On a generalization of Abelian equivalence and complexity of infinite words

Juhani Karhumaki[a,1], Aleksi Saarela[a,1], Luca Q. Zamboni[a,b,2]

[a] *Department of Mathematics and Statistics & FUNDIM, University of Turku, FI-20014 Turku, Finland*
[b] *Université de Lyon, Université Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 boulevard du 11 novembre 1918, F69622 Villeurbanne Cedex, France*

Abstract

In this paper we introduce and study a family of complexity functions of infinite words indexed by $k \in \mathbb{Z}^+ \cup \{+\infty\}$. Let $k \in \mathbb{Z}^+ \cup \{+\infty\}$ and A be a finite non-empty set. Two finite words u and v in A^* are said to be k-Abelian equivalent if for all $x \in A^*$ 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 $\sim_k$ on A^*, bridging the gap between the usual notion of Abelian equivalence (when $k = 1$) and equality (when $k = +\infty$). We show that the number of k-Abelian equivalence classes of words of length n grows polynomially, although the degree is exponential in k. Given an infinite word $\omega \in A^{\mathbb{N}}$, we consider the associated complexity function $\mathcal{P}^{(k)}_\omega : \mathbb{N} \to \mathbb{N}$ which counts the number of k-Abelian equivalence classes of factors of ω of length n. We show that the complexity function $\mathcal{P}^{(k)}$ is intimately linked with periodicity. More precisely we define an auxiliary function $q^k : \mathbb{N} \to \mathbb{N}$ and show that if $\mathcal{P}^{(k)}_\omega(n) < q^k(n)$ for some $k \in \mathbb{Z}^+ \cup \{+\infty\}$ and $n \geq 0$, then ω is ultimately periodic. Moreover if ω is aperiodic, then $\mathcal{P}^{(k)}_\omega(n) = q^k(n)$ if and only if ω is Sturmian. We also study k-Abelian complexity in connection with repetitions in words. Using Szemerédi's theorem, we show that if ω has bounded k-Abelian complexity, then for every $D \subset \mathbb{N}$ with positive upper density and for every positive integer N, there exists a k-Abelian N power occurring in ω at some position $j \in D$.

Keywords: Abelian equivalence, complexity of words, Sturmian words, Szemerédi's theorem.
2010 MSC: 68R15

1. Introduction

Abelian equivalence of words has long been a subject of great interest (see for instance Erdös problem, [3, 4, 5, 6, 19, 26, 28, 29, 31]). Given a finite non-empty set A, let A^* denote the set of all finite words over A. Two words u and v in A^* are *Abelian equivalent,* denoted $u \sim_{\mathrm{ab}} v$, if and only if $|u|_a = |v|_a$ for all $a \in A$, where $|u|_a$ and $|v|_a$ denote the number of

Email addresses: karhumak@utu.fi (Juhani Karhumaki), amsaar@utu.fi (Aleksi Saarela), lupastis@gmail.com (Luca Q. Zamboni)

[1] Partially supported by the Academy of Finland under grant 257857.

[2] Partially supported by a FiDiPro grant (137991) from the Academy of Finland and by ANR grant *SUBTILE.*

occurrences of a in u and v, respectively. It is readily verified that $\sim_{\text{ab}}$ defines an equivalence relation (in fact a congruence) on A^*.

We consider the following natural generalization: Fix $k \in \mathbb{Z}^+ \cup \{+\infty\}$. Two words u and v in A^* are said to be *k-Abelian equivalent*, written $u \sim_k v$, if $|u|_x = |v|_x$ for each non-empty word x with $|x| \leq k$ (where $|x|$ denotes the length of x, and $|u|_x$ and $|v|_x$ denote the number of occurrences of x in u and v, respectively). We note that $u \sim_{+\infty} v$ if and only if $u = v$, while $\sim_1$ corresponds to the usual notion of Abelian equivalence $\sim_{\text{ab}}$. Thus one may regard the notion of k-Abelian equivalence as gradually bridging the gap between Abelian equivalence ($k = 1$) and equality ($k = +\infty$). It is readily verified that $\sim_k$ defines an equivalence relation (in fact a congruence) on A^*. Clearly, if $u \sim_k v$, then $|u| = |v|$ and $u \sim_\ell v$ for each positive integer $\ell \leq k$.

The notion of k-Abelian equivalence was first introduced by the first author in [17] in connection with formal languages and decidability questions of various fundamental problems. It was shown that the well known Parikh Theorem on the equivalence of Parikh images of regular and context-free languages does not hold for k-abelian equivalence. In contrast various highly nontrivial decidability questions including the D0L sequence equivalence problem [8] or the Post Correspondence Problem [25], turned out to be easily decidable in the context of k-Abelian equivalence. Recently k-Abelian equivalence has been studied in the context of avoidance of repetitions in words (see the discussion at the beginning of §5 on k-Abelian powers). In this paper we undergo an investigation of the complexity of infinite words in the framework of k-Abelian equivalence. As is the case with various other notions of complexity of words, we will see that k-Abelian complexity is intimately linked with periodicity and can be used to detect the presence of repetitions.

Let A be a finite non-empty set. For each infinite word $\omega = a_0a_1a_2\ldots$ with $a_i \in A$, we denote by $\mathcal{F}_\omega(n)$ the set of all *factors* of ω of length n, that is, the set of all finite words of the form $a_ia_{i+1}\cdots a_{i+n-1}$ with $i \geq 0$. We set

$$\rho_\omega(n) = \text{Card}(\mathcal{F}_\omega(n)).$$

The function $\rho_\omega : \mathbb{N} \to \mathbb{N}$ is called the *factor complexity function* of ω. Analogously, for each $k \in \mathbb{Z}^+ \cup \{+\infty\}$ we define

$$\mathcal{P}^{(k)}_\omega(n) = \text{Card}\left(\mathcal{F}_\omega(n)/\sim_k\right).$$

The function $\mathcal{P}^{(k)}_\omega : \mathbb{N} \to \mathbb{N}$, which counts the number of k-Abelian equivalence classes of factors of ω of length n, is called the *k-Abelian complexity* of ω. In case $k = +\infty$ we have that $\mathcal{P}^{(+\infty)}_\omega(n) = \rho_\omega(n)$, while if $k = 1$, $\mathcal{P}^{(1)}_\omega(n)$, denoted $\rho^{\text{ab}}_\omega(n)$, corresponds to the usual Abelian complexity of ω.

Most word complexity functions, including factor complexity [24], maximal pattern complexity [16], permutation complexity [1, 10], Abelian complexity [4], and Abelian maximal pattern complexity [15], may be used to detect (and in some cases characterize) ultimately periodic words. For instance, a celebrated result due to Morse and Hedlund [24] states that an infinite word $\omega \in A^\mathbb{N}$ is ultimately periodic if and only if $\rho_\omega(n) \leq n$ for some $n \in \mathbb{Z}^+$. The third author together with T. Kamae proved a similar result in the context of maximal pattern complexity with n replaced by $2n - 1$ (see [16]). Furthermore, amongst all aperiodic (meaning non-ultimately periodic) words, Sturmian words generally have the lowest possible complexity[3]. We show that these same results hold in the framework of k-Abelian complexity. In order to formulate the precise link between aperiodicity and k-Abelian complexity, we define, for each $k \in \mathbb{Z}^+ \cup \{+\infty\}$, an auxiliary function $q^{(k)} : \mathbb{N} \to \mathbb{N}$ by

$$q^{(k)}(n) = \begin{cases} n+1 & \text{for } n \leq 2k-1 \\ 2k & \text{for } n \geq 2k. \end{cases}$$

We prove that for $\omega \in A^{\mathbb{N}}$, if $\mathcal{P}^{(k)}_{\omega}(n_0) < q^{(k)}(n_0)$ for some $k \in \mathbb{Z}^+ \cup \{+\infty\}$ and $n_0 \geq 1$, then ω is ultimately periodic.

This result is already well known in the special cases $k = +\infty$ and $k = 1$ (see [24] and [4] respectively). By the Morse-Hedlund result mentioned earlier, this condition gives a characterization of ultimately periodic words in the special case $k = +\infty$. In contrast, k-Abelian complexity does not yield such a characterization. Indeed, both Sturmian words and the ultimately periodic word $01^{\infty} = 0111\cdots$ have the same constant 2-Abelian complexity. More generally, we shall see that the ultimately periodic word $0^{2k-1}1^{\infty}$ has the same k-Abelian complexity as a Sturmian word. Nevertheless k-Abelian complexity gives a complete characterization of Sturmian words amongst all aperiodic words. More precisely, we prove that for an aperiodic word $\omega \in A^{\mathbb{N}}$, the following conditions are equivalent:

- ω is a balanced binary word, that is, *Sturmian.*
- $\mathcal{P}^{(k)}_{\omega}(n) = q^{(k)}(n)$ for each $k \in \mathbb{Z}^+ \cup \{+\infty\}$ and $n \geq 1$.

Again, the special cases of $k = +\infty$ and $k = 1$ were already known (see [24] and [4] respectively).

Finally we investigate the question of avoidance of k-Abelian N powers: By a k-Abelian N power we mean a word U of the form $U = U_1U_2\ldots U_N$ such that $U_i \sim_k U_j$ for all $1 \leq i, j \leq N$. Using Szemerédi's theorem [32], we show that if ω has bounded k-Abelian complexity, then for every $D \subset \mathbb{N}$ with positive upper density and for every positive integer N, there exists a k-Abelian N power occurring in ω at some position $j \in D$.

The paper is organized as follows: In §2 we recall some basic definitions and notations and establish various basic properties of k-Abelian equivalence of words. Also in §2 we compute the rate of growth of the number of k-Abelian equivalence classes of words in A^n. In §3 we develop the link between k-Abelian complexity and periodicity of words. In §4 we compute the k-Abelian complexity of Sturmian words and show that it completely characterizes Sturmian words amongst all aperiodic words. Finally in §5 we study k-Abelian complexity in the context of repetitions in words.

2. *k*-Abelian equivalence

2.1. Definitions and first properties

Given a finite non-empty set A, we denote by A^* the set of all finite words over A including the empty word, denoted by ε, by A^+ the set of all finite non-empty words over A, by $A^{\mathbb{N}}$

[3] With respect to maximal pattern complexity, and Abelian maximal pattern complexity, Sturmian words are not the only words of lowest complexity.

the set of (right) infinite words over A, and by $A^{\mathbb{Z}}$ the set of bi-infinite words over A. Given a finite word $u = a_1a_2\dots a_n$ with $n \geq 1$ and $a_i \in A$, we denote the length n of u by $|u|$ (by convention we set $|\varepsilon| = 0$.) For each $x \in A^+$, we let $|u|_x$ denote the number of occurrences of x in u. We denote the reversal $a_na_{n-1}\dots a_1$ of u by $\bar{u}$.

A factor u of $\omega = a_0a_1a_2\dots \in A^{\mathbb{N}}$ is called *right special* (respectively *left special*) if there exists distinct symbols $a, b \in A$ such that both ua and ub (respectively au and bu) are factors of ω. We say u is *bispecial* if u is both left and right special. An infinite word $\omega \in A^{\mathbb{N}}$ is said to be *periodic* if there exists a positive integer p such that $a_{i+p} = a_i$ for all indices i. It is said to be *ultimately periodic* if $a_{i+p} = a_i$ for all sufficiently large i. It is said to be *aperiodic* if it is not ultimately periodic. Sturmian words are the *simplest* aperiodic infinite words; Sturmian words are infinite words over a binary alphabet having exactly $n + 1$ factors of length n for each $n \geq 0$. Their origin can be traced back to the astronomer J. Bernoulli III in 1772. A fundamental result due to Morse and Hedlund [24] states that each aperiodic (meaning non-ultimately periodic) infinite word must contain at least $n + 1$ factors of each length $n \geq 0$. Thus Sturmian words are those aperiodic words of lowest factor complexity. They arise naturally in many different areas of mathematics including combinatorics, algebra, number theory, ergodic theory, dynamical systems and differential equations. Sturmian words are also of great importance in theoretical physics and in theoretical computer science and are used in computer graphics as digital approximation of straight lines. If $\omega \in \{a, b\}^{\mathbb{N}}$ is Sturmian, then for each positive integer n there exists a unique right special (respectively left special) factor of length n, and one is the reversal of the other. In particular, if x is a bispecial factor, the x is a *palindrome*, i.e., $x = \bar{x}$. For more on Sturmian words, we refer the reader to [20].

Definition 2.1. Let $k \in \mathbb{Z}^+ \cup \{+\infty\}$. We say two words $u, v \in A^+$ are *k-Abelian equivalent* and write $u \sim_k v$, if $|u|_x = |v|_x$ for all words x of length $|x| \leq k$.

We note that if $u, v \in A^+$ and $|u| = |v| \leq k$, then $u \sim_k v$ if and only if $u = v$.

Example 2.2. *The words $u = 010110$ and $v = 011010$ are 3-Abelian equivalent but not 4-Abelian equivalent since the prefix 0101 of u does not occur in v. The words $u = 0110$ and $v = 1101$ are not 2-Abelian equivalent (since they are not Abelian equivalent) yet for every word x of length 2 we have $|u|_x = |v|_x$.*

The next lemma gives different equivalent ways of defining k-Abelian equivalence. For example, item (1) corresponds to the Definition 2.1 and item (3) corresponds to another common definition: Words u and v of length at least $k - 1$ are k-Abelian equivalent if they share the same prefixes and suffixes of length $k - 1$ and if $|u|_x = |v|_x$ for every word t of length k.

Lemma 2.3. *Let u and v be words of length at least $k - 1$ and let $|u|_t = |v|_t$ for every word t of length k. The following are equivalent:*

1. $|u|_s = |v|_s$ *for all* $s \in A^{\leq k-1}$,
2. $|u|_s = |v|_s$ *for all* $s \in A^{k-1}$,
3. $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ *and* $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$,
4. $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$,
5. $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$,

6. $\mathrm{pref}_i(u) = \mathrm{pref}_i(v)$ *and* $\mathrm{suff}_{k-1-i}(u) = \mathrm{suff}_{k-1-i}(v)$ *for some* $i \in \{0, \ldots, k-1\}$.

Proof. (1) ⇒ (2): Clear.

(2) ⇒ (3): Let $\{t_1, \ldots, t_n\}$ be the multiset of factors of u (and of v) of length k. The multiset of factors of u of length $k-1$ is

$$\{\mathrm{pref}_{k-1}(u)\} \cup \{\mathrm{suff}_{k-1}(t_1), \ldots, \mathrm{suff}_{k-1}(t_n)\},$$

and the multiset of factors of v of length $k-1$ is

$$\{\mathrm{pref}_{k-1}(v)\} \cup \{\mathrm{suff}_{k-1}(t_1), \ldots, \mathrm{suff}_{k-1}(t_n)\}.$$

These multisets must be the same, so $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$. Similarly, $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$.

(3) ⇒ (4), (5): Clear.

(4) or (5) ⇒ (6): Clear.

(6) ⇒ (1): Let $\{t_1, \ldots, t_n\}$ be the multiset of factors of u (and of v) of length k. Every

$$s \in A^{k-1} \smallsetminus \{\mathrm{pref}_{k-1}(u), \mathrm{suff}_{k-1}(u)\}$$

appears in the multiset

$$\{\mathrm{pref}_{k-1}(t_1), \ldots, \mathrm{pref}_{k-1}(t_n)\} \cup \{\mathrm{suff}_{k-1}(t_1), \ldots, \mathrm{suff}_{k-1}(t_n)\} \tag{2.1}$$

$2|u|_s$ times. A word $s \in \{\mathrm{pref}_{k-1}(u), \mathrm{suff}_{k-1}(u)\}$ appears $2|u|_s - 1$ times if $\mathrm{pref}_{k-1}(u) \neq \mathrm{suff}_{k-1}(u)$, and $2|u|_s - 2$ times if $\mathrm{pref}_{k-1}(u) = \mathrm{suff}_{k-1}(u)$. Similarly, every

$$s \in A^{k-1} \smallsetminus \{\mathrm{pref}_{k-1}(v), \mathrm{suff}_{k-1}(v)\}$$

appears $2|v|_s$ times, and a word $s \in \{\mathrm{pref}_{k-1}(v), \mathrm{suff}_{k-1}(v)\}$ appears $2|v|_s - 1$ times if $\mathrm{pref}_{k-1}(v) \neq \mathrm{suff}_{k-1}(v)$, and $2|v|_s - 2$ times if $\mathrm{pref}_{k-1}(v) = \mathrm{suff}_{k-1}(v)$.

If some words appear an odd number of times in (2.1), then these must be $\mathrm{pref}_{k-1}(u)$ and $\mathrm{suff}_{k-1}(u)$, and they must also be $\mathrm{pref}_{k-1}(v)$ and $\mathrm{suff}_{k-1}(v)$. If follows that $|u|_s = |v|_s$ for every $s \in A^{k-1}$. (In this case the assumption (6) was not needed.)

If all words appear an even number of times in (2.1), then necessarily $\mathrm{pref}_{k-1}(u) = \mathrm{suff}_{k-1}(u)$ and $\mathrm{pref}_{k-1}(v) = \mathrm{suff}_{k-1}(v)$. From (6) it follows that $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$ and $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$, and thus $|u|_s = |v|_s$ for every $s \in A^{k-1}$.

The fact that $|u|_s = |v|_s$ also for every s of length less than $k-1$ can be proved in a similar way. □

The next lemma lists some basic facts on k-Abelian equivalence:

Lemma 2.4. *Let* $u, v \in A^*$ *and* $k \geq 1$.

- *If* $|u| = |v| \leq 2k-1$ *and* $u \sim_k v$, *then* $u = v$.
- *If* $u \sim_k v$, *then* $u \sim_{k'} v$ *for all* $k' \leq k$.
- *If* $u_1 \sim_k v_1$ *and* $u_2 \sim_k v_2$, *then* $u_1 u_2 \sim_k v_1 v_2$.

The bound $2k-1$ in Lemma 2.4 is optimal as for each positive integer k there exist words $u \neq v$ of length $2k$ such that $u \sim_k v$. For example, the words $u = 0^{k-1}010^{k-1}$ and $v = 0^{k-1}100^{k-1}$ of length $2k$ are readily verified to be k-Abelian equivalent (see Proposition 2.8).

Lemma 2.5. *Fix $2 \leq k < +\infty$. Suppose $aub \sim_k cvd$ with $a, b, c, d \in A$ and $u, v \in A^*$. Then $u \sim_{k-1} v$.*

Proof. Let $x \in A^*$ with $|x| \leq k-1$. We can assume that $|x| < |aub|$ for otherwise $0 = |u|_x = |v|_x$. If x is neither a prefix nor a suffix of aub, then by Lemma 2.3 x is neither a prefix nor suffix of cvd and hence $|u|_x = |aub|_x = |cvd|_x = |v|_x$. If x is either a prefix of aub or a suffix of aub but not both, the $|u|_x = |aub|_x - 1 = |cvd|_x - 1 = |v|_x$. Finally if x is both a prefix and a suffix of aub then $|u|_x = |aub|_x - 2 = |cvd|_x - 2 = |v|_x$. □

2.2. A first connection to Sturmian words

The next theorem gives a complete classification of pairs of k-Abelian equivalent words of length $2k$ and establishes a first link to Sturmian words:

Theorem 2.6. *Fix a positive integer k, and let $u, v \in A^*$ be distinct words of length $2k$. Then $u \sim_k v$ if and only if there exist distinct letters $a, b \in A$, a Sturmian word $\omega \in \{a, b\}^{\mathbb{N}}$ and a right special factor x of ω of length $k-1$ (or empty in case $k = 1$) such that*

$$u = xab\bar{x} \quad \text{and} \quad v = xba\bar{x}.$$

In particular u and v are both factors of the same Sturmian word ω.

Remark 2.7. It follows that if u and v are distinct k-Abelian equivalent words of length $2k$, then both u and v are on a binary alphabet and in fact factors of the same Sturmian word ω. In fact, if B is a bispecial factor of ω then both $BabB$ and $BbaB$ are factors of ω. Also, if x is a right special factor of ω, then there exists a bispecial factor B of ω with x a suffix of B and $\bar{x}$ a prefix of B. Thus both $xab\bar{x}$ and $xba\bar{x}$ are factors of ω.

We will need the next result applied to Sturmian words, but we prove it more generally for episturmian words. An infinite word is *episturmian* if it has at most one right special factor of each length and if the set of its factors is closed under reversal. We refer the reader to [7] for basic properties of episturmian words.

Proposition 2.8. *Fix a positive integer $k \geq 2$. Let u and v be factors of the same episturmian word ω. Then u and v are k-Abelian equivalent if and only if u and v are $(k-1)$-Abelian equivalent and share a common prefix and a common suffix of length* $\min\{|u|, k-1\}$. *Thus, u and v are k-Abelian equivalent if and only if u and v are Abelian equivalent and share a common prefix and a common suffix of length* $\min\{|u|, k-1\}$.

Proof. One direction follows immediately from Lemma 2.3. Next suppose that u and v are $(k-1)$-Abelian equivalent factors of the same episturmian word ω, and that u and v share a common prefix and a common suffix of length $\min\{|u|, k-1\}$. To prove that $u \sim_k v$ it suffices to show that whenever $axb \in \mathcal{F}_\omega(k)$ (with $a, b \in A$ and $x \in A^*$), we have $|u|_{axb} = |v|_{axb}$. First let us suppose that ax is not a right special factor of ω so that every occurrence in ω of ax is a occurrence of axb. Then, if ax is not a suffix of u (and hence not a suffix of v) we obtain

$$|u|_{axb} = |u|_{ax} = |v|_{ax} = |v|_{axb}.$$

On the other hand if ax is a suffix of u (and hence also a suffix of v) we have

$$|u|_{axb} = |u|_{ax} - 1 = |v|_{ax} - 1 = |v|_{axb}.$$

Similarly, in case xb is not a left special factor of ω we obtain $|u|_{axb} = |v|_{axb}$. Thus it remains to consider the case when ax is right special in ω and xb is left special in ω. In this case x is bispecial and $a = b$. For each $c \in A$, let $n_c = |u|_{axc}$ and $n'_c = |v|_{axc}$. We must show that $n_a = n'_a$. However we know that $n_c = n'_c$ for all $c \neq a$ since xc is not left special in ω. Now, if ax is not a suffix of u (and hence not a suffix of v) we have

$$\sum_{c \in A} n_c = |u|_{ax} = |v|_{ax} = \sum_{c \in A} n'_c$$

whence $n_a = n'_a$. On the other hand if ax is a suffix of u (and hence a suffix of v) then

$$\sum_{c \in A} n_c = |u|_{ax} - 1 = |v|_{ax} - 1 = \sum_{c \in A} n'_c$$

whence $n_a = n'_a$ as required. □

Remark 2.9. The following example illustrates that the assumption in Proposition 2.8 that u and v are factors of the same Sturmian word is necessary: Let $u = aabb$ and $v = abab$. The u and v are Abelian equivalent and share a common prefix and suffix of length 1, yet they are not 2-Abelian equivalent.

Proof of Theorem 2.6. We start by showing that if $\omega \in \{a, b\}^{\mathbb{N}}$ is a Sturmian word, and x a right special factor of ω of length $k-1$, then $u = xab\bar{x}$ and $v = xba\bar{x}$ are k-Abelian equivalent. This follows from Proposition 2.8 since u and v share a common prefix and a common suffix of lengths $k-1$ and are Abelian equivalent.

Next we suppose that u and v are distinct k-Abelian equivalent words of length $2k$ and show that both u and v have the required form. We proceed by induction on k. In case $k = 1$, we have that u and v are distinct Abelian equivalent words of length 2 whence u and v may be written in the form $u = ab$ and $v = ba$ for some $a \neq b$ in A.

Next suppose the result of Theorem 2.6 is true for $k-1$ and we shall prove it for k. So let u and v be distinct k-Abelian equivalent words of length $2k$ with $k > 1$. Then by Lemma 2.3 we can write $u = a'u'b'$ and $v = a'v'b'$ for some $a', b' \in A$ and $u', v' \in A^*$ where $|u'| = |v'| = 2(k-1) \geq 2$. Since u and v are distinct, it follows that $u' \neq v'$. Also, by Lemma 2.5 it follows that $u' \sim_{k-1} v'$. Thus by induction hypothesis, there exist distinct letters $a, b \in A$ and a Sturmian word $\omega \in \{a, b\}^{\mathbb{N}}$ such that u' and v' are both factors of ω of the form $u' = xab\bar{x}$ and $v' = xba\bar{x}$ for some right special factor x of ω of length $k-2$.

Thus we can write $u = a'xab\bar{x}b'$ and $v = a'xba\bar{x}b'$. Since $u \sim_k v$, $|a'xa| = k$, and $a \neq b$ it follows that $a'x$ must occur in v' and hence $a' \in \{a, b\}$. Similarly we deduce that $b' \in \{a, b\}$.

Let us first suppose that $x \neq \bar{x}$. Then $a'xa$ must occur in v' and $a\bar{x}b'$ must occur in u'. Hence both $a'xa$ and $a\bar{x}b'$ are factors of ω. Moreover, since $x \neq \bar{x}$ it follows that x is not left special in ω and $\bar{x}$ is not right special in ω. Hence every occurrence of x in ω is preceded by a' and every occurrence of $\bar{x}$ is ω is followed by b'. Since the factors of ω are closed under reversal, we deduce that $a' = b'$ and $a'x$ is a right special factor of ω. Moreover, since u' and v' are both factors of ω beginning in x and ending in $\bar{x}$, it follows that $u = a'xab\bar{x}a'$ and $v = a'xba\bar{x}a'$ are both factors of ω.

Finally suppose $x = \bar{x}$ so that x is a bispecial factor of ω. We may write the increasing sequence of bispecial factors $\varepsilon = B_0, B_1, \ldots, x = B_n, B_{n+1}, \ldots$ so that x is the nth bispecial factor of ω. We recall that associated to ω is a sequence $(a_i)_{i\geq 0} \in A^{\mathbb{N}}$ (called the *directive word* of ω) defined by a_iB_i is right special in ω. (See for instance [30]).

Without loss of generality we can suppose that $a' = a$. We claim $b' = a$. Suppose to the contrary that $b' = b$. Then both axa and $b\bar{x}b = bxb$ are factors of v' contradicting that ω is balanced. Hence we must have $a' = b' = a$ and so $u = axab\bar{x}a$ and $v = axba\bar{x}a$. Now x is a bispecial factor of the Sturmian word ω. If ax is a right special factor of ω then we are done by Remark 2.7. Otherwise, if bx is a right special factor of ω, then this means that $a_n = b$ where a_n is the nth entry of the directive word of ω. Let ω' be a Sturmian word whose directive word $(b_i)_{i\geq 0}$ is defined by $b_i = a_i$ for $i \neq n$, and $b_n = a$. Then x is a bispecial factor of ω' and ax is a right special factor of ω'. It follows from Remark 2.7 that both u and v are factors of ω'. □

As an immediate consequence of Theorem 2.6 we have:

Corollary 2.10. *Let $u \in A^*$ be of the form $u = vxab\bar{x}w$ where x is a right special factor of length $k-1$ of a Sturmian word. Set $u' = vxba\bar{x}w$. Then $u \sim_k u'$.*

2.3. The number of k-Abelian classes in A^n

Here we shall estimate the number of k-Abelian equivalence classes of words in A^n. Fix $k \geq 1$ and let $m \geq 2$ be the cardinality of the set A.

Lemma 2.11. *The number of k-Abelian equivalence classes of A^{n+1} is at least as large as the number of k-Abelian equivalence classes of A^n.*

Proof. If $k = 1$ or $n < k-1$, then the claim is clear. Otherwise, let B be a set of representatives of the k-Abelian equivalence classes of A^n. The set AB has m times as many words as B. To prove the Lemma, we will show that there can be at most m words in AB that are k-Abelian equivalent.

Let $a \in A$ and let $au_0, \ldots au_m \in AB$ be k-Abelian equivalent. It needs to be shown that some of these words are equal. Two of these words must have the same kth letter, let these be au and av. Because also $\mathrm{pref}_{k-1}(au) = \mathrm{pref}_{k-1}(av)$, it follows that $\mathrm{pref}_k(au) = \mathrm{pref}_k(av)$. If $t \in A^k$, then either $|u|_t = |au|_t = |av|_t = |v|_t$ (if $t \neq \mathrm{pref}_k(au)$), or $|u|_t = |au|_t - 1 = |av|_t - 1 = |v|_t$ (if $t = \mathrm{pref}_k(au)$). Thus u and v are k-Abelian equivalent and, by the definition of B, $u = v$. This proves the claim. □

For $s_1, s_2 \in A^{k-1}$, let

$$S(s_1, s_2, n) = A^n \cap s_1A^* \cap A^*s_2$$

be the set of words of length n that start with s_1 and end with s_2. For every word $w \in S(s_1, s_2, n)$ we can define a function

$$f_w : A^k \to \{0, \ldots, n-k+1\}, \ f_w(t) = |w|_t.$$

If $u, v \in S(s_1, s_2, n)$, then $u \sim_k v$ if and only if $f_u = f_v$. To count the number of k-Abelian equivalence classes, we need to count the number of the functions f_w. Not every function $f : A^k \to \{0, \ldots, n-k+1\}$ is possible. It must be

$$\sum_{t\in A^k} f(t) = n - k + 1, \tag{2.2}$$

and there are also other restrictions, which are determined in Lemma 2.12.

If a function $f : A^k \to \mathbb{N}$ is given, then a directed multigraph G_f can be defined as follows: The set of vertices is A^{k-1}, and if $t = pa = bq$, where $a, b \in A$ and $p, q \in A^{k-1}$, then there are $f(t)$ edges from p to q. If $f = f_w$, then this multigraph is related to the Rauzy graph of w. For more on Rauzy graphs, see, e.g., [11]. In the next lemma, $\deg^-$ denotes the indegree and $\deg^+$ the outdegree of a vertex in G_f.

Lemma 2.12. *For a function $f : A^k \to \mathbb{N}$ and words $u, v \in A^{k-1}$, the following are equivalent:*

(i) *There is a number n and a word $w \in S(u, v, n)$ such that $f = f_w$.*

(ii) *There is an Eulerian path from u to v in G_f.*

(iii) *The underlying graph of G_f is connected, except possibly for some isolated vertices, and $\deg^-(s) = \deg^+(s)$ for every vertex s, except that if $u \neq v$, then $\deg^-(u) = \deg^+(u) - 1$ and $\deg^-(v) = \deg^+(v) + 1$.*

(iv) *The underlying graph of G_f is connected, except possibly for some isolated vertices, and*

$$\sum_{a \in A} f(as) = \sum_{a \in A} f(sa) + c_s \qquad (s \in A^{k-1}), \tag{2.3}$$

where

$$c_s = \begin{cases} -1 & \text{if } s = u \neq v \\ 1 & \text{if } s = v \neq u \\ 0 & \text{otherwise.} \end{cases}$$

Proof. (i) ⇔ (ii): $w = a_1 \dots a_n \in S(u, v, n)$ and $f = f_w$ if and only if

$$u = a_1 \dots a_{k-1} \to a_2 \dots a_k \to \cdots \to a_{n-k+2} \dots a_n = v$$

is an Eulerian path in G_f.

(ii) ⇔ (iii): This is well known.

(iii) ⇔ (iv): (iv) is just a reformulation of (iii) in terms of the function f. □

In the next lemma we consider the independence of homogeneous systems related to the equations (2.3) and (2.2).

Lemma 2.13. *Let x_t, where $t \in A^k$, be m^k unknowns. The system of equations*

$$\sum_{a \in A} x_{as} = \sum_{a \in A} x_{sa} \qquad (s \in A^{k-1}) \tag{2.4}$$

is not independent, but all of its proper subsystems are. If we add the equation

$$\sum_{t \in A^k} x_t = 0 \tag{2.5}$$

to one of these independent systems, then the system remains independent.

Proof. The sum of the equations (2.4) is a trivial identity $\sum_{t \in A^k} x_t = \sum_{t \in A^k} x_t$, so every one of these equations follows from the other $m^{k-1} - 1$ equations. If $s_1, s_2 \in A^{k-1}$ are two different words, then $x_t = |s_1 s_2|_t$ for all t is a solution of all the equations, except those with $s = s_1$ or $s = s_2$. This proves that all subsystems are independent. Addition of (2.5) keeps them independent, because $x_t = 1$ for all t is a solution of the system (2.4) but not of (2.5). □

Theorem 2.14. *Let $k \geq 1$ and $m \geq 2$ be fixed numbers and let A be an m-letter alphabet. The number of k-Abelian equivalence classes of A^n is $\Theta(n^{m^k - m^{k-1}})$.*

Proof. Let $n \geq 2k-2$, $f : A^k \to \{0, \dots, n-k+1\}$ and $u, v \in A^{k-1}$. By Lemma 2.12, there is a word $w \in S(u,v,n)$ such that $f = f_w$ only if f satisfies (2.2) and (2.3). Consider the system formed by these equations. The function f_w satisfies the equations for every $w \in S(u,v,n)$, so the system has a solution. By Lemma 2.13, the rank of the coefficient matrix of the system is m^{k-1}, so the general solution of this system is of the form

$$f(r_i) = \sum_{j=1}^{m^k - m^{k-1}} a_{ij} f(s_j) + b_i \qquad (i = 1, \dots, m^{k-1}),$$

where the words r_i and s_j form the set A^k and a_{ij}, b_i are rational numbers. Because $0 \leq f(s_j) \leq n-k+1$, there are $O(n^{m^k - m^{k-1}})$ possible functions f.

Let $u = v$ and consider the system of equations (2.3). By Lemma 2.13, the general solution of this homogeneous system is of the form

$$f(r_i) = \sum_{j=1}^{m^k - m^{k-1}+1} a_{ij} f(s_j) \qquad (i = 1, \dots, m^{k-1}-1), \tag{2.6}$$

where the words r_i and s_j form the set A^k and a_{ij} are rational numbers. The coefficients a_{ij} do not depend on n. Let

$$c = \max\left\{ \textstyle\sum_{j=1}^{m^k - m^{k-1}+1} |a_{ij}| \mid 1 \leq i \leq m^{k-1}-1 \right\}$$

and let d be the least common multiple of the denominators of the numbers a_{ij}. Every constant function f satisfies the system of equations. In particular, $f(t) = \lfloor n/2m^k \rfloor$ for all t is a solution of the system and thus

$$\left\lfloor \frac{n}{2m^k} \right\rfloor = \sum_{j=1}^{m^k - m^{k-1}+1} a_{ij} \left\lfloor \frac{n}{2m^k} \right\rfloor \qquad (i = 1, \dots, m^{k-1}-1). \tag{2.7}$$

Let

$$f(s_j) = \left\lfloor \frac{n}{2m^k} \right\rfloor + b_j, \tag{2.8}$$

where b_j can be any integers such that

$$|b_j| < \frac{n}{2cm^k} - 1 \quad \text{and} \quad d | b_j.$$

Since c, m, k and d do not depend on n, the number of possible values for each b_j is $\Theta(n)$. Because we are interested in the asymptotic behavior as n grows, we can assume that $n > 2cm^k$. The numbers

$$f(r_i) = \left\lfloor \frac{n}{2m^k} \right\rfloor + \sum_{j=1}^{m^k - m^{k-1}+1} a_{ij} b_j$$

given by (2.6), (2.7) and (2.8) are integers and, moreover,

$$\Big| \sum_{j=1}^{m^k - m^{k-1}+1} a_{ij} b_j \Big| \le \sum_{j=1}^{m^k - m^{k-1}+1} |a_{ij} b_j| \le \sum_{j=1}^{m^k - m^{k-1}+1} |a_{ij}| \left(\frac{n}{2cm^k} - 1 \right)$$
$$\le c \left(\frac{n}{2cm^k} - 1 \right) \le \frac{n}{2m^k} - 1.$$

Thus $1 \le f(t) \le n/m^k - 1$ for all $t \in A^k$. Because $f(t) \ge 1$ for all t, the underlying graph of G_f is connected, so by Lemma 2.12 there is a word $w \in S(u, v, |w|)$ such that $f = f_w$. Because $f(t) \le n/m^k - 1$ for all t, we get

$$|w| = \sum_{t \in A^k} f(t) + k - 1 \le n - m^k + k - 1 < n.$$

There are $\Theta(n^{m^k - m^{k-1}+1})$ ways to choose the numbers b_j. Every choice gives a different function $f = f_w$ for some $w \in S(u, v, |w|)$ such that $|w| < n$. Let these words be $w_1, \ldots, w_N$. No two of them are k-Abelian equivalent. Among these words there are at least N/n words of equal length. By Lemma 2.11, there are at least N/n words of length n such that no two of them are k-Abelian equivalent, and $N/n = \Omega(n^{m^k - m^{k-1}})$. □

3. *k*-Abelian complexity & periodicity

In this section we prove that if $\mathcal{P}^{(k)}_\omega(n_0) < q^{(k)}(n_0)$ for some $k \in \mathbb{Z}^+ \cup \{+\infty\}$ and $n_0 \ge 1$, then ω is ultimately periodic (see Corollary 3.3 below). For this purpose we introduce an auxiliary family of equivalence relations $\mathcal{R}_k$ on A^* defined as follows: Let $k \in \mathbb{Z}^+ \cup \{+\infty\}$. Give $u, v \in A^*$ we write $u\mathcal{R}_k v$, if and only if $u \sim_1 v$ (i.e., $u \sim_{\mathrm{ab}} v$) and u and v share a common prefix and a common suffix of lengths $k-1$. In case $|u| < k-1$, then $u\mathcal{R}_k v$ means $u = v$.

It follows immediately from Lemma 2.3 that

$$u \sim_k v \Longrightarrow u\mathcal{R}_k v. \tag{3.1}$$

In general the converse is not true: For example, taking $u = 0011$ and $v = 0101$ we see that $u\mathcal{R}_2 v$ yet u and v are not 2-Abelian equivalent. However, in view of Proposition 2.8 we have:

Corollary 3.1. *Let u and v be two factors of a Sturmian word ω, and $k \in \mathbb{Z}^+ \cup \{+\infty\}$. Then $u \sim_k v$ if and only if $u\mathcal{R}_k v$.*

Let $\omega \in A^{\mathbb{N}}$. Associated to the relation $\mathcal{R}_k$ is a complexity function, denoted $\rho^{(k)}_\omega(n)$, which counts the number of distinct $\mathcal{R}_k$ equivalence classes of factors of ω of length n. It follows from (3.1) above that for each n we have

$$\rho^{(k)}_\omega(n) \le \mathcal{P}^{(k)}_\omega(n). \tag{3.2}$$

We recall the function $q^{(k)} : \mathbb{N} \to \mathbb{N}$ ($k \in \mathbb{Z}^+ \cup \{+\infty\}$) defined by

$$q^{(k)}(n) = \begin{cases} n+1 & \text{for } n \le 2k-1 \\ 2k & \text{for } n \ge 2k. \end{cases}$$

Theorem 3.2. *Let $\omega = a_0a_1a_2\ldots \in A^{\mathbb{N}}$ and $k \in \mathbb{Z}^+ \cup \{+\infty\}$. If $\rho_\omega^{(k)}(n_0) < q^{(k)}(n_0)$ for some $n_0 \geq 1$, then ω is ultimately periodic.*

Proof. The result is well known in case $k = +\infty$ (see [24]). For $k \in \mathbb{Z}^+$, we proceed by induction on k. In case $k = 1$, then $\mathcal{R}_1$ is simply the usual notion of Abelian equivalence and the result follows from [4].

Now suppose $k > 1$ and that $\rho_\omega^{(k)}(n_0) < q^{(k)}(n_0)$ for some $n_0 \geq 1$. It follows immediately from the definition of $\mathcal{R}_k$ that if $u\mathcal{R}_kv$ and $|u| \leq 2k-1$, then $u = v$. Thus, if $\rho_\omega^{(k)}(n_0) < q^{(k)}(n_0)$ where $n_0 \leq 2k-1$, then $\rho_\omega(n_0) < n_0 + 1$ and so ω is ultimately periodic by the well known result of Morse and Hedlund in [24].

Thus we suppose that $\rho_\omega^{(k)}(n_0) < 2k$ for some $n_0 \geq 2k$. We claim that ω must be ultimately periodic. Suppose to the contrary that ω is aperiodic. We shall show that this implies that $\rho_\nu^{(k-1)}(n_0-2) < 2(k-1)$ where $\nu = a_0^{-1}\omega$ denotes the first shift of ω, i.e., the word obtained from ω by removing the first letter of ω. Since $n_0 - 2 \geq 2(k-1)$ we deduce that $\rho_\nu^{(k-1)}(n_0-2) < q^{(k-1)}(n_0-2)$. But then by induction hypothesis on k, it follows that ν (and hence ω) is ultimately periodic, a contradiction.

Consider the map

$$\Psi : \mathcal{F}_\omega(n_0)/\mathcal{R}_k \longrightarrow \mathcal{F}_\nu(n_0-2)/\mathcal{R}_{k-1}$$

defined by

$$\Psi([aub]_k) = [u]_{k-1}$$

where $a, b \in A$, and $u \in A^*$ of length $n_0 - 2$. Here $[u]_k$ denotes the $\mathcal{R}_k$ equivalence class of u. To see that Ψ is well defined, suppose $aub\mathcal{R}_kcud$. Then since $k > 1$, it follows that $a = c$ and $b = d$ and thus that $u\mathcal{R}_1v$. Moreover as aub and cud share a common prefix and suffix of length k, it follows that u and v share a common prefix and suffix of length $k-1$. Thus $u\mathcal{R}_{k-1}v$ as required. Clearly the mapping Ψ is surjective, in fact for each $u \in \mathcal{F}_\nu(n_0-2)$ there exist $a, b \in A$ such that $aub \in \mathcal{F}_\omega(n_0)$. This is the reason for replacing ω by ν.

We now show that either there exist distinct classes $[u]_{k-1}, [v]_{k-1} \in \mathcal{F}_\nu(n_0-2)/\mathcal{R}_{k-1}$ for which

$$\min\{\mathrm{Card}\left(\Psi^{-1}([u]_{k-1})\right), \mathrm{Card}\left(\Psi^{-1}([v]_{k-1})\right)\} \geq 2, \tag{3.3}$$

or there exists a class $[u]_{k-1} \in \mathcal{F}_\nu(n_0-2)/\mathcal{R}_{k-1}$ for which

$$\mathrm{Card}\left(\Psi^{-1}([u]_{k-1})\right) \geq 3. \tag{3.4}$$

In either case it follows that

$$\mathrm{Card}\,(\mathcal{F}_\nu(n_0-2)/\mathcal{R}_{k-1}) \leq \mathrm{Card}\,(\mathcal{F}_\omega(n_0)/\mathcal{R}_k) - 2 < 2(k-1).$$

Since ω is assumed to be aperiodic, ω contains both a left special factor of the form uc and a right special factor of the form dv of lengths $n_0 - 1$ for some choice of $c, d \in A$ and $u, v \in A^*$. Thus there exist distinct letters $a, b \in A$ such that auc and buc are factors of ω. Moreover since $a \neq b$, it follows that $[auc]_k \neq [buc]_k$. Thus $\mathrm{Card}\left(\Psi^{-1}([u]_{k-1})\right) \geq 2$. Similarly, there exist distinct letters $a', b' \in A$ such that dva' and dvb' are factors of ω, and since $a' \neq b'$, it follows that $[dva']_k \neq [dvb']_k$. Thus $\mathrm{Card}\left(\Psi^{-1}([v]_{k-1})\right) \geq 2$. In case $[u]_{k-1} \neq [v]_{k-1}$, we obtain the desired inequality (3.3). In case $[u]_{k-1} = [v]_{k-1}$, since $a \neq b$ and $a' \neq b'$ it follows that

$$\mathrm{Card}\{[auc]_k, [buc]_k, [dua']_k, [dub']_k\} \geq 3$$

which yields the inequality (3.4). This completes the proof of Theorem 3.2 □

Corollary 3.3. *Let $\omega \in A^{\mathbb{N}}$ and $k \in \mathbb{Z}^+ \cup \{+\infty\}$. If $\mathcal{P}_\omega^{(k)}(n_0) < q^{(k)}(n_0)$ for some $n_0 \geq 1$ then ω is ultimately periodic.*

Proof. As a consequence of the inequality (3.2), if $\mathcal{P}_\omega^{(k)}(n_0) < q^{(k)}(n_0)$ then $\rho_\omega^{(k)}(n_0) < q^{(k)}(n_0)$, whence by Theorem 3.2 it follows that ω is ultimately periodic. □

The same method of proof of Theorem 3.2 can be used to prove the following:

Corollary 3.4. *Let ω be a bi-infinite word over the alphabet A and $k \in \mathbb{Z}^+ \cup \{+\infty\}$. If $\mathcal{P}_\omega^{(k)}(n_0) < q^{(k)}(n_0)$ for some $n_0 \geq 1$, then ω is periodic.*

We conclude this section with a few remarks:

Remark 3.5. In the special case $k = +\infty$, the condition given in Corollary 3.3 gives a characterization of ultimately periodic words by means of factor complexity: $\omega \in A^{\mathbb{N}}$ is ultimately periodic if and only if $\rho_\omega(n_0) < n_0+1$ for some $n_0 \geq 1$. However, k-Abelian complexity does not yield such a characterization. Indeed, both Sturmian words and the ultimately periodic word $01^\infty = 0111\cdots$ have the same Abelian complexity. More generally, the ultimately periodic word $0^{2k-1}1^\infty \ldots$ has the same k-Abelian complexity as a Sturmian word (see Theorem 4.1 below).

Remark 3.6. The result of Corollary 3.4 is already known to be true in the special cases $k = +\infty$ (see [24]) and $k = 1$ (see Remark 4.07 in [4]). In these special cases, the converse is also true. But for general $2 \leq k < +\infty$ the converse is false. For instance, let $\mathrm{Card}(A) = 5$, and let u be a word containing at least one occurrence of every $x \in A^3$. Let ω be the periodic word $\omega = \ldots uuuu \ldots$. Then $\rho_\omega^{(2)}(n) \geq 5$ for every $n \geq 1$.

4. *k*-Abelian complexity of Sturmian words

In this section we determine the k-Abelian complexity of Sturmian words and show that for each k, the complexity function $\mathcal{P}^{(k)}$ completely characterizes Sturmian words amongst all aperiodic words. More precisely:

Theorem 4.1. *Fix $k \in \mathbb{Z}^+ \cup \{+\infty\}$. Let $\omega \in A^{\mathbb{N}}$ be an aperiodic word. The following conditions are equivalent:*

- *ω is a balanced binary word, that is, Sturmian.*
- $\mathcal{P}_\omega^{(k)}(n) = q^{(k)}(n) = \begin{cases} n+1 & \text{for } 0 \leq n \leq 2k-1 \\ 2k & \text{for } n \geq 2k. \end{cases}$

Our proof of Theorem 4.1 will make use of the following functions g_i, which transform binary words by changing the letters around a specific point. For words $w \in \{0,1\}^n$ we define $g_1, \ldots, g_n$ as follows:

$$g_i(w) = \begin{cases} u10v & \text{if } i < n,\ w = u01v \text{ and } |u0| = i \\ u1 & \text{if } i = n \text{ and } w = u0. \end{cases}$$

Lemma 4.2. *Let $n \geq 1$ and let $w \in \{0,1\}^{\mathbb{N}}$ be Sturmian. There is a word $u_1 \in \{0,1\}^n$ and a permutation σ of $\{1,\ldots,n\}$ such that if $u_{i+1} = g_{\sigma(i)}(u_i)$ for $i = 1,\ldots,n$, then $u_1,\ldots,u_{n+1}$ are the factors of w of length n.*

Proof. Let $u_1,\ldots,u_{n+1}$ be the factors of w of length n in lexicographic order. If follows from Theorem 1.1. in [2] that for every i there is an m such that $u_{i+1} = g_m(u_i)$. It needs to be proved that the m's are all different. Let $u_{i+1} = g_m(u_i)$ and $u_{i'+1} = g_m(u_i')$. For every j

$$|\mathrm{pref}_m(u_j)|_1 \leq |\mathrm{pref}_m(u_{j+1})|_1$$

and for $j \in \{i,i'\}$

$$|\mathrm{pref}_m(u_j)|_1 < |\mathrm{pref}_m(u_{j+1})|_1.$$

If $i \neq i'$, then

$$|\mathrm{pref}_m(u_1)|_1 + 2 \leq |\mathrm{pref}_m(u_{n+1})|_1$$

which contradicts the balance property of Sturmian words. □

Example 4.3. *The factors of the Fibonacci word of length six are*

$$u_1 = 001001, \quad u_2 = 001010 = g_5(u_1), \quad u_3 = 010010 = g_2(u_2), \quad u_4 = 010100 = g_4(u_3),$$
$$u_5 = 100100 = g_1(u_4), \quad u_6 = 100101 = g_6(u_5), \quad u_7 = 101001 = g_3(u_6).$$

We have $u_2 \sim_2 u_3 \sim_2 u_4$ and $u_6 \sim_2 u_7$. There are no other 2-Abelian equivalences between these factors.

Proof of Theorem 4.1. First let us suppose $\omega \in \{0,1\}^{\mathbb{N}}$ is Sturmian and let $1 \leq k \leq +\infty$. Let $n \leq 2k-1$. By Lemma 2.4 two factors u and v of ω of length n are k-Abelian equivalent if and only $u = v$. Thus $\mathcal{P}_w^{(k)}(n) = n+1$ as required.

Next let $n \geq 2k$ and let $u_1,\ldots,u_{n+1}$ and σ be as in Lemma 4.2. If $k \leq \sigma(i) \leq n-k$, then there are words $s,t \in \{0,1\}^*$ and $u,v \in \{0,1\}^{k-1}$ and letters $a,b \in \{0,1\}$ so that $u_i = su01vt$ and $u_{i+1} = g_{\sigma(i)}(u_i) = su10vt$. We prove that $u_i \sim_k u_{i+1}$. The prefixes and suffixes of u_i and u_{i+1} of length $k-1$ are the same. The factors of u_i of length k are the factors of su, $u01v$ and vt of length k, and the factors of u_{i+1} of length k are the factors of su, $u10v$ and vt of length k. Because $u01v$ and $u10v$ are factors of w, it follows that u is right special and v is left special and hence equal to the reversal of u. By Theorem 2.6, $u01v$ and $u10v$ are k-Abelian equivalent. This proves that $u_i \sim_k u_{i+1}$ if $k \leq \sigma(i) \leq n-k$. Thus the words $u_1,\ldots,u_{n+1}$ are in at most $2k$ different k-Abelian equivalence classes and $\mathcal{P}_\omega^{(k)}(n) \leq 2k$. By Corollary 3.3, $\mathcal{P}_\omega^{(k)}(n) = 2k$.

Next let $1 \leq k \leq +\infty$ and let $\omega \in A^{\mathbb{N}}$ be aperiodic and

$$\mathcal{P}_\omega^{(k)}(n) = q^{(k)}(n) = \begin{cases} n+1 & \text{for } 0 \leq n \leq 2k-1 \\ 2k & \text{for } n \geq 2k. \end{cases}$$

Taking $n = 1$ we see that ω is binary, (say $\omega \in \{0,1\}^{\mathbb{N}}$). We must show that ω is balanced. We first recall some basic facts concerning factors of Sturmian words (see for instance [30]): Let $\eta \in \{0,1\}^{\mathbb{N}}$ be a Sturmian word, and let $\mathcal{F}_\eta(n)$ denote the factors of η of length n. The set $\mathcal{F}_\eta(n+1)$ is completely determined from the set $\mathcal{F}_\eta(n)$ unless η has a bispecial factor B of length $n-1$ in which case both $0B$ and $1B$ are factors of η and exactly one of the two is right special. If $0B$ is right special, then every occurrence of $1B$ in η is an occurrence of $1B0$. If v is a factor of η and u a prefix of v, we write $u \vdash v$ if every occurrence of u in η is an occurrence of v. Thus if $0B$ is right special, then $1B \vdash 1B0$, and similarly if $1B$ is right special, then $0B \vdash 0B1$.

Now suppose to the contrary that the aperiodic binary word ω is not Sturmian. Then there exists a smallest positive integer $n \geq 1$ and a Sturmian word η such that $\mathcal{F}_\omega(n) = \mathcal{F}_\eta(n)$ but $\mathcal{F}_\omega(n+1) \neq \mathcal{F}_{\eta'}(n+1)$ for every choice of Sturmian word η'. This means that ω has a bispecial factor B of length $n-1$ and both $0B$ and $1B$ are in $\mathcal{F}_\omega(n)$ and one of the following must occur: i) Neither $0B$ nor $1B$ is right special in ω; ii) There exists a unique $a \in \{0,1\}$ such that aB is right special, and $(1-a)B \vdash (1-a)B(1-a)$; iii) Both $0B$ and $1B$ are right special in ω. We will show that since ω is aperiodic, only case iii) is in fact possible. Clearly, if neither $0B$ nor $1B$ were right special, then $\mathrm{Card}(\mathcal{F}_\omega(n)) = \mathrm{Card}(\mathcal{F}_\omega(n+1))$ whence ω is ultimately periodic, a contradiction. Next suppose case ii) occurs. We may suppose without loss of generality that $0B$ is right special and $1B \vdash 1B1$. If $1 \vdash 1B$ (and hence $1 \vdash 1B1$), then we would have $1 \vdash 1(B1)^n$ for every $n \geq 1$ from which it follows that the tail of ω corresponding to the first occurrence of 1 on ω is periodic. Thus if $\neg(1 \vdash 1B)$, then there exists a bispecial factor B' of ω with $0 < |B'| < |B|$ such that $1B'$ is right special and $1B'1 \vdash 1B$ and hence $1B'1 \vdash 1B1$. Writing $1B1 = 1B'1V$ we have $1B'1 \vdash 1B'1V$. We next show by induction on n that $1B'1V^n$ is a palindrome for each $n \geq 1$. Clearly this is true for $n = 1$ since $1B'1V = 1B1$. Next suppose $1B'1V^n$ is a palindrome. Then

$$\overline{1B'1V^{n+1}} = \overline{V}^{n+1}\overline{1B'1} = \overline{V}\,\overline{V}^n\overline{1B'1} = \overline{V}1B'1V^n = \overline{V}\,\overline{1B'1}\,V^n = 1B'1VV^n = 1B'1V^{n+1}.$$

Having established that $1B'1V^n$ is a palindrome, it follows that $1B'1$ is a suffix of $1B'1V^n$ and hence $1B'1V^n \vdash 1B'1V^{n+1}$ for each $n \geq 0$. Whence as before ω is ultimately periodic. Thus if ω is not Sturmian, case iii) must occur. This implies that

$$\mathcal{F}_\omega(n+1) = \mathcal{F}_\eta(n+1) \cup \{0B0, 1B1\}$$

and $\mathrm{Card}(\mathcal{F}_\eta(n+1) \cap \{0B0, 1B1\}) = 1$. Since η is Sturmian, the number of k-Abelian classes of factors of η of length $n+1$ is equal to $q^{(k)}(n+1)$. But the additional factor aBa of ω of length $n+1$ introduces a new k-Abelian class since it is not even Abelian equivalent to any other factor of η (and hence ω) of length $n+1$. Thus $\mathcal{P}^{(k)}_\omega(n+1) = q^{(k)}(n+1) + 1$, a contradiction. Thus ω is Sturmian. □

Remark 4.4. In view of Corollary 3.3, within the class of aperiodic words, Sturmian words have the lowest possible k-Abelian complexity. See [1, 15, 16, 24] for other instances in which Sturmian words have the lowest complexity amongst all aperiodic words.

5. Bounded k-Abelian complexity & k-Abelian repetitions

There is great interest in avoidability of repetitions in infinite words. This originated with the classical work of Thue [33] and [34], in which he established the existence of an infinite binary (resp. ternary) word avoiding cubes (resp. squares). It was later shown that to avoid Abelian cubes or Abelian squares, one needs 3-letter or 4-letter alphabets respectively (see [6] and [19]). The corresponding problems for k-abelian repetitions turned out to be quite nontrivial. It follows easily that the smallest alphabet where k-abelian cubes can be avoided is either 2 or 3, and similarly the smallest alphabet where k-abelian squares can be avoided is either 3 or 4. In the latter case for $k = 2$ a computer verification revealed that the correct value is 4, as in the case of Abelian repetitions: Each ternary 2-abelian square-free word is of length at most 537 [13]. In the case of binary alphabets and cubes it was shown in a sequence of papers that an infinite word avoiding k-abelian cubes can be constructed for $k = 8$, $k = 5$ and for $k = 3$ (see [14], [21] and [22] respectively). The only remaining case $k = 2$ has apparently been solved recently in [27], where the existence of a 2-abelian cube-free binary word is proved. For avoiding k-abelian squares in a ternary alphabet the situation has been equally challenging. The avoidability in infinite words of k-abelian squares in a ternary alphabet was first proved for large values of k ($k \geq 64$) [12]. A construction and proof for $k = 3$ can be found in [27].

In this section we prove that k-Abelian repetitions are unavoidable in words having bounded k-Abelian complexity. For each positive integer k we set

$$A^{\leq k} = \{x \in A^* \, : \, |x| \leq k\}.$$

Given an infinite word $\omega = a_0a_1a_2 \ldots \in A^{\mathbb{N}}$, for each $0 \leq i < j < +\infty$ we denote by $\omega[i,j]$ the factor $a_ia_{i+1} \cdots a_j$.

Definition 5.1. *Let k and B be positive integers and $\omega \in A^{\mathbb{N}}$. We say ω is (k, B)-balanced if and only if for all factors u and v of ω of equal length, and for all $x \in A^{\leq k}$ we have $||u|_x - |v|_x| \leq B$. We say ω is arbitrarily k-imbalanced if ω is not (k, B)-balanced for any positive integer B.*

An elementary, but key observation is that

Lemma 5.2. *Let k be a positive integer and $\omega \in A^{\mathbb{N}}$. Then ω has bounded k-Abelian complexity if and only if ω is (k, B)-balanced for some positive integer B.*

Proof. Clearly if $\mathcal{P}^{(k)}_{\omega}$ is bounded, say by B, then ω is $(k, B-1)$-balanced. Conversely, if ω is (k, B)-balanced, then for each positive integer n and for each $x \in A^*$ with $|x| \leq k$ we have

$$\text{Card}\{|u|_x : \, u \in \mathcal{F}_{\omega}(n)\} \leq B + 1.$$

It follows that

$$\mathcal{P}^{(k)}_{\omega}(n) \leq (B + 1)^K$$

where $K = \text{Card}(A^{\leq k})$. □

Fix a positive integer k. It follows from Theorem 4.1 and Lemma 5.2 that each Sturmian word is (k, B)-balanced for some positive integer B (depending on k). Actually, I. Fagnot and L. Vuillon proved in [9] that every Sturmian word is (k, k)-balanced.

Definition 5.3. *Fix $k \in \mathbb{Z}^+ \cup \{+\infty\}$, and N a positive integer. By a k-Abelian N-power we mean a word U of the form $U = U_1U_2 \cdots U_N$ such that $U_i \sim_k U_j$ for all $1 \leq i, j \leq N$.*

In this section we shall prove the following result:

Theorem 5.4. *Fix $k \in \mathbb{Z}^+ \cup \{+\infty\}$. Let $\omega = a_0a_1a_2\ldots \in A^{\mathbb{N}}$ be an infinite word on a finite alphabet A having bounded k-Abelian complexity. Let $D \subseteq \mathbb{N}$ be a set of positive upper density, that is*

$$\limsup_{n\to\infty} \frac{\operatorname{Card}(D \cap \{1,2,\ldots,n\})}{n} > 0.$$

Then, for every positive integer N, there exist i and ℓ such that $\{i, i+\ell, i+2\ell, \ldots, i+\ell N\} \subset D$ and the N consecutive blocks $(\omega[i+j\ell, i+(j+1)\ell-1])_{0\leq j\leq N-1}$ of length ℓ are pairwise k-Abelian equivalent. In particular, ω contains arbitrarily high k-Abelian powers.

Remark 5.5. The result in Theorem 5.4 is already known in the special case of $D = \mathbb{N}$ and $k = +\infty$ and $k = 1$ (see [24] and [29] respectively).

Before proving Theorem 5.4 we give some immediate consequences:

Corollary 5.6. *Let k and N be positive integers, and ω an infinite word avoiding k-Abelian N-powers. Then ω is arbitrarily k-imbalanced.*

Proof. This follows immediately from Lemma 5.2 and Theorem 5.4. □

Corollary 5.7. *Let ω be a Sturmian word. Then ω contains k-Abelian N-powers for all positive integers k and N.*

Proof. This follows immediately from Theorems 4.1 and 5.4; in fact the k-Abelian complexity $\mathcal{P}_\omega^{(k)}$ is bounded (by $2k$) for each positive integer k. □

Remark 5.8. It is known that a Sturmian word ω contains an N-power for each positive integer N if and only if the sequence of partial quotients in the continued fraction expansion of the slope of ω is unbounded. So, a Sturmian word whose corresponding slope has bounded partial quotients (e.g., the Fibonacci word) will not contain N-powers for N sufficiently large (e.g., the Fibonacci word contains no 4-powers [18, 23]). However, every Sturmian word will contain arbitrarily high k-Abelian powers.

Our proof of Theorem 5.4 will make use of the following well known result first conjectured by Erdős and Turan and later proved by to E. Szemerédi:

Theorem 5.9. [Szemerédi's theorem [32]] *Let $D \subseteq \mathbb{N}$ be a set of positive upper density. Then D contains arbitrarily long arithmetic progressions.*

Proof of Theorem 5.4. Let $D \subseteq \mathbb{N}$ be a set of positive upper density. First we consider the case $k = +\infty$. By assumption $\mathcal{P}_\omega^{(+\infty)}(n)$ is bounded. This is equivalent to saying that ω has bounded factor complexity. It follows by Morse-Hedlund that ω is ultimately periodic, i.e., $\omega = UV^\infty$ for some $U, V \in A^*$. For each $i \geq 0$, set $D_i = D \cap \{i + j|V| : j = 1, 2, 3, \ldots\}$. Pick $i > |U|$ such that the set D_i has positive upper density. Then an arithmetic progression of length $N + 1$ in D_i (guaranteed by Szemerédi's theorem) determines the Nth power of some cyclic conjugate of V.

Next let us fix positive integers k and N and assume that $\mathcal{P}_\omega^{(k)}(n)$ is bounded. It follows by Lemma 5.2 that ω is (k, B)-balanced for some positive integer B. We recall the following lemma proved in [29]:

Lemma 5.10. [Lemma 5.4 in [29]] *Let k and B be positive integers. There exist positive integers α_x for each $x \in A^{\leq k}$ and a positive integer M such that whenever*

$$\sum_{x \in A^{\leq k}} c_x \alpha_x \equiv 0 \pmod{M}$$

for integers c_x with $|c_x| \leq B$ for each $x \in A^{\leq k}$, then $c_x = 0$ for each $x \in A^{\leq k}$.

Set

$$\mathcal{D} = (D-1) \cap \{k, k+1, k+2 \ldots\}.$$

Then $\mathcal{D}$ is of positive upper density. We now define a finite coloring

$$\Phi : \mathcal{D} \longrightarrow \{0, 1, 2, \ldots, M-1\} \times \mathcal{F}_\omega(2k)$$

as follows

$$\Phi(n) \doteqdot \left(\sum_{x \in A^{\leq k}} |\omega[1,n]|_x \alpha_x \, (\text{mod} M) \, ; \omega[n-k+1, n+k] \right)$$

where α_x and M are as in Lemma 5.10. Note that the second coordinate of $\Phi(n)$ is the suffix of length $2k$ of $\omega[1, n+k]$. We note also that if $\Phi(m) = \Phi(n)$ for some $m < n$, then by considering the first coordinate of Φ one has

$$\sum_{x \in A^{\leq k}} |\omega[1,n]|_x \alpha_x - \sum_{x \in A^{\leq k}} |\omega[1,m]|_x \alpha_x \equiv 0 \pmod{M} \tag{5.1}$$

$$\sum_{x \in A^{\leq k}} (|\omega[1,n]|_x - |\omega[1,m]|_x) \, \alpha_x \equiv 0 \pmod{M} \tag{5.2}$$

$$\sum_{x \in A^{\leq k}} |\omega[m - |x| + 2, n]|_x \alpha_x \equiv 0 \pmod{M}. \tag{5.3}$$

Φ defines a finite partition of $\mathcal{D}$ where two elements r and s in $\mathcal{D}$ belong to the same class of the partition if and only if $\Phi(r) = \Phi(s)$. Clearly at least one class of this partition of $\mathcal{D}$ has positive upper density. Thus by Szemerédi's theorem, there exist positive integers r and t with $r \geq k$ such that

$$\{r, r+t, r+2t, \ldots, r+Nt\} \subset \mathcal{D}$$

and

$$\Phi(r) = \Phi(r+t) = \Phi(r+2t) = \cdots = \Phi(r+Nt).$$

We now claim that the N consecutive blocks of length t

$$\omega[r+1, r+t]\omega[r+t+1, r+2t]\omega[r+2t+1, r+3t] \ldots \omega[r+(N-1)t+1, r+Nt]$$

are pairwise k-Abelian equivalent. This would prove that ω contains a k-Abelian N-power in position $r+1 \in D$.

To prove the claim, let $0 \leq i, j \leq N-1$. We will show that

$$\omega[r+it+1, r+(i+1)t] \sim_k \omega[r+jt+1, r+(j+1)t].$$

By (5.3) first taking $n = r + (i+1)t$ and $m = r + it$, then $n = r + (j+1)t$ and $m = r + jt$

$$\sum_{x\in A^{\leq k}} |\omega[r+it-|x|+2, r+(i+1)t]|_x\alpha_x \equiv \sum_{x\in A^{\leq k}} |\omega[r+jt-|x|+2, r+(j+1)t]|_x\alpha_x \equiv 0 \pmod{M}$$

and hence

$$\sum_{x\in A^{\leq k}} \left(|\omega[r+it-|x|+2, r+(i+1)t]|_x - |\omega[r+jt-|x|+2, r+(j+1)t]|_x\right)\alpha_x \equiv 0 \pmod{M}.$$

But since

$$|\omega[r+it-|x|+2, r+(i+1)t]| = |\omega[r+jt-|x|+2, r+(j+1)t]| = |x|+t-1$$

and ω is (k,B)-balanced, it follows that

$$||\omega[r+it-|x|+2, r+(i+1)t]|_x - |\omega[r+jt-|x|+2, r+(j+1)t]|_x| \leq B$$

whence by Lemma 5.10 we deduce that for each $x \in A^{\leq k}$

$$|\omega[r+it-|x|+2, r+(i+1)t]|_x = |\omega[r+jt-|x|+2, r+(j+1)t]|_x. \tag{5.4}$$

Since $\Phi(r+it) = \Phi(r+jt)$, the second coordinate of Φ gives

$$\omega[r+it-k+1, r+it+k] = \omega[r+jt-k+1, r+jt+k].$$

Together with (5.4) we deduce that for each $x \in A^{\leq k}$

$$|\omega[r+it+1, r+(i+1)t]|_x = |\omega[r+jt+1, r+(j+1)t]|_x.$$

In other words

$$\omega[r+it+1, r+(i+1)t] \sim_k \omega[r+jt+1, r+(j+1)t]$$

as required. This completes our proof of Theorem 5.4 □

References

[1] S.V. Avgustinovich, A. Frid, T. Kamae, P. Salimov, Infinite permutations of lowest maximal pattern complexity, *Theoret. Comput. Sci.,* 412 (2011) 2911–2921.

[2] M. Bucci, A. De Luca, L.Q. Zamboni, Some characterizations of Sturmian words in terms of the lexicographic order, *Fund. Inform.*, 116 (2012) 25–33.

[3] J. Cassaigne, G. Richomme, K. Saari, L.Q. Zamboni, Avoiding Abelian powers in binary words with bounded Abelian complexity, *Internat. J. Found. Comput. Sci.*, 22 (2011) 905–920.

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

[5] J. Currie, N. Rampersad. Recurrent words with constant Abelian complexity, *Adv. in Appl. Math.,* 47 (2011) 116–124.

[6] F.M. Dekking, Strongly non-repetitive sequences and progression-free sets, *J. Combin. Theory Ser. A,* 27 (1979) 181–185.

[7] X. Droubay, J. Justin, G. Pirillo, Episturmian words and some constructions of de Luca and Rauzy, *Theoret. Comput. Sci.,* 255 (2001) 539–553.

[8] A. Ehrenfeucht, G. Rozenberg, Elementary homomorphisms and a solution of the D0L sequence equivalence problem, *Theoret. Comput. Sci.,* 7 (1978) 169–183.

[9] I. Fagnot, L. Vuillon, Generalized balances in Sturmian words, *Discrete Appl. Math.,* 121 (2002) 83–101.

[10] D.G. Fon-Der-Flaass, A. Frid, On periodicity and low complexity of infinite permutations, *European J. Combin.,* 28 (2007) 2106–2114.

[11] A. Frid, On factor graphs of DOL words, *Discrete Appl. Math.,* 114 (2001) 121-130.

[12] M. Huova, Existence of infinite ternary k-Abelian square free words, Preprint, 2013.

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

[14] M. Huova, J. Karhumäki, A. Saarela, Problems in between words and abelian words: k-abelian avoidability, *Theoret. Comput. Sci.* 454 (2012) 172–177.

[15] T. Kamae, S. Widmer, L.Q. Zamboni, Abelian maximal pattern complexity of words, *Ergodic Theory Dynam. Systems* (To appear).

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

[17] J. Karhumäki, Generalized Parikh mappings and homomorphisms, *Information and Control,* 47 (1980) 155–165.

[18] J. Karhumäki, On cube free ω-words generated by binary morphisms, *Discrete Appl. Math.,* 5 (1983) 279–297.

[19] V. Keränen. Abelian squares are avoidable on 4 letters. In W. Kuich, editor, *Proceedings of ICALP'1992 (International Conference on Automata, Languages and Programming - Vienna 1992)*, volume 623 of *Lecture Notes in Comput. Sci.*, pages 41–52. Springer, Berlin, 1992.

[20] M. Lothaire, *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and its Applications.* Addison-Wesley, 1983. Reprinted in the *Cambridge Mathematical Library*, Cambridge University Press, UK, 1997.

[21] R. Mercaş, A. Saarela, 5-abelian cubes are avoidable on binary alphabets, In *Proceedings of the 14th Mons Days of Theoretical Computer Science*, 2012.

[22] R. Mercaş, A. Saarela, 3-abelian cubes are avoidable on binary alphabets. In *Proceedings of the 17th International Conference on Developments in Language Theory*, volume 7907 of *Lecture Notes in Comput. Sci.*, pages 374–383. Springer, Berlin, 2013.

[23] F. Mignosi, G. Pirillo, Repetitions in the Fibonacci infinite word, *RAIRO Theor. Inform. Appl.,* 26 (1992) 199-204.

[24] M. Morse, G.A. Hedlund, Symbolic Dynamics II: Sturmian trajectories, *Amer. J. Math.,* 62 (1940) 1–42.

[25] E. Post, A variant of a recursively unsolvable problem, *Bull. Amer. Math. Soc.,* 52 (1946) 264–268.

[26] S. Puzynina, L.Q. Zamboni, Abelian returns in Sturmian words, *J. Combin. Theory Ser. A,* 120 (2013) 390-408.

[27] M. Rao, On some generalizations of abelian power avoidability, manuscript, available at http://perso.ens-lyon.fr/michael.rao/publi/kab.pdf (2013).

[28] G. Richomme, K. Saari, L.Q. Zamboni, Balance and Abelian complexity of the Tribonacci word, *Adv. Appl. Math.,*, 45 (2010) 212–231.

[29] G. Richomme, K. Saari, L.Q. Zamboni, Abelian complexity of minimal subshifts, *J. London Math. Soc. (2),* 83 (2011) 79–95.

[30] R. Risley, L.Q. Zamboni. A generalization of Sturmian sequences; combinatorial structure and transcendence, *Acta Arith.,* XCV.2 (2000) 167–184.

[31] A. Saarela, Ultimately constant abelian complexity of infinite words, *J. Autom. Lang. Comb.,* 14 (2009) 255–258.

[32] E. Szemerédi, On sets of elements containing no k elements in arithmetic progressions, *Acta Arith.,* 27 (1975) 299–345.

[33] A. Thue, Über unendliche zeichenreihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl.,* 7 (1906) 1–22.

[34] A. Thue, Über die gegenseitige lage gleicher teile gewisser zeichen-reihen, *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl.,* 1 (1912) 1–67.