Theoretical Computer Science 469 (2013) 77–91

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Abelian complexity and abelian co-decomposition

Ondřej Turek *

Laboratory of Physics, Kochi University of Technology, Tosa Yamada, Kochi 782-8502, Japan

ARTICLE INFO

Article history:
Received 10 January 2012
Accepted 23 October 2012
Communicated by J. Karhumaki

Keywords:
Abelian complexity
Recurrent word
Parry number
Tribonacci word

ABSTRACT

We propose a technique for exploring the abelian complexity of recurrent infinite words, focusing particularly on infinite words associated with Parry numbers. Using that technique, we give an affirmative answer to the open question posed by Richomme, Saari and Zamboni, whether the abelian complexity of the Tribonacci word attains each value in $\{4, 5, 6\}$ infinitely many times.

1. Introduction

Abelian complexity is now a widely studied property of infinite words. The first appearance of the idea dates back to the seventies, when Coven and Hedlund realized that periodic words and Sturmian words can be alternatively characterized using Parikh vectors [1]. Their results have been recently generalized by Richomme, Saari and Zamboni in [2], where the term "abelian complexity" itself has been introduced. That work initiated a systematic study of abelian properties of words: [2] was quickly followed by a series of related papers, both on general topics in abelian complexity [3,4] and on abelian complexity of particular infinite words [5] as well as of certain families of words [6,7].

In general, evaluating the abelian complexity function $\mathrm{AC}(n)$ for a given infinite word is a difficult problem. Only a few results, as well as effective methods, are known so far. Even the simpler question, whether a given value k of the function AC is attained finitely or infinitely many times, is usually hard to answer. This applies especially to k different from the extremal values $\max \mathrm{AC}$, $\min \mathrm{AC}$. For example, it is known that the abelian complexity of the Tribonacci word $\mathbf{t}$ (recall that $\mathbf{t}$ is the fixed point of the substitution $0 \mapsto 01$, $1 \mapsto 02$, $2 \mapsto 0$) satisfies $\mathrm{AC}_{\mathbf{t}}(n) \in \{3, 4, 5, 6, 7\}$ for all n, but only for the values 3 and 7 it is proved that they are attained infinitely many times; see [5]. Similarly, for $\mathbf{u}^{(p)}$ being the fixed point of the substitution $L \mapsto L^pS$, $S \mapsto M$, $M \mapsto L^{p-1}S$ for an arbitrary $p \geq 2$, it has been proved $\mathrm{AC}_{\mathbf{u}^{(p)}}(n) \in \{3, 4, 5, 6, 7\}$, but so far only the value 7 is known to be attained infinitely many times [6] (for additional information on those words, see [8,9]).

In this paper we develop a method for dealing with the abelian complexity of recurrent infinite words, which is fitted especially to infinite words associated with Parry numbers. It can be used for an effective calculation of $\mathrm{AC}(n)$ for a given n, as well as for proving that a certain value of AC is attained infinitely many times. To demonstrate that, we consider an open question posed by Richomme, Saari and Zamboni in [5], whether the abelian complexity of the Tribonacci word attains each value in $\{4, 5, 6\}$ infinitely often; with the help of the method, we obtain the affirmative answer.

2. Preliminaries

An alphabet $\mathcal{A}$ is a finite set of symbols called *letters*. Any concatenation of letters from $\mathcal{A}$ is called a *word*. The set $\mathcal{A}^*$ of all finite words over $\mathcal{A}$ including the empty word ε is a free monoid. For any $w = w_0w_1w_2\cdots w_{n-1} \in \mathcal{A}^*$, the *length* of w is defined as $|w| = n$. The length of the empty word is by definition $|\varepsilon| = 0$.

* Tel.: +81 887 57 3505.
E-mail addresses: ondrej.turek@kochi-tech.ac.jp, oturek@centrum.cz.

0304-3975/$ – see front matter
doi:10.1016/j.tcs.2012.10.034

An infinite sequence of letters from $\mathcal{A}$ is called an *infinite word* and the set of all infinite words over $\mathcal{A}$ is denoted by $\mathcal{A}^{\mathbb{N}}$.

A finite word w is a *factor* of a (finite or infinite) word v if there exists a finite word x and a (finite or infinite, respectively) word y such that $v = xwy$. The word w is called a *prefix* of v if $x = \varepsilon$, and a *suffix* of v, if $y = \varepsilon$.

An infinite word v is *recurrent* if every factor of v occurs infinitely many times in v.

If $w \in \mathcal{A}^*$ and $k \in \mathbb{N}$, w^k stands for the concatenation of k words w, thus $w^k = \underbrace{ww\cdots w}_{k \text{ times } w}$. We also set $w^0 = \varepsilon$. One can introduce negative powers as well. If a word $v \in \mathcal{A}^*$ has the prefix w^k, $k \in \mathbb{N}$, then the symbol $w^{-k}v$ denotes the word satisfying $w^k w^{-k} v = v$. Similarly, if a $v \in \mathcal{A}^{\mathbb{N}}$ has the suffix w^k for a $k \in \mathbb{N}$, then vw^{-k} denotes the word with the property $vw^{-k}w^k = v$.

2.1. Parikh vectors, abelian complexity, and relative Parikh vectors

Let $\mathcal{A} = \{0, 1, 2, \ldots, m-1\}$. For any $\ell \in \mathcal{A}$ and for any $w \in \mathcal{A}^*$, the symbol $|w|_\ell$ denotes the number of occurrences of the letter ℓ in the word w. The *Parikh vector* of w is the m-tuple $\Psi(w) = (|w|_0, |w|_1, \ldots, |w|_{m-1})$; note that $|w|_0 + |w|_1 + \cdots + |w|_{m-1} = |w|$.

For any given infinite word $\mathbf{u}$, we set

$$\mathcal{P}_{\mathbf{u}}(n) = \{\Psi(w) \mid w \text{ is a factor of } \mathbf{u}, |w| = n\},$$

thus $\mathcal{P}_{\mathbf{u}}(n)$ denotes the set of all Parikh vectors corresponding to factors of $\mathbf{u}$ having the length n. The *abelian complexity* of the word $\mathbf{u}$ is the function $\mathrm{AC}_{\mathbf{u}} : \mathbb{N} \to \mathbb{N}$ defined as

$$\mathrm{AC}_{\mathbf{u}}(n) = \#\mathcal{P}_{\mathbf{u}}(n), \tag{1}$$

where # denotes the cardinality. Let us introduce two new terms.

Definition 2.1. Let $\mathbf{u}_{[n]}$ denote the prefix of $\mathbf{u}$ of length $n \in \mathbb{N}_0$.

- If w is a factor of $\mathbf{u}$ of length n, then the *relative Parikh vector* of w is defined as

  $$\Psi^{\mathrm{rel}}(w) = \Psi(w) - \Psi(\mathbf{u}_{[n]}).$$

- The *set of relative Parikh vectors* corresponding to the length n is the set

  $$\mathcal{P}^{\mathrm{rel}}_{\mathbf{u}}(n) := \left\{ \Psi^{\mathrm{rel}}(w) \;\middle|\; w \text{ is a factor of } \mathbf{u}, |w| = n \right\}.$$

Remark 2.2. The idea of transforming the Parikh vectors $\Psi(w)$ into the relative Parikh vectors $\Psi^{\mathrm{rel}}(w)$ slightly resembles a technique of Adamczewski used in [10], where frequencies of letters were employed for a simplification of the study of balance properties of fixed points of primitive substitutions.

Since prefixes of $\mathbf{u}$ will play an important role in the sequel, the symbol $\mathbf{u}_{[n]}$ will be used in the same meaning throughout the whole paper.

Note that the cardinality of $\mathcal{P}^{\mathrm{rel}}_{\mathbf{u}}(n)$ is equal to the cardinality of $\mathcal{P}_{\mathbf{u}}(n)$, whence we obtain, with regard to (1),

$$\mathrm{AC}_{\mathbf{u}}(n) = \#\mathcal{P}^{\mathrm{rel}}_{\mathbf{u}}(n). \tag{2}$$

An infinite word $\mathbf{u}$ is said to be *c-balanced* if for every $\ell \in \mathcal{A}$ and for every pair of factors v, w of $\mathbf{u}$ such that $|v| = |w|$, it holds $||v|_\ell - |w|_\ell| \le c$.

Observation 2.3. *For all $n \in \mathbb{N}$, the set of relative Parikh vectors has the following properties.*

(i) $\vec{0} \in \mathcal{P}^{\mathrm{rel}}_{\mathbf{u}}(n)$ *for all n.*
(ii) If $(\psi'_0, \psi'_1, \ldots, \psi'_{m-1}) \in \mathcal{P}^{\mathrm{rel}}_{\mathbf{u}}(n)$ *and* $\mathbf{u}$ *is c-balanced, then* $|\psi'_\ell| \le c$ *for all* $\ell \in \mathcal{A}$.

Proof. (i) It is easy to see that $\vec{0}$ is the relative Parikh vector of $\mathbf{u}_{[n]}$, hence $\vec{0} \in \mathcal{P}^{\mathrm{rel}}_{\mathbf{u}}(n)$.
(ii) Let w be a factor of $\mathbf{u}$ such that $(\psi'_0, \psi'_1, \ldots, \psi'_{m-1}) = \Psi^{\mathrm{rel}}(w)$, $|w| = n$. Then $|\psi'_\ell| = \big||w|_\ell - |\mathbf{u}_{[n]}|_\ell\big|$. Since $\mathbf{u}$ is c-balanced, for any pair v, w of factors of $\mathbf{u}$, it holds $||w|_\ell - |v|_\ell| \le c$ for all $\ell \in \mathcal{A}$. Particular choice $v = \mathbf{u}_{[n]}$ gives the statement (ii). □

Remark 2.4. Let $\mathbf{u}$ be a c-balanced word. With regard to the part (ii) of Observation 2.3, the main advantage of dealing with relative Parikh vectors instead of with "standard" Parikh vectors is twofold.

- The components of $\Psi(w)$ grow to infinity with growing $|w|$ (because $|w|_0 + |w|_1 + \cdots + |w|_{m-1} = |w|$), whereas the components of $\Psi^{\mathrm{rel}}(w)$ are bounded.
- The set of all Parikh vectors $\{\Psi(w) | w \text{ is a factor of } \mathbf{u}\}$ is infinite, whereas the set of all relative Parikh vectors $\left\{\Psi^{\mathrm{rel}}(w) \middle| w \text{is a factor of} \mathbf{u}\right\}$ is finite. (This is a very important advantage.)