Acta Cybernetica **23** (2017) 175–189.

Variations of the Morse-Hedlund Theorem for k-Abelian Equivalence*

Juhani Karhumäki[a], Aleksi Saarela[a], and Luca Q. Zamboni[b]

Abstract

In this paper we investigate local-to-global phenomena for a new family of complexity functions of infinite words indexed by $k \geq 0$. Two finite words u and v are said to be k-abelian equivalent if for all words x of length less than or equal to k, the number of occurrences of x in u is equal to the number of occurrences of x in v. This defines a family of equivalence relations, bridging the gap between the usual notion of abelian equivalence (when $k = 1$) and equality (when $k = \infty$). Given an infinite word w, we consider the associated complexity function which counts the number of k-abelian equivalence classes of factors of w of length n. As a whole, these complexity functions have a number of common features: Each gives a characterization of periodicity in the context of bi-infinite words, and each can be used to characterize Sturmian words in the framework of aperiodic one-sided infinite words. Nevertheless, they also exhibit a number of striking differences, the study of which is one of the main topics of our paper.

1 Introduction

A fundamental problem in both mathematics and computer science is to describe local constraints which imply global regularities. A splendid example of this phenomena may be found in the framework of combinatorics on words. In their seminal papers [19, 20], G. A. Hedlund and M. Morse proved that a bi-infinite word w is periodic if and only if for some positive integer n, the word w contains at most n distinct factors of length n. In other words, it describes the exact borderline between periodicity and aperiodicity of words in terms of the *factor complexity function* which counts the number of distinct factors of each length n. An analogous result was established some thirty years later by E. Coven and G. A. Hedlund in the framework of abelian equivalence. They show that a bi-infinite word is periodic if

*Partially supported by the Academy of Finland grants 257857 and 137991 (FiDiPro), by ANR grant *SUBTILE*, and by the Vilho, Yrjö and Kalle Väisälä Foundation.

[a]Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland, E-mail: {karhumak,amsaar}@utu.fi

[b]Institut Camille Jordan, Université Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France, E-mail: lupastis@gmail.com

DOI: 10.14232/actacyb.23.1.2017.11

and only if for some positive integer n all factors of w are abelian equivalent. Thus once again it is possible to distinguish between periodic and aperiodic words on a local level by counting the number of abelian equivalence classes of factors of length n.

In this paper we study the local-to-global behavior for a new family of complexity functions $\mathcal{P}_w^k$ of infinite words indexed by $k \in \mathbb{Z}_+ \cup \{\infty\}$ where $\mathbb{Z}_+ = \{1, 2, 3, \ldots\}$ denotes the set of positive integers. Let $k \in \mathbb{Z}_+ \cup \{\infty\}$ and A be a finite non-empty set. Two finite words $u, v \in A^*$ are said to be *k-abelian equivalent* if for all $x \in A^*$ of length at most k, the number of occurrences of x in u is equal to the number of occurrences of x in v. This defines a family of equivalence relations $\sim_k$ on A^*, bridging the gap between the usual notion of abelian equivalence (when $k = 1$) and equality (when $k = \infty$). Abelian equivalence of words has long been a subject of great interest (see, for instance, Erdős's problem, [5, 6, 7, 9, 17, 22, 23, 24, 26]). Although the notion of k-abelian equivalence is quite new, there are already a number of papers on the topic [11, 12, 13, 14, 15, 18].

Given an infinite word $w \in A^\omega$, we consider the associated complexity function $\mathcal{P}_w^k : \mathbb{Z}_+ \to \mathbb{Z}_+$ which counts the number of k-abelian equivalence classes of factors of w of length n. Thus $\mathcal{P}_w^\infty$ corresponds to the usual factor complexity (sometimes called subword complexity in the literature) while $\mathcal{P}_w^1$ corresponds to abelian complexity. As it turns out, each intermediate complexity function $\mathcal{P}_w^k$ can be used to detect periodicity of words. As a starting point of our research, we list two classical results on factor and abelian complexity in connection with periodicity, and their k-abelian counterparts proved by the authors in [15]. We note that in each case, the first two items are included in the third.

Theorem 1. *Let w be a bi-infinite word over a finite alphabet. Then the following properties hold:*

- *(M. Morse, G. A. Hedlund, [19]) The word w is periodic if and only if $\mathcal{P}_w^\infty(n) < n + 1$ for some $n \geq 1$.*
- *(E. M. Coven, G. A. Hedlund, [6]) The word w is periodic if and only if $\mathcal{P}_w^1(n) < 2$ for some $n \geq 1$.*
- *The word w is periodic if and only if $\mathcal{P}_w^k(n) < \min\{n + 1, 2k\}$ for some $k \in \mathbb{Z}_+ \cup \{\infty\}$ and $n \geq 1$.*

Also, each complexity provides a characterization for an important class of binary words, the so-called *Sturmian* words:

Theorem 2. *Let w be an aperiodic one-sided infinite word. Then the following properties hold:*

- *(M. Morse, G. A. Hedlund, [20]). The word w is Sturmian if and only if $\mathcal{P}_w^\infty(n) = n + 1$ for all $n \geq 1$.*
- *(E. M. Coven, G. A. Hedlund, [6]). The word w is Sturmian if and only if $\mathcal{P}_w^1(n) = 2$ for all $n \geq 1$.*

- *The word w is Sturmian if and only if* $\mathcal{P}_w^k(n) = \min\{n+1, 2k\}$ *for all* $k \in \mathbb{Z}_+ \cup \{\infty\}$ *and* $n \geq 1$.

However, in other respects, these various complexities exhibit radically different behaviors. For instance, in the context of one-sided infinite words, the first item in Theorem 1 gives rise to a characterization of ultimately periodic words, while for the other two, the result holds in only one direction: If $\mathcal{P}_w^k(n) < \min\{n+1, 2k\}$ for some $k \in \mathbb{Z}_+$ and $n \geq 1$ then w is ultimately periodic, but not conversely (see [15]). For instance in the simplest case when $k = 1$, it is easy to see that if w is the ultimately periodic word 01^ω, then for each positive integer n there are precisely two abelian classes of factors of w of length n. However, the same is true of the (aperiodic) *infinite Fibonacci word* $w = 0100101001001\cdots$ defined as the fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 0$. Analogously, in Theorem 2 the first item holds without the added assumption that w be aperiodic, while the other two items do not. Another striking difference between them is in their rate of growth. Consider for instance the binary Champernowne word $\mathcal{C} = 011011100101110111\cdots$ obtained by concatenating the binary representation of the consecutive natural numbers. Let w denote the morphic image of $\mathcal{C}$ under the Thue–Morse morphism $0 \mapsto 01$, $1 \mapsto 10$. Then while $\mathcal{P}_w^\infty(n)$ has exponential growth, it can be shown that $\mathcal{P}_w^1(n) \leq 3$ for all n. Yet another fundamental disparity concerns the difference $\mathcal{P}_w^k(n+1) - \mathcal{P}_w^k(n)$. For factor complexity, one always has $\mathcal{P}_w^\infty(n+1) - \mathcal{P}_w^\infty(n) \geq 0$, while for general k this inequality is far from being true.

A primary objective in this paper is to study the asymptotic lower and upper complexities defined by

$$\mathcal{L}_w^k(n) = \min_{m \geq n} \mathcal{P}_w^k(m) \qquad \text{and} \qquad \mathcal{U}_w^k(n) = \max_{m \leq n} \mathcal{P}_w^k(m).$$

Surprisingly these quantities can deviate from one another quite drastically. Indeed, one of our main results is to compute these values for the famous Thue–Morse word. We show that the upper limit is logarithmic, while the lower limit is just constant, in fact at most 8 in the case $k = 2$. This is quite unexpected considering the Thue–Morse word is both pure morphic and abelian periodic (of period 2). If we however allow more general words, then we obtain much stronger evidence of the non-existence of gaps in low k-abelian complexity classes. We construct uniformly recurrent infinite words having arbitrarily low upper limit and just constant lower limit. The concept of k-abelian complexity also leads to many interesting open questions. We conclude the paper in Section 6 by mentioning some of these problems.

This is an extended version of an article that was presented at the 18th conference on Developments in Language Theory [16].

2 Preliminaries

Let Σ be a finite non-empty set called the *alphabet*. The set of all finite words over Σ is denoted by Σ^* and the set of all (right) infinite words is denoted by Σ^ω. The set of positive integers is denoted by $\mathbb{Z}_+$. A function $f : \mathbb{Z}_+ \to \mathbb{R}$ is *increasing* if $f(m) \leq f(n)$ for all $m < n$, and *strictly increasing* if $f(m) < f(n)$ for all $m < n$.

Let $w \in \Sigma^\omega$. The word w is *periodic* if there is $u \in \Sigma^*$ such that $w = u^\omega$, and *ultimately periodic* if there are $u, v \in \Sigma^*$ such that $w = vu^\omega$. If w is not ultimately periodic, then it is *aperiodic*. Let $u = a_0 \cdots a_{m-1}$ and $a_0, \ldots, a_{m-1} \in \Sigma$. The *prefix* of length n of u is $\mathrm{pref}_n(u) = a_0 \cdots a_{n-1}$ and the *suffix* of length n of u is $\mathrm{suff}_n(u) = a_{m-n} \cdots a_{m-1}$. If $0 \leq i \leq m$, then the notation $\mathrm{rfact}_n^i(u) = a_i \cdots a_{i+n-1}$ is used. The length of a word u is denoted by $|u|$ and the number of occurrences of another word x as a factor of u by $|u|_x$. As a trivial boundary case, $|u|_\varepsilon = |u| + 1$. Two words $u, v \in \Sigma^*$ are *abelian equivalent* if $|u|_a = |v|_a$ for all $a \in \Sigma$.

Let $k \in \mathbb{Z}_+$. Two words $u, v \in \Sigma^*$ are *k-abelian equivalent* if $|u|_x = |v|_x$ for all words x of length at most k. k-abelian equivalence is denoted by $\sim_k$. If the length of u and v is at least $k-1$, then $u \sim_k v$ if and only if $|u|_x = |v|_x$ for all words x of length k and $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ and $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$. This gives an alternative definition for k-abelian equivalence. A proof can be found in [15].

Let $w \in \Sigma^\omega$. The set of factors of w of length n is denoted by $\mathcal{F}_w(n)$. The *factor complexity* of w is the function $\mathcal{P}_w^\infty : \mathbb{Z}_+ \to \mathbb{Z}_+$ defined by

$$\mathcal{P}_w^\infty(n) = \#\mathcal{F}_w(n),$$

where $\#$ is used to denote the cardinality of a set. Let $k \in \mathbb{Z}_+$. The *k-abelian complexity* of w is the function $\mathcal{P}_w^k : \mathbb{Z}_+ \to \mathbb{Z}_+$ defined by

$$\mathcal{P}_w^k(n) = \#(\mathcal{F}_w(n)/\sim_k).$$

Factor complexity functions are always increasing, and even strictly increasing for aperiodic words. For k-abelian complexity this is not true. This is why we define *upper k-abelian complexity* $\mathcal{U}_w^k$ and *lower k-abelian complexity* $\mathcal{L}_w^k$:

$$\mathcal{U}_w^k(n) = \max_{m \leq n} \mathcal{P}_w^k(m) \qquad \text{and} \qquad \mathcal{L}_w^k(n) = \min_{m \geq n} \mathcal{P}_w^k(m).$$

These two functions can be significantly different. For example, if w is the Thue–Morse word and $k \geq 2$, then $\mathcal{U}_w^k(n) = \Theta(\log n)$ and $\mathcal{L}_w^k(n) = \Theta(1)$. This will be proved in Section 4.

When using Θ-notation, the parameter k and the size of the alphabet are assumed to be fixed, so the implied constants of the Θ-notation can depend on them.

The abelian complexity of a binary word $w \in \{0,1\}^\omega$ can be determined by using the formula (see [24])

$$\mathcal{P}_w^1(n) = \max\{|u|_1 \mid u \in \mathcal{F}_n(w)\} - \min\{|u|_1 \mid u \in \mathcal{F}_n(w)\} + 1. \tag{1}$$

For $k \in \mathbb{Z}_+ \cup \{\infty\}$, we define

$$q^k : \mathbb{Z}_+ \to \mathbb{Z}_+, q^k(n) = \min\{n+1, 2k\}.$$

The significance of this function is that if w is Sturmian, then $\mathcal{P}_w^k = q^k$. This is further discussed in Section 3.

There are large classes of words for which the k-abelian complexities are of the same order for many values of k. This is shown in the next two lemmas. Thus when analyzing the growth rate of the k-abelian complexity of a word, it may be sufficient to analyze the abelian or 2-abelian complexity.

Lemma 1. *Let $w \in \{0,1\}^\omega$ be such that every factor of w of length k contains at most one occurrence of 1. Then $\mathcal{P}_w^k(n) = \Theta(\mathcal{P}_w^1(n))$.*

Proof. Clearly $\mathcal{P}_w^k(n) \geq \mathcal{P}_w^1(n)$. Let u be a factor of w of length n. Let $x = 0^i 1 0^{k-i-1}$. Every factor of w of length k except 0^k is of this form, because every factor of w of length k contains at most one occurrence of 1. For the same reason, $|u|_x = |u|_1 - a$, where $a \in \{0,1,2\}$ depending on $\mathrm{pref}_{k-1}(u)$ and $\mathrm{suff}_{k-1}(u)$. It follows that the k-abelian equivalence class of u is determined by $\mathrm{pref}_{k-1}(u)$, $\mathrm{suff}_{k-1}(u)$, and $|u|_1$. The number of possible pairs $(\mathrm{pref}_{k-1}(u), \mathrm{suff}_{k-1}(u))$ is at most k^2, and the number of possible values for $|u|_1$ is $\mathcal{P}_w^1(n)$, so $\mathcal{P}_w^k(n) \leq k^2 \mathcal{P}_w^1(n)$. □

Lemma 2. *Let $k, m \geq 2$ and let w be a fixed point of an m-uniform morphism h. Let i be such that $m^i \geq k-1$. Then $\mathcal{P}_w^k(m^i(n+1)) = O(\mathcal{P}_w^2(n))$.*

Proof. Every factor of w of length $m^i(n+1)$ can be written as $ph^i(u)q$, where u is a factor of w of length n and $|pq| = m^i$. The k-abelian equivalence class of $ph^i(u)q$ is determined by p, q, and the 2-abelian equivalence class of u. The number of possible pairs (p,q) is $O(1)$, and the number of possible values for the 2-abelian equivalence class of u is $\mathcal{P}_w^2(n)$. The claim follows. □

In particular, Lemma 2 can be applied to the Thue–Morse word to analyze its k-abelian complexity once the behavior of its 2-abelian complexity is known.

It has been shown that there are many words for which the k-abelian and $(k+1)$-abelian complexities are similar, but there are also many words for which they are very different. For example, there are words having bounded k-abelian complexity but linear $(k+1)$-abelian complexity. These words can even be assumed to be k-abelian periodic, meaning that they are of the form $u_1 u_2 \cdots$, where $u_1, u_2, \ldots$ are k-abelian equivalent. This is shown in the next lemma.

Lemma 3. *For every $k \geq 1$, there is a k-abelian periodic word w such that $\mathcal{P}_w^{k+1}(n) = \Theta(n)$.*

Proof. Let $W \in \{0,1\}^\omega$ be a word with linear abelian complexity (e.g., the Champernowne word) and let h be the morphism defined by

$$h(0) = 0^{k+1} 1 0^{k-1} 1, \qquad h(1) = 0^k 1 0^k 1.$$

Then the word $w = h(W)$ is k-abelian periodic of period $2k+2$. If $u, v \in \{0,1\}^*$ are not abelian equivalent, then $h(u)$ and $h(v)$ are not $(k+1)$-abelian equivalent because the factor $1 0^{k-1} 1$ appears only inside $h(0)$. On the other hand, if $u, v \in \{0,1\}^*$ are abelian equivalent and $p, q \in \{0,1\}^*$, then $ph(u)q$ and $ph(v)q$ are $(k+1)$-abelian equivalent. It follows that

$$\mathcal{P}_w^{k+1}((2k+2)n) = \Theta(\mathcal{P}_W^1(n)) = \Theta(n). \tag{2}$$

We know that

$$\mathcal{P}_w^{k+1}(n+1) \leq 2\mathcal{P}_w^{k+1}(n) \tag{3}$$

for all n (this would work for all words w if 2 would be replaced by the size of the alphabet). Every n can be written as $(2k+2)n' + r$, where $0 \leq r < 2k+2$, so from (2) and (3) it follows that

$$\mathcal{P}_w^{k+1}(n) = \mathcal{P}_w^{k+1}((2k+2)n' + r) \leq 2^r \mathcal{P}_w^{k+1}((2k+2)n') = \Theta(n') = \Theta(n).$$

Similarly, every n can be written as $(2k+2)n' - r$, where $0 \leq r < 2k+2$, so from (2) and (3) it follows that

$$\mathcal{P}_w^{k+1}(n) = \mathcal{P}_w^{k+1}((2k+2)n' - r) \geq 2^{-r} \mathcal{P}_w^{k+1}((2k+2)n') = \Theta(n') = \Theta(n).$$

The claim follows. □

3 Minimal k-Abelian Complexities

In this section classes of words with small k-abelian complexity are studied. Some well-known results about factor complexity are compared to results on k-abelian complexity proved in [15]. It should be expected that ultimately periodic words have low complexity, and this is indeed true for k-abelian complexity, although the k-abelian complexity of some ultimately periodic words is higher that the k-abelian complexity of some aperiodic words. For many complexity measures, Sturmian words have the lowest complexity among aperiodic words. This is also true for k-abelian complexity.

We recall the famous theorem of Morse and Hedlund [19] characterizing ultimately periodic words in terms of factor complexity. This theorem can be generalized for k-abelian complexity: If $\mathcal{P}_w^k(n) < q^k(n)$ for some n, then w is ultimately periodic, and if w is ultimately periodic, then $\mathcal{P}_w^\infty(n)$ is bounded. This was proved in [15].

If k is finite, then this generalization does not give a characterization of ultimately periodic words, because the function q^k is bounded. In fact, it is impossible to characterize ultimately periodic words in terms of k-abelian complexity. For example, the word $0^{2k-1}1^\omega$ has the same k-abelian complexity as every Sturmian word. On the other hand, for every ultimately periodic word w there is a finite k such that $\mathcal{P}_w^k(n) < q^k(n)$ for all sufficiently large n.

The theorem of Morse and Hedlund has a couple of immediate consequences. The words w with $\mathcal{P}_w^\infty(n) = n + 1$ for all n are, by definition, Sturmian words. Thus the following classification is obtained:

- w is ultimately periodic $\Leftrightarrow \mathcal{P}_w^\infty$ is bounded.

- w is Sturmian $\Leftrightarrow \mathcal{P}_w^\infty(n) = n + 1$ for all n.
- w is aperiodic and not Sturmian $\Leftrightarrow \mathcal{P}_w^\infty(n) \geq n+1$ for all n and $\mathcal{P}_w^\infty(n) > n+1$ for some n.

This can be generalized for k-abelian complexity if the equivalences are replaced with implications:

- w is ultimately periodic $\Rightarrow \mathcal{P}_w^k$ is bounded.
- w is Sturmian $\Rightarrow \mathcal{P}_w^k = q^k$.
- w is aperiodic and not Sturmian $\Rightarrow \mathcal{P}_w^k(n) \geq q^k(n)$ for all n and $\mathcal{P}_w^k(n) > q^k(n)$ for some n.

For $k = 1$ this follows from the theorem of Coven and Hedlund [6]. For $k \geq 2$ it follows from a theorem in [15].

The above result means that one similarity between factor complexity and k-abelian complexity is that Sturmian words have the lowest complexity among aperiodic words. Another similarity between them is that ultimately periodic words have bounded complexity, and the largest values can be arbitrarily high: For every n, there is a finite word u having every possible factor of length n. Then $\mathcal{P}_{u^\omega}^k(n)$ is as high as it can be for any word, i.e., the number of k-abelian equivalence classes of words of length n.

Another direct consequence of the theorem of Morse and Hedlund is that there is a gap between constant complexity and the complexity of Sturmian words. For k-abelian complexity there cannot be a gap between bounded complexities and q^k, because the function q^k itself is bounded. However, the question whether there is a gap above bounded complexity is more difficult. The answer is that there is no such gap, even if only uniformly recurrent words are considered. This is proved in Section 5.

4 k-Abelian Complexity of the Thue–Morse Word

In this section the k-abelian complexity of the Thue–Morse word is analyzed. Before that, the abelian complexity of a closely related word is determined.

Let σ be the morphism defined by $\sigma(0) = 01, \sigma(1) = 00$. Let

$$S = 0100010101000100010001010100010 1\cdots$$

be the *period-doubling word*, which is the fixed point of σ; see, e.g., [8].

The abelian complexity of S is completely determined by the recurrence relations in the next lemma and by the first value $\mathcal{P}_S^1(1) = 2$. These relations were proved independently in [3]. It is an easy consequence that the abelian complexity of S is 2-regular (2-regular sequences were defined in [2]). The 2-abelian complexity of the Thue–Morse word has been conjectured to be 2-regular [25], and this is proved in [10] and [21].

Lemma 4. *For $n \geq 1$,*

$$\mathcal{P}_S^1(2n) = \mathcal{P}_S^1(n) \qquad \text{and} \qquad \mathcal{P}_S^1(4n \pm 1) = \mathcal{P}_S^1(n) + 1.$$

Proof. Let

$$p_n = \min\{|u|_1 \mid u \in \mathcal{F}_n(S)\} \qquad \text{and} \qquad q_n = \max\{|u|_1 \mid u \in \mathcal{F}_n(S)\}.$$

Let $\overline{0} = 1$ and $\overline{1} = 0$. For $a \in \{0,1\}$, $\sigma(a) = 0\overline{a}$ and $\sigma^2(a) = 010a$. Because

$$\mathcal{F}_{2n}(S) = \{\sigma(u) \mid u \in \mathcal{F}_n(S)\} \cup \{\overline{a}\sigma(u)0 \mid au \in \mathcal{F}_n(S)\},$$

it can be seen that $p_{2n} = n - q_n$ and $q_{2n} = n - p_n$. Because

$$\begin{aligned}\mathcal{F}_{4n-1}(S) = &\left\{\sigma^2(u)010 \mid u \in \mathcal{F}_{n-1}(S)\right\} \cup \left\{10a\sigma^2(u) \mid au \in \mathcal{F}_n(S)\right\} \cup \\ &\left\{0a\sigma^2(u)0 \mid au \in \mathcal{F}_n(S)\right\} \cup \left\{a\sigma^2(u)01 \mid au \in \mathcal{F}_n(S)\right\},\end{aligned}$$

it can be seen that

$$\begin{aligned}p_{4n-1} &= \min\{p_{n-1} + n, p_n + n, p_n + n - 1, p_n + n\} = p_n + n - 1,\\ q_{4n-1} &= \max\{q_{n-1} + n, q_n + n, q_n + n - 1, q_n + n\} = q_n + n.\end{aligned}$$

Because

$$\begin{aligned}\mathcal{F}_{4n+1}(S) = &\left\{\sigma^2(u)0 \mid u \in \mathcal{F}_n(S)\right\} \cup \left\{10a\sigma^2(u)01 \mid au \in \mathcal{F}_n(S)\right\} \cup \\ &\left\{0a\sigma^2(u)010 \mid au \in \mathcal{F}_n(S)\right\} \cup \left\{a\sigma^2(u) \mid au \in \mathcal{F}_{n+1}(S)\right\}\end{aligned}$$

it can be seen that

$$\begin{aligned}p_{4n+1} &= \min\{p_n + n, p_n + n + 1, p_n + n, p_{n+1} + n - 1\} = p_n + n,\\ q_{4n+1} &= \max\{q_n + n, q_n + n + 1, q_n + n, q_{n+1} + n - 1\} = q_n + n + 1.\end{aligned}$$

The claim follows because $\mathcal{P}_S^1(n) = q_n - p_n + 1$ for all n by (1). $\square$

Theorem 3. *For $n \geq 1$ and $m \geq 0$,*

$$\mathcal{P}_S^1(n) = O(\log n), \quad \mathcal{P}_S^1((2 \cdot 4^m + 1)/3) = m + 2, \quad \mathcal{P}_S^1(2^m) = 2.$$

Proof. Follows from Lemma 4 by induction. $\square$

The abelian complexity of S has a logarithmic upper bound and a constant lower bound. These bounds are the best possible increasing bounds.

Corollary 1. $\mathcal{U}_S^1(n) = \Theta(\log n)$ *and* $\mathcal{L}_S^1(n) = 2$.

Let τ be the Thue–Morse morphism defined by $\tau(0) = 01, \tau(1) = 10$. Let

$$T = 0110100110010110100101100110100 1\cdots$$

be the Thue–Morse word, which is a fixed point of τ. The first values of $\mathcal{P}_T^2$ are

$$2, 4, 6, 8, 6, 8, 10, 8, 6, 8, 8, 10, 10, 10, 8, 8, 6, 8, 10, 10.$$

The 2-abelian equivalence of factors of T can be determined with the help of the following lemma.

Lemma 5. *Words $u, v \in \{0,1\}^*$ are 2-abelian equivalent if and only if*

$$|u| = |v|, \qquad |u|_{00} = |v|_{00}, \qquad |u|_{11} = |v|_{11}, \qquad \textit{and} \qquad \mathrm{pref}_1(u) = \mathrm{pref}_1(v).$$

Proof. The "only if" direction follows immediately from the alternative definition of 2-abelian equivalence. For the other direction, it follows from the assumptions that $|u|_{01} + |u|_{10} = |v|_{01} + |v|_{10}$. In any word $w \in \{0,1\}^*$, the numbers $|w|_{01}$ and $|w|_{10}$ can differ by at most one. If $|w|_{01} + |w|_{10}$ is even, then $|w|_{01} = |w|_{10}$. If it is odd and $\mathrm{pref}_1(w) = 0$, then $|w|_{01} = |w|_{10} + 1$. If it is odd and $\mathrm{pref}_1(w) = 1$, then $|w|_{01} + 1 = |w|_{10}$. This means that $|u|_{01} = |v|_{01}$ and $|u|_{10} = |v|_{10}$ and u and v are 2-abelian equivalent. □

The following lemma states that if u is a factor of T, then the numbers $|u|_{00}$ and $|u|_{11}$ can differ by at most one.

Lemma 6. *In the image of any word under τ, between any two occurrences of* 00 *there is an occurrence of* 11 *and vice versa.*

Proof. 00 can only occur in the middle of $\tau(10)$, and 11 can only occur in the middle of $\tau(01)$. The claim follows because 10's and 01's alternate in all binary words. □

Let u be a factor of T. If $|u|$ and $|u|_{00} + |u|_{11}$ are given, then there are at most 4 possibilities for the 2-abelian equivalence class of u. This is stated in a more precise way in the next lemma. First we define a function ϕ as follows. If $w = a_1 \cdots a_n$, then $\phi(w) = b_1 \cdots b_{n-1}$, where $b_i = 0$ if $a_i a_{i+1} \in \{01, 10\}$ and $b_i = 1$ if $a_i a_{i+1} \in \{00, 11\}$. If $w = a_1 a_2 \cdots$ is an infinite word, then $\phi(w) = b_1 b_2 \cdots$ is defined in an analogous way.

Lemma 7. *Let $u_1, \ldots, u_n$ be factors of T. Let $\phi(u_1), \ldots, \phi(u_n)$ be abelian equivalent and $|\phi(u_1)|_1 = m$. If m is even, then $u_1, \ldots, u_n$ are in at most 2 different 2-abelian equivalence classes, and if m is odd, then $u_1, \ldots, u_n$ are in at most 4 different 2-abelian equivalence classes.*

Proof. We have $|u_i|_{00} + |u_i|_{11} = |\phi(u_i)|_1 = m$ for all i. By Lemma 6, we have $\{|u_i|_{00}, |u_i|_{11}\} = \{\lfloor m/2 \rfloor, \lceil m/2 \rceil\}$. If m is even, there are at most two different possible values for the triples $(|u_i|_{00}, |u_i|_{11}, \mathrm{pref}_1(u_i))$, and if m is odd, there are at most four different possible values. The claim follows from Lemma 5. □

Now it can be proved that the 2-abelian complexity of T is of the same order as the abelian complexity of $\phi(T)$. It is known that $\phi(T)$ is actually the period-doubling word S [1].

Lemma 8. *For $n \geq 2$,*

$$\mathcal{P}_S^1(n-1) \leq \mathcal{P}_T^2(n) \leq 3\mathcal{P}_S^1(n-1) + \begin{cases} 0 & \textit{if } \mathcal{P}_S^1(n-1) \textit{ is even} \\ 1 & \textit{if } \mathcal{P}_S^1(n-1) \textit{ is odd.} \end{cases}$$

Proof. If the factors of T of length n are $u_1, \ldots, u_m$, then the factors of $\phi(T)$ of length $n-1$ are $\phi(u_1), \ldots, \phi(u_m)$. If u_i and u_j are 2-abelian equivalent, then $\phi(u_i)$ and $\phi(u_j)$ are abelian equivalent, so the first inequality follows. The second inequality follows from Lemma 7, because the number of different values $|\phi(u_i)|_1$ is $\mathcal{P}_S^1(n-1)$, and at least $\lfloor \mathcal{P}_S^1(n-1)/2 \rfloor$ of these different values are even. □

Theorem 4. *For $n \geq 1$ and $m \geq 0$,*

$$\mathcal{P}_T^2(n) = O(\log n), \quad \mathcal{P}_T^2((2 \cdot 4^m + 4)/3) = \Theta(m), \quad \mathcal{P}_T^2(2^m + 1) \leq 6.$$

Proof. Follows from Lemma 8 and Theorem 3. □

With the help of Lemma 2, we see that the k-abelian complexity of T behaves in a similar way as the abelian complexity of S.

Corollary 2. *Let $k \geq 2$. Then $\mathcal{U}_T^k(n) = \Theta(\log n)$ and $\mathcal{L}_T^k(n) = \Theta(1)$.*

5 Arbitrarily Slowly Growing k-Abelian Complexities

In this section we study whether there is a gap above bounded k-abelian complexity. This question can be formalized in several different ways:

1. Does there exist an increasing unbounded function $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ such that for every infinite word w, either $\mathcal{P}_w^k$ is bounded or $\mathcal{P}_w^k = \Omega(f)$?
2. Does there exist an increasing unbounded function $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ such that for every infinite word w, either $\mathcal{P}_w^k$ is bounded or $\mathcal{P}_w^k \neq O(f)$?
3. Does there exist an increasing unbounded function $f : \mathbb{Z}_+ \to \mathbb{Z}_+$ such that for every infinite word w, either $\liminf \mathcal{P}_w^k < \infty$ or $\mathcal{P}_w^k \neq O(f)$?

The first question has already been answered negatively in Section 4. The answers to the second and third question are also negative. In the case of the second question, we prove this by a uniformly recurrent construction, and in the case of the third question, we prove this by a recurrent construction.

First, consider the second question. Let $n_1, n_2, \ldots$ be a sequence of integers greater than 1. Let $m_j = \prod_{i=1}^{j} n_i$ for $j = 0, 1, 2, \ldots$. Let $a_i = 0$ if the greatest j such that $m_j | i$ is even and $a_i = 1$ otherwise. Let $U = a_1 a_2 a_3 \cdots$. The idea is that the faster the sequence $n_1, n_2, \ldots$ grows, the slower the k-abelian complexity of the word U grows.

The word U could also be described by a Toeplitz-type construction: Start with the word $(0^{n_1 - 1} \diamond)^\omega$, then replace the $\diamond$'s by the letters of $(1^{n_2 - 1} \diamond)^\omega$, then replace the remaining $\diamond$'s by the letters of $(0^{n_3 - 1} \diamond)^\omega$, then replace the remaining $\diamond$'s by the letters of $(1^{n_4 - 1} \diamond)^\omega$, and keep repeating this procedure so that U is obtained as a limit. It follows from the construction that $U \in (\mathrm{pref}_{m_j - 1}(U)\{0, 1\})^\omega$ for all j.

Lemma 9. *The word U is uniformly recurrent.*

Proof. For every factor u of U, there is a j such that u is a factor of $\mathrm{pref}_{m_j-1}(U)$. Because $U \in \{\mathrm{pref}_{m_j-1}(U)0, \mathrm{pref}_{m_j-1}(U)1\}^\omega$, every factor of U of length $2m_j - 2$ contains u. $\square$

Lemma 10. *For every $n \geq 2$, let n' be such that $m_{n'-1} < n \leq m_{n'}$. Then*

$$\mathcal{P}_U^1(n) \leq n' + 1.$$

For all $J \geq 1$, if $n = 2\sum_{j=1}^{J}(m_{2j} - m_{2j-1})$, then

$$\mathcal{P}_U^1(n) \geq \frac{n'+1}{2}.$$

For all $j \geq 1$,

$$\mathcal{P}_U^1(m_j) = 2.$$

Proof. Formula (1) will be used repeatedly in this proof. Another important simple fact is that if a, b, c are integers and c divides a, then $\lfloor (a+b)/c \rfloor = a/c + \lfloor b/c \rfloor$.

For all $n \geq 1$,

$$|\mathrm{pref}_n(U)|_1 = \sum_{i=1}^{\infty} (-1)^{i+1} \left\lfloor \frac{n}{m_i} \right\rfloor,$$

and for all $n \geq 1$ and $l \geq 0$,

$$|\mathrm{rfact}_n^l(U)|_1 = |\mathrm{pref}_{n+l}(U)|_1 - |\mathrm{pref}_l(U)|_1 = \sum_{i=1}^{\infty} (-1)^{i+1} \left(\left\lfloor \frac{n+l}{m_i} \right\rfloor - \left\lfloor \frac{l}{m_i} \right\rfloor \right).$$

For all i,

$$\left\lfloor \frac{(n+l)}{m_i} \right\rfloor - \left\lfloor \frac{l}{m_i} \right\rfloor \in \left\{ \left\lfloor \frac{n}{m_i} \right\rfloor, \left\lceil \frac{n}{m_i} \right\rceil \right\}.$$

Moreover, for every n and l there is an i' such that, for $i \geq n'$,

$$\left\lfloor \frac{n+l}{m_i} \right\rfloor - \left\lfloor \frac{l}{m_i} \right\rfloor = \begin{cases} 1 & \text{if } n' \leq i < i' \\ 0 & \text{if } i \geq i' \end{cases},$$

so

$$\sum_{i=n'}^{\infty} (-1)^{i+1} \left(\left\lfloor \frac{n+l}{m_i} \right\rfloor - \left\lfloor \frac{l}{m_i} \right\rfloor \right) \in \left\{ 0, (-1)^{n'+1} \right\}.$$

Thus there are at most $n'+1$ possible values for $|\mathrm{rfact}_n^l(U)|_1$ and $\mathcal{P}_U^1(n) \leq n'+1$.

Consider the second claim. Let $n = 2\sum_{j=1}^{J}(m_{2j} - m_{2j-1})$. The sequence (m_j) is increasing and, moreover, $m_{j+1} \geq 2m_j$ for all j, so by standard estimates for alternating sums,

$$m_{2J} \leq 2(m_{2J} - m_{2J-1}) < n < 2m_{2J} \leq m_{2J+1}.$$

Thus $n' = 2J + 1$. Let $l = m_{2J+1} - n/2$. Then

$$|\mathrm{rfact}_n^l(U)|_1 - |\mathrm{pref}_n(U)|_1 = \sum_{i=1}^{\infty}(-1)^{i+1}\left(\left\lfloor\frac{n+l}{m_i}\right\rfloor - \left\lfloor\frac{l}{m_i}\right\rfloor - \left\lfloor\frac{n}{m_i}\right\rfloor\right)$$

and for $i \leq 2J$ (recall that $m_i | m_j$ when $j \geq i$)

$$\begin{aligned}
&\left\lfloor\frac{(n+l)}{m_i}\right\rfloor - \left\lfloor\frac{l}{m_i}\right\rfloor - \left\lfloor\frac{n}{m_i}\right\rfloor \\
=&\frac{m_{2J+1} + \sum_{(i+1)/2 \leq j \leq J}(m_{2j} - m_{2j-1})}{m_i} + \left\lfloor\frac{\sum_{1 \leq j < (i+1)/2}(m_{2j} - m_{2j-1})}{m_i}\right\rfloor \\
&- \frac{m_{2J+1} - \sum_{(i+1)/2 \leq j \leq J}(m_{2j} - m_{2j-1})}{m_i} - \left\lfloor - \frac{\sum_{1 \leq j < (i+1)/2}(m_{2j} - m_{2j-1})}{m_i}\right\rfloor \\
&- \frac{2\sum_{(i+1)/2 \leq j \leq J}(m_{2j} - m_{2j-1})}{m_i} - \left\lfloor\frac{2\sum_{1 \leq j < (i+1)/2}(m_{2j} - m_{2j-1})}{m_i}\right\rfloor \\
=&\left\lfloor\frac{s}{m_i}\right\rfloor - \left\lfloor-\frac{s}{m_i}\right\rfloor - \left\lfloor\frac{2s}{m_i}\right\rfloor,
\end{aligned}$$

where $s = \sum_{1 \leq j < (i+1)/2}(m_{2j} - m_{2j-1})$. If i is even, then $m_i/2 \leq s < m_i$, and if i is odd and $i > 1$, then $m_{i-1}/2 \leq s < m_{i-1}$. Thus

$$\left\lfloor\frac{s}{m_i}\right\rfloor - \left\lfloor-\frac{s}{m_i}\right\rfloor - \left\lfloor\frac{2s}{m_i}\right\rfloor = \begin{cases} 0 & \text{if } i \text{ is even or } i = 1 \\ 1 & \text{if } i \text{ is odd and } i > 1 \end{cases}$$

and

$$\begin{aligned}
\mathcal{P}_U^1(n) &\geq |\mathrm{rfact}_n^l(U)|_1 - |\mathrm{pref}_n(U)|_1 + 1 \\
&= \sum_{i'=2}^{J}(-1)^{(2i'-1)+1} + \sum_{i=2J+1}^{\infty}(-1)^{i+1}\left(\left\lfloor\frac{n+l}{m_i}\right\rfloor - \left\lfloor\frac{l}{m_i}\right\rfloor - \left\lfloor\frac{n}{m_i}\right\rfloor\right) + 1 \\
&= J + 1 = \frac{n'+1}{2}.
\end{aligned}$$

Consider the third claim. Because $U \in \{\mathrm{pref}_{m_j-1}(U)0, \mathrm{pref}_{m_j-1}(U)1\}^\omega$, every factor of U of length m_j is abelian equivalent to either the word $\mathrm{pref}_{m_j-1}(U)0$ or the word $\mathrm{pref}_{m_j-1}(U)1$. Thus $\mathcal{P}_U^1(m_j) \leq 2$. Both $\mathrm{pref}_{m_j-1}(U)0$ and $\mathrm{pref}_{m_j-1}(U)1$ are factors of U, so $\mathcal{P}_U^1(m_j) = 2$. □

If $n_i = 2$ for all i, then the word U is the period-doubling word S. Thus Lemma 10 gives an alternative proof for Corollary 1.

Theorem 5. *For every increasing unbounded function $f : \mathbb{Z}_+ \to \mathbb{Z}_+$, there is a uniformly recurrent word $w \in \{0,1\}^\omega$ such that $\mathcal{P}_w^k(n) = O(f(n))$ but $\mathcal{P}_w^k(n)$ is not bounded.*

Proof. Follows from Lemmas 1, 9 and 10. □

Consider the third question. Let $m_0, m_1, \ldots$ be a sequence of positive integers. Let $v_0 = 0^{m_0}1$ and $v_n = v_{n-1}v_{n-1}0^{m_i}$ for $n \geq 1$. Let V be the limit of the sequence $v_0, v_1, v_2, \ldots$. Again, the idea is that the faster the sequence $m_0, m_1, \ldots$ grows, the slower the k-abelian complexity of the word V grows.

Lemma 11. *The word V is recurrent and* $\liminf \mathcal{P}_V^1(n) = \infty$.

Proof. Every factor of V is a factor of v_n for some n, and $v_n v_n$ is a prefix of V, so every factor appears at least twice in V. Thus V is recurrent.

The word V has factors 0^i for all i, so by (1), $\mathcal{P}_V^1$ is increasing. Moreover, the word V has factors with arbitrarily many 1's, so $\liminf \mathcal{P}_V^1(n) = \infty$. □

Lemma 12. *For every $n \geq m_0 + 2$, let n' be such that $|v_{n'-1}| < n \leq |v_{n'}|$. Then*

$$\mathcal{P}_V^1(n) \leq 2^{n'} + 1.$$

Proof. The word V has factors 0^i for all i, so by (1),

$$\mathcal{P}_V^1(n) = \max\{|v|_1 \mid v \in \mathcal{F}_n(V)\} + 1.$$

Because $V \in (\{v_{n'}\} \cup 0^*)^\omega$,

$$\max\{|v|_1 \mid v \in \mathcal{F}_n(V)\} \leq |v_{n'}|_1 = 2^{n'}.$$

The claim follows. □

Theorem 6. *For every increasing unbounded function $f : \mathbb{Z}_+ \to \mathbb{Z}_+$, there is a recurrent word $w \in \{0,1\}^\omega$ such that $\mathcal{P}_w^k(n) = O(f(n))$ but* $\liminf \mathcal{P}_w^k(n) = \infty$.

Proof. Follows from Lemmas 1, 11 and 12. □

6 Conclusion

In this paper we have investigated some generalizations of the results of Morse and Hedlund and those of Coven and Hedlund for k-abelian complexity. We have pointed out many similarities but also many differences. We have studied the k-abelian complexity of the Thue–Morse word and proved that there are uniformly recurrent words with arbitrarily slowly growing k-abelian complexities.

There are many open questions and possible directions for future work. Inspired by Lemma 3, the relations of k-abelian complexities for different values of k could be studied. In fact, several questions related to this idea were answered in [4]. Another interesting topic would be the k-abelian complexities of morphic words. For example, for a morphic (or pure morphic) word w, how slowly can $\mathcal{U}_w^k(n)$ grow without being bounded? Can it grow slower than logarithmically? More generally, can the possible k-abelian complexities of some subclass of morphic words be classified?

References

[1] Allouche, Jean-Paul, Arnold, André, Berstel, Jean, Brlek, Srečko, Jockusch, William, Plouffe, Simon, and Sagan, Bruce E. A relative of the Thue-Morse sequence. *Discrete Math.*, 139(1–3):455–461, 1995.

[2] Allouche, Jean-Paul and Shallit, Jeffrey. The ring of k-regular sequences. *Theoret. Comput. Sci.*, 98(2):163–197, 1992.

[3] Blanchet-Sadri, Francine, Currie, James, Rampersad, Narad, and Fox, Nathan. Abelian complexity of fixed point of morphism $0 \mapsto 012, 1 \mapsto 02, 2 \mapsto 1$. *Integers*, 14:A11, 2014.

[4] Cassaigne, Julien, Karhumäki, Juhani, and Saarela, Aleksi. On growth and fluctuation of k-abelian complexity. In *Proceedings of the 10th CSR*, volume 9139 of *LNCS*, pages 109–122. Springer, 2015.

[5] Cassaigne, Julien, Richomme, Gwénaël, Saari, Kalle, and Zamboni, Luca Q. Avoiding Abelian powers in binary words with bounded Abelian complexity. *Internat. J. Found. Comput. Sci.*, 22(4):905–920, 2011.

[6] Coven, Ethan M. and Hedlund, Gustav A. Sequences with minimal block growth. *Math. Systems Theory*, 7:138–153, 1973.

[7] Currie, James and Rampersad, Narad. Recurrent words with constant Abelian complexity. *Adv. in Appl. Math.*, 47(1):116–124, 2011.

[8] Damanik, David. Local symmetries in the period-doubling sequence. *Discrete Appl. Math.*, 100(1–2):115–121, 2000.

[9] Dekking, Michel. Strongly nonrepetitive sequences and progression-free sets. *J. Combin. Theory Ser. A*, 27(2):181–185, 1979.

[10] Greinecker, Florian. On the 2-abelian complexity of the Thue–Morse word. *Theoret. Comput. Sci.*, 593:88–105, 2015.

[11] Huova, Mari and Karhumäki, Juhani. Observations and problems on k-abelian avoidability. In *Combinatorial and Algorithmic Aspects of Sequence Processing (Dagstuhl Seminar 11081)*, pages 2215–2219, 2011.

[12] Huova, Mari, Karhumäki, Juhani, and Saarela, Aleksi. Problems in between words and abelian words: k-abelian avoidability. *Theoret. Comput. Sci.*, 454:172–177, 2012.

[13] Huova, Mari, Karhumäki, Juhani, Saarela, Aleksi, and Saari, Kalle. Local squares, periodicity and finite automata. In Calude, Cristian, Rozenberg, Grzegorz, and Salomaa, Arto, editors, *Rainbow of Computer Science*, volume 6570 of *LNCS*, pages 90–101. Springer, 2011.

[14] Karhumäki, Juhani, Puzynina, Svetlana, and Saarela, Aleksi. Fine and Wilf's theorem for k-abelian periods. *Internat. J. Found. Comput. Sci.*, 24(7):1135–1152, 2013.

[15] Karhumäki, Juhani, Saarela, Aleksi, and Zamboni, Luca Q. On a generalization of Abelian equivalence and complexity of infinite words. *J. Combin. Theory Ser. A*, 120(8):2189–2206, 2013.

[16] Karhumäki, Juhani, Saarela, Aleksi, and Zamboni, Luca Q. Variations of the Morse-Hedlund theorem for k-abelian equivalence. In *Proceedings of the 18th DLT*, volume 8633 of *LNCS*, pages 203–214. Springer, 2014.

[17] Keränen, Veikko. Abelian squares are avoidable on 4 letters. In *Proceedings of the 19th ICALP*, volume 623 of *LNCS*, pages 41–52. Springer, 1992.

[18] Mercaş, Robert and Saarela, Aleksi. 3-abelian cubes are avoidable on binary alphabets. In *Proceedings of the 17th DLT*, volume 7907 of *LNCS*, pages 374–383. Springer, 2013.

[19] Morse, Marston and Hedlund, Gustav A. Symbolic dynamics. *Amer. J. Math.*, 60(4):815–866, 1938.

[20] Morse, Marston and Hedlund, Gustav A. Symbolic dynamics II: Sturmian trajectories. *Amer. J. Math.*, 62(1):1–42, 1940.

[21] Parreau, Aline, Rigo, Michel, Rowland, Eric, and Vandomme, Elise. A new approach to the 2-regularity of the l-abelian complexity of 2-automatic sequences. *Electron. J. Combin.*, 22(1):P1.27, 2015.

[22] Puzynina, Svetlana and Zamboni, Luca Q. Abelian returns in Sturmian words. *J. Combin. Theory Ser. A*, 120(2):390–408, 2013.

[23] Richomme, Gwénaël, Saari, Kalle, and Zamboni, Luca Q. Balance and Abelian complexity of the Tribonacci word. *Adv. in Appl. Math.*, 45(2):212–231, 2010.

[24] Richomme, Gwénaël, Saari, Kalle, and Zamboni, Luca Q. Abelian complexity of minimal subshifts. *J. Lond. Math. Soc. (2)*, 83(1):79–95, 2011.

[25] Rigo, Michel and Vandomme, Elise. 2-abelian complexity of the Thue–Morse sequence, 2012. http://hdl.handle.net/2268/135841.

[26] Saarela, Aleksi. Ultimately constant abelian complexity of infinite words. *J. Autom. Lang. Comb.*, 14(3–4):255–258, 2009.