps one in a non-periodic d-regular word is finite.*

Lemma 9. *Let w be an infinite word. If the subsequence w_k^k is canonically p-regular for some k and a prime p, then so are $(w_m^m)_k^k$ for all m.*

The next lemma is also easy.

Lemma 10. *Consider $w = T(R)$, where $R \in \mathcal{P}^q_{p^k}$, and some $d \equiv 1 \pmod{qp^k}$. Then $w_d^d = w$.*

Proof. Let us denote by $\overline{n} \in \{0, \dots, qp^k - 1\}$ the residue of a number n modulo qp^k; then $\overline{d} = 1$. Note also that for all m we have $w_{mp^k} = w_m$, and if $m \not\equiv 0 \pmod{p^k}$, then $w_m = w_{\overline{m}}$. Now let us fix an arbitrary integer $i > 0$ and define l as the maximal integer such that $i = i'p^{lk}$ for some i'. Then

$$(w_d^d)_i = w_{di} = w_{di'p^{lk}} = w_{di'} = w_{\overline{di'}} = w_{\overline{\overline{d}\,\overline{i'}}} = w_{\overline{i'}} = w_{i'} = w_{i'p^{lk}} = w_i.$$

So, for all i the ith symbols of w_d^d and of w are equal. □

The next lemma gives us not all information on an arithmetical subsequence of a Toeplitz word: we could prove more, but this is what we shall need in the end of proof of the main theorem.

Lemma 11. *If a word $w \in \mathcal{O}(T(\mathcal{P}^q_{p^k}))$ is non-periodic, then any its arithmetical subsequence w_d^b with $\gcd(d, pq) = 1$ is also non-periodic and p^k-regular. The maximal order of each of its symbols in w_d^b is equal to its maximal order in w.*

Proof. Let us choose some $w' \in T(\mathcal{P}^q_{p^k})$ such that $w \in \mathcal{O}(w')$. Clearly, such w' exists and is not periodic. Since w and w' are p^k-regular, for all n we can uniquely find in w and w' non-periodic subsequences of difference p^{kn} starting not later than at the position p^{kn}; moreover, in w' this is $(w')_{p^{kn}}^{p^{kn}}$. Let us substitute these non-periodic subsequences of symbols of nth order by gaps. The obtained infinite words $W(n)$ and $W'(n)$ are qp^{kn}-periodic, and their sets of factors coincide. So, there exists some j_n such that all shifts of $W(n)$ by $j_n + lqp^{nk}$ symbols, $l \geqslant 0$, are equal to $W'(n)$, i.e., $(W(n))_1^{j_n+lqp^{kn}+1} = W'(n)$. Let us choose some l_n so that $w_{j_n+l_nqp^{kn}}$ is a symbol of w_d^b: this is possible since $\gcd(d, qp^{nk}) = 1$. Note that $w_{j_n+l_nqp^{kn}}$ is of nth order in w, i.e., $W(n)_{j_n+l_nqp^{kn}}$ is a gap. Let us denote $j_n + l_nqp^{kn} = B_n = b + (m_n + 1)d$.

Let us fix an $m \not\equiv m_n \pmod{p^{kn}}$ and show that $(w_d^b)_{p^{kn}}^m$ is periodic. Indeed, $(w_d^b)_{p^{kn}}^m = w_{dp^{kn}}^{b+(m-1)d} = (w_{p^{kn}}^{b+(m-1)d})_d^1$, the word $w_{p^{kn}}^{b+(m-1)d}$ is periodic since $b + (m-1)d \not\equiv$

$B_n \pmod{p^{kn}}$, and thus $(w_d^b)^m_{p^{kn}}$ is periodic due to Lemma 1. Since n and $m \not\equiv m_n \pmod{p^{kn}}$ were chosen arbitrarily, we have proved that w_d^b is p^k-regular and its symbols of maximal order less than n are of order less than n in w. Since this is true for all n, we see that the maximal order of a symbol in w_d^b is equal to its maximal order in w.

Moreover, for all $c \in \{1, \ldots, p^{kn}-1\}$ we have $(w_d^b)_{m_n+c} = w_{b+(m_n+c-1)d} = w_{B_n+cd} = (W(n))_{B_n+cd} = (W'(n))_{cd} = (w')_{cd}$. Thus, $(w_d^b)_{m_n+1}\,(w_d^b)_{m_n+2}\cdots(w_d^b)_{m_n+p^{kn}-1}$ coincides with the prefix of $(w')_d^d$ of length $p^{kn}-1$. Since the choice of w' does not depend on d and b, and w_d^b and $(w')_d^d$ are uniformly recurrent due to Lemma 4, this implies $F(w_d^b) = F((w')_d^d)$ for all b and d: we see that the language of factors of w_d^b does not depend on b.

It remains to prove that w_d^b is not periodic. Suppose the opposite: let w_d^b is periodic. Since $\gcd(d, pq) = 1$, there exists c such that $dc \equiv 1 \pmod{qp^k}$. Consider a subsequence of w_d^b of difference c. It also must be periodic due to Lemma 1. On the other hand, it is a subsequence of w of difference dc; as it has been shown above, its language of factors is equal to $F((w')_{dc}^{dc})$. But $(w')_{dc}^{dc} = w'$ due to Lemma 10. We must conclude that w' is periodic. A contradiction. □

Lemmas from this section will be used only in the proof of the "only if" part of Theorem 1, but the proof of the "if" part will resemble the proof of Lemma 11.

6. The "if" proof

In this section, we show that if a word w is canonically p-regular for some prime p and $w^{p^a}_{p^a} = w^{p^b}_{p^b}$ for some $a \neq b$, then its arithmetical complexity grows linearly. Due to Lemma 5, we can consider w as defined by $w = P \cdot T(R_1 \cdot \ldots \cdot R_k)$, where all patterns R_i are p-regular and P is regular.

We shall divide the proof into two statements: first we shall show that arithmetical complexity of $T(R)$, where $R = R_1 \cdot \ldots \cdot R_k$, grows linearly (Lemma 12), and then that applying a regular pattern to an infinite word does not increase order of growth of arithmetical complexity (Lemma 13). Clearly, these statements imply what we need.

Lemma 12. *Let $R \in \mathcal{P}^q_{p^k}$ be a p^k-regular pattern, where p is prime and $q, k > 0$ are arbitrary. Then $a_{T(R)}(n) = \mathrm{O}(n)$.*

Note that for $k = 1$, this statement was proved in [3] as a particular case of Theorem 3. The proof below is structured like that from [3], with just one additional argument needed for $k > 1$.

Proof of Lemma 12. Note that it is sufficient to prove the lemma for the case when all symbols of the pattern R are distinct and equal to their positions in it, i.e., for

$$\begin{aligned} R &= R^q_{p^k} \\ &= 12\cdots(p^k-1)?(p^k+1)\cdots(2p^k-1)?\cdots?((q-1)p^k+1)\cdots(qp^k-1)?. \end{aligned}$$

Indeed, any other pattern $R' \in \mathcal{P}^q_{p^k}$ can be obtained from $R^q_{p^k}$ by identifying symbols, that is, by some *coding* c. If $R' = c(R^q_{p^k})$, then clearly $T(R') = c(T(R^q_{p^k}))$. Thus, any arithmetical factor of $T(R')$ can be obtained by identifying symbols from an arithmetical factor of $T(R^q_{p^k})$, and $a_{T(R')}(n) \leqslant a_{T(R^q_{p^k})}(n)$.

In what follows, we consider $R = R^q_{p^k}$. Note that as well as any pattern from $\mathcal{P}^q_{p^k}$, it can be naturally decomposed as a product of k p-regular patterns: $R^q_{p^k} = R_1 \cdot \ldots \cdot R_k$. In this particular case, symbols of R_i not equal to ? are successive numbers from 1 to qp^k which are divided by p^{i-1} but not divided by p^i. For example, $R^3_4 = 123?567?9\ 10\ 11\ ? = (1?3?5?7?9?11?) \cdot (2?6?10?)$.

To show that arithmetical complexity of $u = T(R^q_{p^k})$ grows linearly, let us consider an arbitrary arithmetical subsequence u^i_d of u and show that it belongs to the orbit of a Toeplitz word from a finite set and has linear subword complexity. Note that if $\gcd(d, qp^k) = 1$, then a part of the statement we need has been already proved in Lemma 11. But now we need to consider the general case of arbitrary d.

Consider a subsequence $v = u^i_d$ and suppose first that $\gcd(d, p) = 1$. Then exactly 1 of each p^k successive symbols of v is of order 1 in u, exactly 1 of each p^{2k} successive symbols (and one of p^k symbols of order 1) is of order 2, and so on. Let us say that a factor $s \in F(v)$ is *n-canonical* if its length is at least p^{kn}, and there exists its *canonical* occurrence to v such that the symbol of s numbered p^{kn} is of order n in u, i.e., lies in u at a position numbered mp^{kn} for some m. Clearly, n-canonical words exist for any n. Moreover, each n-canonical word s is $(n-1)$-canonical. Indeed, symbols of s numbered $p^{k(n-1)}$ and p^{kn} in its canonical occurrence to v lie in u at the distance $dp^{k(n-1)}(p^k - 1)$, i.e., $p^{k(n-1)}$th symbol of s lie in u at the position numbered $mp^{kn} - dp^{k(n-1)}(p^k - 1) = (mp^k - dp^k + d)p^{k(n-1)}$, which is of $(n-1)$th order.

Thus, there exists a sequence of n-canonical words, $n \to \infty$, tending to an infinite word $t \in \mathcal{O}(v)$. Since v is uniformly recurrent due to Lemma 4, $F(v) = F(t)$. For all n, the prefix of t of length p^{kn} is n-canonical. Let us fix n and m and consider symbols of t numbered mp^{kn} and $(m + qp^k)p^{kn}$. In each occurrence of the prefix of length $(m + qp^k)p^{kn}$ of t to v, these symbols lie at the distance $dqp^{k(n+1)}$ in u. So, if $p^k \nmid m$, these symbols are equal. This means that t is canonically p^k-regular, moreover, $t = S_1 \cdot S_2 \cdot \ldots \cdot S_n \cdot \ldots \cdot$, where $S_n \in \mathcal{P}^q_{p^k}$ for all n. Since in the initial pattern $R^q_{p^k}$ all symbols are distinct, each of S_i is uniquely determined by its first symbol s_i and the residue d' of d modulo qp^k: if $m + 1 = ip^k$, then the $(m+1)$th symbol of S_i is ?, and otherwise it is $s_i + dm \equiv s_i + d'm \pmod{qp^k}$. So, t is uniquely determined by d' and the sequence $\{s_i\}_{i=1}^{\infty}$.

Now let us show that the sequence $\{s_i\}_{i=1}^{\infty}$ is periodic and completely determined by the symbol s_1. To do it, consider a canonical occurrence in v of the prefix of t of length p^{kn} (recall that it is n-canonical). By definition, its last symbol is s_n, and the $p^{k(n-1)}$th one is s_{n-1}. In u, these symbols lie at the distance $dp^{k(n-1)}(p^k - 1)$. Here s_{n-1} lies at the position of the form $p^{k(n-1)}(mqp^k + s_{n-1})$ for some m, and s_n lies at the position of the form $p^{kn}(m'qp^k + s_n)$. We see that $p^{k(n-1)}(mqp^k + s_{n-1}) = p^{kn}(m'qp^k + s_n) - dp^{k(n-1)}(p^k - 1)$; after dividing by $p^{k(n-1)}$, this means $p^k(mq + d) + s_{n-1} - d = p^k(m'qp^k + s_n)$. Modulo qp^k, this gives $s_{n-1} \equiv p^k(s_n - d') + d'$. So, s_{n-1} is uniquely determined by s_n. Since the sequence $\{s_n\}_{n=1}^{\infty}$

is infinite, we see that it is periodic, and all symbols in its minimal period are distinct. So, it is completely determined by s_1 and d', and the same is true for $t = S_1 \cdot \ldots \cdot S_n \cdot \ldots \cdot$ as a whole: we can write $t = t(d', s_1)$.

Now let us consider the case of $v = u_d^i$ with $p|d$: more precisely, suppose that $d = p^m d'$, where $\gcd(d', p) = 1$. Two cases are possible. First, if $p^m \nmid i$, then v is qp^{k-1}-periodic. Second, if $p^m | i$, then $v = (u_{p^m}^{p^m})_{d'}^{i'}$ is an arithmetical subsequence of $u_{p^m}^{p^m} = T(R_{m+1} \cdot \ldots \cdot R_{m+k})$, where indices are taken modulo k. The difference of v as a subsequence of $u_{p^m}^{p^m}$ is coprime with p, so, it can be considered analogously to the previous case, and $F(v) = F(t)$, where $t = t(d', s_1, m')$ is a regular Toeplitz word which depends only on $d' = d \bmod qp^k$, $m' = m \bmod k$, and the initial symbol s_1. In particular, if $m = 0$, then $t(d', s_1, 0)$ is $t(d', s_1)$ defined in the previous paragraph.

Summarizing these arguments for all $v = u_d^i$, $d, i > 0$, we see that

$$A(u) = \left[\bigcup_{m=0}^{k-1} \bigcup_{d=1}^{qp^k-1} \bigcup_{s=1}^{qp^k-1} F(t(d, s, m)) \right] \bigcup Per,$$

where the unions for d and s exclude the cases when these parameters are divided by p^k. Here all $t(d, s, m)$ are p^k-regular Toeplitz words, and their subword complexity grows linearly [5,13], and *Per* is the union of sets of factors of qp^{k-1}-periodic words u_d^i corresponding to $p|(d/i)$. Subword complexity of *Per* is ultimately constant, and thus the arithmetical complexity of u grows linearly. Lemma 12 is proved. □

Example 6. Consider $w = T(R_2^3) = T(1?3?5?)$. If, for instance, $d \equiv 3 \pmod 6$, then $F(w_d^i)$ is equal either to $F(T(1? \cdot 5?))$, like for w_3^1, or to $F(T(5? \cdot 1?))$, like for w_3^2, or to $F(3^\omega)$, like for w_3^3.

Lemma 13. *Let $P \in \mathcal{P}_d^q$ be a regular pattern, u be an infinite word, and w be defined as $P \cdot u$. Then*

$$\begin{aligned} a_w(n) &\leqslant q^2 \sum_{k|d, k>1} \varphi(k) \left(k a_u \left(\left\lfloor \frac{n}{k} \right\rfloor + 1 \right) + d - k \right) + q^2(d-1) + a_u(n) \\ &\leqslant a_u(n) \cdot \mathrm{O}(q^2 d^3). \end{aligned}$$

Proof. Let us fix residues $i, j \in \{0, \ldots, qd-1\}$ and consider for all $m \geqslant 0$ arithmetical factors of w which are prefixes of length n of subsequences w_{mqd+j}^i. If $\gcd(d, j)$ divides i, then such a prefix contains $\lfloor (n/d) \gcd(d, j) \rfloor$ or $\lfloor (n/d) \gcd(d, j) \rfloor + 1$ successive symbols of $u_{mqd+j/\gcd(d,j)}^k$ for some k, situated in w_{mqd+j}^i at the distance $d/\gcd(j, d)$. There are $qd/\gcd(j, d)$ such values of i. For remaining $qd(1-1/\gcd(d, j))$ cases, sequences w_{mqd+j}^i are periodic and do not contain symbols of u at all. They do not depend on m. Summarizing these arguments for all j, we obtain that

$$a_w(n) \leqslant \sum_{j \in \{1, \ldots, qd\}} \left[\frac{qd}{\gcd(j, d)} a_u \left(\left\lfloor \frac{n}{d} \gcd(j, d) \right\rfloor + 1 \right) + qd \left(1 - \frac{1}{\gcd(j, d)} \right) \right].$$

For each j, let us define $k = d/\gcd(j, d)$. Then the formula above can be rewritten as

$$a_w(n) \leqslant q \sum_{k|d} \left(k a_u \left(\left\lfloor \frac{n}{k} \right\rfloor + 1 \right) + d - k \right) N_k,$$

where N_k is the number of values of $j \in \{1, \ldots, qd\}$ such that $d/\gcd(j, d) = k$. It can be easily seen that $N_k = q\varphi(k)$, where φ is the Euler function. Note also that for $k = 1$ and $d|i$, the arithmetical factors of w we count are already arithmetical factors of u, and we do not need to count them q^2 times instead of one. So, the resulting formula is

$$a_{P\cdot u}(n) \leqslant q^2 \sum_{k|d, k>1} \varphi(k) \left(k a_u \left(\left\lfloor \frac{n}{k} \right\rfloor + 1 \right) + d - k \right) + q^2(d-1) + a_u(n).$$

For all $k \geqslant 2$, we can roughly estimate that $a_u(\lfloor n/k \rfloor + 1) \leqslant a_u(n)$ and $\sum_{k|d} k\varphi(k) = \mathrm{O}(d^3)$. This gives the final estimate of

$$a_{P\cdot u}(n) = \mathrm{O}(q^2 d^3) a_u(n).$$

The lemma and thus the "if" part of Theorem 1 are proved. □

The remaining part of the paper is devoted to the "only if" part. We start with two sections of auxiliary statements and notions. At the end of the next section, we give a sketch of the "only if" proof.

7. Special words

Recall that a language is called *factorial* if it is closed under taking factors. Clearly, languages $F(w)$ and $A(w)$ are factorial for any word w. If w is uniformly recurrent, then they are also *prolongable*, which means that each element of either of them can be prolonged to another element of the same language by adding symbols both to the left and to the right.

One of the main techniques for computing subword complexity of a word or a factorial language is counting its special factors. A finite word u is called *special* in a factorial language F if $au \in F$ and $bu \in F$ for some distinct symbols a and b.

Let us denote the subword complexity (that is, the number of elements of length n) of a factorial language F by $f_F(n)$; the subword complexity of a word w is $f_{F(w)}(n) = f_w(n)$. If is well known that for each prolongable factorial language F, the subword complexity satisfies the inequality $f_F(n+1) \geqslant f_F(n) + s_F(n)$, where $s_F(n)$ is the number of special words of length n in F. For precise formulas involving special words see, e.g. Cassaigne [4].

Note that a prefix of a special word is also special, so, special words of a language F constitute a prefixial tree. Each of its infinite branches corresponds to a unique infinite word having the respective series of prefixes. We call this infinite word an infinite *special* word of F and denote the set of such words by $\mathcal{S}(F)$. An infinite word which is special in its language of factors is called simply *special*.

Recall that the arithmetical complexity of a word w is the subword complexity of its arithmetical closure, so, the previous formula applied for it gives $a_w(n+1) \geqslant a_w(n) +$

$s_{A(w)}(n)$. Suppose that the arithmetical closure of a word w has an infinite number of special infinite words. Then the function $s_{A(w)}(n)$ tends to infinity, and thus $a_w(n)$ grows faster than linearly. We have obtained the following:

Lemma 14. *If $a_w(n) = \mathrm{O}(n)$, then the set $\mathcal{S}(A(w))$ is finite.*

Now let us consider a special word $u = u_1 \cdots u_{mk} \in A(w)$. Suppose that au and bu are in $A(w)$, where $a, b \in \Sigma$. Then so are $au_k^k = au_k u_{2k} \cdots u_{mk}$ and $bu_k^k = bu_k u_{2k} \cdots u_{mk}$. Passing to $m \to \infty$, we obtain

Lemma 15. *If $u \in \mathcal{S}(A(w))$, then $u_k^k \in \mathcal{S}(A(w))$ for all k.*

Hence we shall say that subsequences of the form u_k^k are *special* subsequences of u.

These statements are sufficient to show by an example that if $w = T(R)$, where R is a d-regular pattern and d is not prime, then the arithmetical complexity of w is not in general linear.

Example 7. Let us consider a canonically 6-regular pattern $R = 00100?$ and the Toeplitz word $w = T(R) = 001000001000001001\ldots$. Let us show that it is special. Indeed, since it is not periodic and R contains only one gap, each prefix $w(k)$ of w of length $6^k - 1$, followed by a gap, is the minimal prefix period of $W_k = \underbrace{R \cdot \ldots \cdot R}_{k} \cdot ?$. So, each gap in T_k is followed by $w(k)$, and gaps in T_k can be substituted in T_{k+1} both by 0's and 1's. This means that $w(k)$ is special for all k, and so is w.

Due to Lemma 15, all sequences $w_{2^k}^{2^k}$ belong to $\mathcal{S}(A(w))$. The first 1 in w is its 3rd symbol. Then, $w_2^2 = 00? \cdot w$, and thus the first 1 in w_2^2 is its 9th symbol. Analogously, $w_4^4 = 00? \cdot w_2^2$, and the first 1 in it is its 27th symbol, etc. Thus, all $w_{2^k}^{2^k}$ are distinct, $\mathcal{S}(A(w))$ is infinite, and $a_w(n)$ grows faster than $\mathrm{O}(n)$ due to Lemma 14.

The following lemma is easy:

Lemma 16. *If a p-regular sequence is special then it is canonically p-regular.*

Note that the converse in not true: a canonically p-regular word may be non-special.

Example 8. The canonically 2-regular word $w = T(1?3?5?) = 113153113553\cdots$ is not special since $w_3^3 = 3^\omega$ and thus w could be prolonged to the left only by 3. But due to Lemmas 6 and 16, all special infinite words from the orbit of w are also canonically 2-regular; in fact, they are $T(3?5?1? \cdot 5?1?3?) = 35531135\cdots$ and $T(5?1?3? \cdot 3?5?1?) = 53153553\cdots$.

Lemma 17. *If a special canonically p-regular word v has linear arithmetical complexity, then it belongs to $\mathcal{P}_{p^k} \cdot T(\mathcal{P}_{p^r})$ for some k and r.*

Proof. It is sufficient to note that the set $\{v_{p^a}^{p^a}\}_{a \geqslant 0} \subseteq \mathcal{S}(A(W))$ is finite due to Lemma 15 and to use Lemma 5. □

Note that each infinite special word of $F(w)$ is also an infinite special word of $A(w)$. Such special word exists for each non-periodic w and has the same set of factors as w since w is uniformly recurrent. So, it has the same arithmetical complexity. If we prove that for some w with linear arithmetical complexity one of these special words, denoted by w', is canonically p-regular for some p, this will prove the theorem due to Lemma 17. So, without loss of generality we can consider $w = w'$, i.e., assume that our uniformly recurrent infinite word of linear arithmetical complexity is special.

In the next technical section, we prove Lemma 20 stating that a word of linear arithmetical complexity cannot simultaneously contain non-periodic p-regular and p'-regular arithmetic subsequences for prime $p \neq p'$. Then we shall pass to the main part of the proof: given a special uniformly recurrent word w of linear arithmetical complexity, we first prove in Section 9 the principal Lemma 21 asserting that w contains a special canonically p-regular subsequence w^m_m for some m and prime p. To do it, we have to split symbols of w and to pass to a sequence v on the alphabet $\mathcal{S}(A(w))$, then to find in it a special subsequence v^m_m with a needed symmetric structure, and then find in v^m_m an infinite periodic subsequence with a prime difference, denoted by p. After that we prove that v^m_m is canonically p-regular, and thus so is w^m_m.

After that in Section 10 we use Lemma 21 together with Lemmas 11 and 20 to show that w itself is canonically p-regular. Due to Lemmas 5 and 17 this will prove the theorem.

8. Some more technical lemmas

The following two lemmas will be used for the proof of Lemma 20:

Lemma 18. *For all n and D, each non ultimately periodic infinite word w contains at least $(n+1)/D$ distinct words of length n occurring in it starting with positions equal to 1 modulo D.*

Proof. Let us divide w to blocks of length D starting from the first symbol and consider these blocks as symbols of a new alphabet. The obtained word is non ultimately periodic and thus for all m contains at least $m+1$ distinct words of length m. So, the word w has at least $m+1$ words mentioned in the statement of the lemma of lengths mD to $(m+1)D-1$, and the lemma is proved. □

Lemma 19. *Let a word v occur as a factor in a word $w \in \mathcal{O}(T(\mathcal{P}^q_d))$ starting with position numbered k, and the order in w of symbols of v is bounded by $m-1 = m(v,k)-1$. Then v occurs in w as a factor starting with all positions congruent to k modulo qd^m.*

Proof. Let $w \in \mathcal{O}(T(R))$, where $R \in \mathcal{P}^q_d$. Let us consider the word $W(m)$ obtained from w by substituting all symbols of order at least m by gaps. By the definition of the order, it is periodic; on the other hand, it belongs to the orbit of $\underbrace{R \cdot R \cdot \ldots \cdot R}_{m} \cdot ?^\omega$, and thus its minimal

period $|\underbrace{R\cdot R\cdot\ldots\cdot R}_{m}|$ divides qd^m. The occurrence of v starting with position numbered k occurs already in $W(m)$ and thus in all positions of $W(m)$ (and w) congruent to k modulo qd^m. □

Lemma 20. *If an infinite word w contains non-periodic p- and p'-regular arithmetical subsequences for prime $p \neq p'$, then its arithmetical complexity grows faster than linearly.*

Proof. Suppose by contrary that $a_w(n) = \mathrm{O}(n)$. First let us show that w contains non-periodic subsequences from $\mathcal{O}(T(\mathcal{P}_{p^k}))$ and $\mathcal{O}(T(\mathcal{P}_{p'^{k'}}))$ for some $k, k' > 0$. Indeed, let us consider the p-regular non-periodic subsequence v of w and pass to a special word $v' \in \mathcal{O}(v)$. Due to Lemmas 6 and 16, it is canonically p-regular. Its arithmetical complexity $a_{v'}(n) \leqslant a_w(n) = \mathrm{O}(n)$, and due to Lemma 15, the set $\{(v')^{p^a}_{p^a} | a > 0\}$ is finite. Thus, due to Lemma 5, $v' \in \mathcal{P}\cdot T(\mathcal{P}_{p^k})$ for some k and v' has a special non-periodic subsequence $u' \in T(\mathcal{P}^q_{p^k})$ for some q. Then $v \in \mathcal{O}(v')$ contains $u \in \mathcal{O}(u') \subset \mathcal{O}(T(\mathcal{P}^q_{p^k}))$ due to Lemma 3. We have proved that w contains a subsequence from $\mathcal{O}(T(\mathcal{P}^q_{p^k}))$ (let us denote it by w^b_a); analogously, it contains a subsequence $w^d_c \in \mathcal{O}(T(\mathcal{P}^{q'}_{p'^{k'}}))$. Without loss of generality, we assume that $b \leqslant d$. In what follows we shall prove that even the subword complexity of w grows at least quadratically, contradicting to our assertion.

Since p and p' are coprime, $\gcd(aqp^h, cq'p'^{h'})$ is stabilized for all sufficiently large h, h'. Let us denote this $\lim_{h,h'\to\infty} \gcd(aqp^h, cq'p'^{h'})$ by D. Let us fix an n and consider an arbitrary word u of length $nc+1$ occurring in w^b_a on a position equal to 1 modulo D (say, at position $hD+1$); such words are at least $(nc+2)/D$ due to Lemma 18. Analogously let us consider a word u' of length $na+1$ occurring in w^d_c at a position congruent to 1 modulo D (say, at position $h'D+1$); such words are at least $(na+2)/D$. We shall prove that if neither u nor u' contain a symbol of infinite order in w, then there exists a subword $v \in F(w)$ of length $nac+d-b+1$ such that $v^1_a = u$ and $v^{d-b+1}_c = u'$. Since u and u' were chosen arbitrarily, and there is at most one symbol of infinite order in w, it will mean that $f_w(nac+d-b+1) \geqslant ((na+2)/D-1)((nc+2)/D-1) = \mathrm{O}(n^2)$, which is sufficient for the lemma to be proved.

To find the desired word v, we note that due to Lemma 19, the word u occurs in w^b_a at all positions equal to $hD+1$ modulo $qp^{km(u,hD+1)}$; in w, these are positions equal to $ahD+b$ modulo $aqp^{km(u,hD+1)}$. Analogously, u' starts in w^d_c with all positions equal to $h'D+1$ modulo $q'p'^{k'm(u',h'D+1)}$; in w, they are positions equal to $ch'D+d$ modulo $cq'p'^{k'm(u',h'D+1)}$. Thus, the needed word v is any subword of w starting with a position x, where

$$\begin{aligned} x &\equiv ahD+b \ (\mathrm{mod}\, aqp^{km(u,hD+1)}),\\ x+d-b &\equiv ch'D+d \ (\mathrm{mod}\, cq'p'^{k'm(u',h'D+1)}). \end{aligned}$$

This system always has a solution because $\gcd(aqp^{km(u,hD+1)}, cq'p'^{k'm(u',h'D+1)})$ divides $ahD-ch'D$ by the definition of D. So, the needed word v exists, and the lemma is proved. □

9. The principal lemma

The main difficulty in the proof of the "only if" part of Theorem 1 is hidden in the following:

Lemma 21. *Let $w = w_1 \cdots w_n \cdots$ be a uniformly recurrent special infinite word on an alphabet Σ having linear arithmetical complexity. Then there exist some m and a prime p such that w_m^m is canonically p-regular.*

Proof. Due to Lemmas 14 and 15, the sequences w_k^k take only a finite number of values for all $k > 0$: their set $\mathcal{S}$ is a subset of $\mathcal{S}(A(w))$. Let us consider $\mathcal{S}$ as a new alphabet and define the sequence $v = v_1 \cdots v_n \cdots \in \mathcal{S}^\omega$ by $v_k = w_k^k$ for all k.

Note that $v_k = v_l$ implies $w_k = w_l$ for all $k, l > 0$. So, w is obtained from v by applying a symbol-to-symbol morphism $c : \mathcal{S} \to \Sigma$, i.e., $w = c(v)$. Furthermore, note that $w_k^k = w_l^l$ implies $v_k^k = v_l^l$: indeed, if $w_k^k = w_l^l$, then for all n we have $(w_k^k)_n^n = w_{kn}^{kn} = v_{kn} = v_{ln} = w_{ln}^{ln} = (w_l^l)_n^n$.

The same argument shows that the symbol v_{kn} for all k and n is determined by v_k and v_n. So, we can define a commutative operation $\times_0 : \mathcal{S} \to \mathcal{S}$ such that for all $k, n > 0$ the equality holds $v_k \times_0 v_n = v_{kn}$. The initial symbol v_1 works as the unit element with respect to $\times_0$. The operation $\times_0$ is not obliged to be a group one: the symbol v_1 can be absent in some $v_{k_1}^{k_1}$, and thus it may happen that $v_{k_1} \times_0 a \neq v_1$ for all $a \in \mathcal{S}$. However, in this case we can pass from v to its special subsequence $v' = v_{k_1}^{k_1}$ which has a strictly less number of different special subsequences (and thus symbols which occur in it), and define a respective operation $\times_1$.

Since this procedure strictly decreases the number of symbols occurring in the considered word, we can continue the process until some operation $\times_l = \times$ defined according to the subsequence $v^{(l)} = (v^{(l-1)})_{k_l}^{k_l} = v_m^m$ is a group one. We denote v_m^m by z. The respective subalphabet of $\mathcal{S}$ is an abelian group with respect to $\times$. In particular, this means that for all symbols a, b, c from it the equality $a \times b = a \times c$ implies $b = c$. So, $z_{nk} = z_{nl}$ for any k, l, and n implies $z_l = z_k$ and thus $z_l^l = z_k^k$.

Note that the least possible number of distinct symbols in z is 2 because a constant infinite word cannot be special.

Let us say that the sequences z' and z'' are *clones* if they are obtained from each other by a permutation involving all symbols: $(z')_l = (z')_k$ if and only if $(z'')_l = (z'')_k$, and $(z')_k = (z'')_k$ if and only if $z' = z''$. We see that any two special subsequences of z are clones, and each of special subsequences of z can be reconstructed from any its symbol. Moreover, since $z_{nk}^{nl} = (z_n^n)_k^l$, and z and z_n^n are clones, we see that so are z_{nk}^{nl} and z_k^l for all k, l, n. In what follows we call z_{nk}^{nl} the *n-clone* of z_k^l.

Let us denote the subsequence w_m^m by y: so, $y = c(z)$. The relations among w, v, z and y are depicted below.

$$\begin{array}{ccc} v & \xrightarrow{()_m^m} & z \\ {\scriptstyle c}\downarrow & & {\scriptstyle c}\downarrow \\ w & \xrightarrow{()_m^m} & y \end{array}$$

Due to Lemma 7, to prove that y is canonically p-regular (and thus the claim holds), it is sufficient to show that z is canonically p-regular. First let us prove the following:

Claim 1. *The sequence z contains a periodic arithmetical subsequence z_d^r.*

Proof. Let us return to y. It is uniformly recurrent due to Lemma 4 and has linear arithmetical complexity $a_y(n) \leqslant a_w(n)$. Let us consider the set of all its non-periodic subsequences. Each of them is uniformly recurrent due to Lemma 4, so, it is not ultimately periodic, and the maximal power of a symbol occurring in it is finite. The language of factors of each of these subsequences has a special infinite word which has the same set of factors and belongs to $\mathcal{S}(A(w))$. Since the latter set is finite, there exists a maximal power M of a symbol occurring in all non-periodic subsequences of y.

Now let us pass again to z. It is obtained from y by splitting symbols, hence if y_d^r is non-periodic, then so is z_d^r, and the maximal power of a symbol occurring in it is also bounded by M. At the same time, $M+1$ successive equal symbols occur in some arithmetical subsequence z_d^r of z by the Van der Waerden theorem. Thus, the respective subsequence y_d^r of y is periodic. Let us consider all sequences y_{nd}^{nr}, where $n > 0$. They take a finite number of values because they are similar subsequences of special sequences y_n^n, and each of them after splitting symbols gives the n-clone z_{nd}^{nr} of z_d^r. If some of y_{nd}^{nr} is non-periodic, then it contains at most M successive equal symbols. Consequently, so do z_{nd}^{nr} and its clone z_d^r, a contradiction. Thus, all y_{nd}^{nr} are ultimately periodic and take a finite number of values. Moreover, they are periodic since uniformly recurrent.

The kth symbol of z_d^r for every k is determined by the sequence of kth symbols of $\{y_{nd}^{nr}\}_{n=1}^{\infty}$. So, the sequence z_d^r is periodic. □

Note that we could always choose $r < d$. Indeed, let $r > d$, then all $y_{nd}^{n(r-d)}$ are uniformly recurrent and thus periodic and fit to our conditions. So, we can subtract d from the number of the first symbol of the subsequence until we have $d \geqslant r$. Here we cannot have $d = r$ because the sequence y_d^d is special which implies that ay_d^d and by_d^d belong to orbits of some arithmetical subsequences of y for some $a, b \in \Sigma$, $a \neq b$. If y_d^d were periodic, then at least one of these prolonged sequences would not be uniformly recurrent, contradicting to Lemma 4.

Let us choose the difference d of a periodic subsequence of z to be minimal, and r be minimal for the given d (as we have shown, in this case we have $r < d$). Note that $\gcd(d, r) = 1$ because z_d^r is the $\gcd(d, r)$-clone of $z_{d/\gcd(d,r)}^{r/\gcd(d,r)}$ which is thus also periodic, and $\gcd(d, r) > 1$ would contradict to the minimality of d.

Claim 2. *For each s, the fact that $\gcd(d, s) = 1$ implies that z_d^s is periodic.*

Proof. Let us consider an arbitrary $s \in \{1, \ldots, d-1\}$ coprime with d. Let $t \in \{0, \ldots, t-1\}$ be the number satisfying the congruence $st \equiv r \pmod{d}$, i.e., $st = cd + r$ for some c; such t always exists. Then for every $n > 0$ the tth symbol of z_{s+nd}^{s+nd} is $z_{ts+tnd} = z_{r+d(tn+c)}$. These symbols constitute the arithmetical subsequence $z_{td}^{ts} = z_{td}^{r+dc} = (z_d^r)_t^{c+1}$ of z_d^r, which is periodic by Lemma 1. But z_{td}^{ts} is the t-clone of z_d^s, which is thus also periodic. □

As a corollary of Claim 2, we see that z_d^1 is always periodic.

Claim 3. *The minimal difference d is prime.*

Proof. Suppose by contrary that it is composite: $d = p^\alpha q$, where p is prime, $\alpha > 0$, p does not divide q, and either $\alpha > 1$ or $q > 1$. Let us fix some $k \in \{0, \dots, p-1\}$ and consider the subsequence $(z_d^p)_p^{k+1} = z_{dp}^{p+kd}$. It is the p-clone of $z_d^{1+kp^{\alpha-1}q}$ and thus has the same properties.

Case 1: Suppose that $\alpha > 1$. Then for all k we have $\gcd(1 + kp^{\alpha-1}q, d) = 1$. So, $z_d^{1+kp^{\alpha-1}q} = z_d^{1+kd/p}$ is periodic for all k due to Claim 2. But then the sequence $z_{d/p}^1 = z_d^1 \sqcup z_d^{1+d/p} \sqcup \cdots \sqcup z_d^{1+(p-1)d/p}$ is also periodic by Lemma 2, contradicting to the minimality of d. So, this case is impossible.

Case 2: Suppose that $\alpha = 1$. In this case, $d/p = q$ is coprime with p, and we have $\gcd(1 + k_1 p^{\alpha-1}q, d) = \gcd(1 + k_1 q, d) > 1$ for some unique $k_1 \in \{0, \dots, p-1\}$ because $k_1 q \equiv -1 \pmod p$. For this k_1 we have $\gcd(1 + k_1 q, d) = p$. For all $k \in \{0, \dots, p-1\}\backslash\{k_1\}$, the sequences z_{pd}^{p+kd} are periodic p-clones of $z_d^{1+kd/p}$, which are periodic due to Claim 2. Let us consider the sequence $z_{\mathrm{pd}}^{p+k_1 d}$. Positions in z of all its elements are divided by p^2; we see that p^2 divides its difference and p^3 does not. So, exactly one of each its p elements occurs at a position whose number is divided by p^3; these elements constitute the subsequence $z_{p^2 d}^{p+k_1 d+k_2 pd}$, where $k_2 \in \{0, \dots, p-1\}$. At the same time, any other subsequence of the form $z_{p^2 d}^{p+k_1 d+kpd}$, where $k \in \{0, \dots, p-1\}\backslash\{k_2\}$, is periodic since it is the p^2-clone of one of sequences z_d^l, where $\gcd(d, l) = 1$. Continuing the process, we see that for all n, only one of subsequences $z_{p^n d}^{p+(k_1+k_2 p+\cdots+k_n p^{n-1})d} = (z_d^p)_{p^n}^{[k_n \cdots k_1]_p+1}$, where $k_1, \dots, k_n \in \{0, \dots, p-1\}$, can be non-periodic. So, z_d^p is by definition p-regular with $i_n = [k_n \cdots k_1]_p + 1$. It is the p-clone of $z_{d/p}^1$, which is non-periodic due to minimality of d, so, it is non-periodic too.

Now let us prove that some of subsequences y_{nd}^{np}, $n = 1, 2, \dots$, is non-periodic p-regular. These subsequences take a finite number of values because $y_{nd}^{np} = (y_n^n)_d^p$, p and d are fixed, and y_n^n, $n = 1, 2, \dots$, take a finite number of values due to Lemma 14. Suppose that all sequences y_{nd}^{np} are periodic; since the kth symbol of z_d^p for all k depends only on the sequence of the kth symbols of y_{nd}^{np}, in this case z_d^p also would be periodic. A contradiction. So, for some n the sequence y_{nd}^{np} is non-periodic. Since it is obtained from the n-clone z_{nd}^{np} of z_d^p by applying a letter-to-letter morphism, it is p-regular due to Lemma 7. We have found a non-periodic p-regular subsequence of y.

But by our assumption, d is composite and another prime p' divides it. At the same time, $(p')^2$ does not divide d since Case 1 is impossible. So, analogously we can prove that y contains a non-periodic p'-regular subsequence. But then by Lemma 20 we have $a_y(n) \geqslant \mathrm{O}(n^2)$. Since $a_w(n) \geqslant a_y(n)$, this contradicts to the assumption that $a_w(n)$ grows linearly. Thus, Case 2 is also impossible, the minimal difference d cannot be composite, and the Claim is proved. □

We have proved that d is a prime number, $d = p$. Then Claim 2 means that *all* subsequences z_p^s, where $s < p$, are periodic. As for z_p^p, it is the p-clone of z and thus has the same property, as well as $z_{p^2}^{p^2}$, etc. Since subsequences of periodic z_p^s are also periodic, this means that z is canonically p-regular. By Lemma 7, so is $y = w_m^m$, and the statement of the lemma follows. □

10. The remaining part of the proof

Let w be a special uniformly recurrent infinite word of linear arithmetical complexity. Due to Lemma 21, some its special subsequence w_m^m is canonically p-regular for some prime p. To prove the theorem, we must show that w itself is canonically p-regular. First, let us note that it is sufficient to verify that all subsequences $w_p^1, w_p^2, \ldots, w_p^{p-1}$ are periodic. Indeed, the subsequence $(w_p^p)_m^m = (w_m^m)_p^p$ is canonically p-regular due to Lemma 9. So, w_p^p fits to the same conditions that w, and if we prove that all w_p^i, for $1 \leqslant i < p$, are periodic, then we can prove the same for respective subsequences of w_p^p, $(w_p^p)_p^p = w_{p^2}^{p^2}$, etc. The fact that w is canonically p-regular will follow by induction on the power of p.

Thus, let us consider some w_p^i for $1 \leqslant i < p$ and prove that it is periodic. Suppose by contrary that it is not. Since it is uniformly recurrent due to Lemma 4, we can pass to a special word v from its orbit without changing the set of factors and apply Lemma 21 to v: it has a canonically p'-regular non-periodic special subsequence for some prime p'. Due to Lemma 17, this subsequence has a subsequence from $T(\mathcal{P}_{p'^r}^q)$. Returning to w_p^i, we use Lemmas 3 and 6 and see that it also has a p'-regular non-periodic subsequence which in its turn has a subsequence from $\mathcal{O}(T(\mathcal{P}_{p'^r}^q))$. If $p' \neq p$, then $a_w(n)$ grows faster than linearly due to Lemma 20. So, $p' = p$, and there is a non-periodic subsequence from $\mathcal{O}(T(\mathcal{P}_{p^r}^q))$ for some r and q in w_p^i. Let us denote it by $(w_p^i)_b^{c+1} = w_{bp}^{i+cp}$.

Let us define $D = \gcd(i + cp, bp)$ and consider w_D^D. Due to Lemma 9, its special subsequence $(w_D^D)_m^m$ is canonically p-regular, and w_{bp}^{i+cp} is its subsequence $(w_D^D)_{bp/D}^{(i+cp)/D}$ whose difference and position of the first symbol are coprime. So, we can consider w_{bp}^{i+cp} as a subsequence of w_D^D instead of w to find a contradiction to the fact that w_{bp}^{i+cp} is not periodic. Since w_D^D has the same properties that w, for the sake of simplicity we can assume that $D = 1$, or, equivalently, that $\gcd(i + cp, bp) = 1$.

Let the kth order symbols be situated in w_{bp}^{i+cp} at positions equal to i_k modulo p^{rk}. Then they constitute the arithmetical subsequence $w_{bp^{rk+1}}^{i+cp+(i_k-1)bp}$ of w. Note that $\gcd(i + cp, bp) = 1$ implies $\gcd(i + cp + (i_k - 1)bp, bp^{rk+1}) = 1$. So, by the Dirichlet theorem, each of arithmetical progressions of positions in w of symbols occurring in $w_{bp^{rk+1}}^{i+cp+(i_k-1)bp}$ contains an infinite number of primes. Due to Lemma 8, we can always choose a prime number l_k from this progression, coprime with q (and, by the construction, with b and p) such that the maximal order of w_{l_k} in w is finite.

Let us consider the sequence $w_{l_k}^{l_k}$ and its intersection $v(k)$ with the p-regular sequence w_{bp}^{i+cp}: note that due to the choice of l_k, it starts with the first symbol of $w_{l_k}^{l_k}$. On one hand, since $\gcd(l_k, bp) = 1$, the sequence $v(k)$ is a subsequence of w_{bp}^{i+cp} of prime difference l_k. Since $\gcd(l_k, pq) = 1$ and due to Lemma 11, $v(k)$ is non-periodic, p^r-regular, and its first symbol (coinciding with w_{l_k}) is of the same maximal order that in w_{bp}^{i+cp}, greater than or equal to k. So, the set $\{v(k)\}_{k=1}^{\infty}$ is infinite. On the other hand, $v(k) = (w_{l_k}^{l_k})_{bp}^{1}$, i.e., $v(k)$ is the subsequence of a special subsequence of w of fixed difference bp and initial position 1. Since there is a finite number of special sequences of w due to Lemmas 14 and 15, the set $\{v(k)\}_{k=1}^{\infty}$ must also be finite. A contradiction. So, w_p^i is periodic for all $1 \leqslant i < p$, and by induction, w is canonically p-regular. The reference to Lemma 17 completes the proof of Theorem 1. □

References

[1] J.-P. Allouche, J. Shallit, The ring of k-regular sequences, Theoret. Comput. Sci. 98 (1992) 163–197.

[2] S.V. Avgustinovich, J. Cassaigne, A.E. Frid, Sequences of low arithmetical complexity, submitted, currently available at http://www.math.nsc.ru/LBRT/k4/Frid/acf_low_ar_compl.ps.

[3] S.V. Avgustinovich, D.G. Fon-Der-Flaass, A.E. Frid, Arithmetical complexity of infinite words, in: M. Ito, T. Imaoka (Eds.), Words, Languages and Combinatorics III, World Scientific, Singapore, 2003, pp. 51–62.

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

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

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

[7] A. Ehrenfeucht, K.P. Lee, G. Rozenberg, Subword complexities of various classes of deterministic developmental languages without interaction, Theoret. Comput. Sci. 1 (1975) 59–75.

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

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

[10] A. Frid, On possible growth of arithmetical complexity, submitted, currently available at http://www.math.nsc.ru/LBRT/k4/Frid/frid_for_ita.ps.

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

[12] T. Kamae, L. Zamboni, Sequence entropy and the maximal pattern complexity of infinite words, Ergodic Theory Dynam. Systems 22 (2002) 1191–1199.

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

[14] A. Restivo, S. Salemi, Binary patterns in infinite binary words, in: W. Brauer et al. (Eds.): Formal and Natural Computing, Lecture Notes in Computer Science, Vol. 2300, 2002, pp. 107–116.