Regularity of k-Abelian Equivalence Classes of Fixed Cardinality

Juhani Karhumäki and Markus A. Whiteland(✉)

Department of Mathematics and Statistics, University of Turku,
20014 Turku, Finland
{karhumak,mawhit}@utu.fi

Abstract. Two words u and v are said to be k-Abelian equivalent if, for each word x of length at most k, the number of occurrences of x as a factor of u is the same as for v. In this note we continue the analysis of k-Abelian equivalence classes. In particular, we show that, for any fixed integer $r \geq 1$, the language of words representing equivalence classes of cardinality r is regular.

Keywords: k-Abelian equivalence · Regular languages

1 Introduction

The k-Abelian equivalence, originally introduced in [9], is an equivalence relation in between equality and Abelian equivalence of words. It identifies words u and v which contain all words of length at most k equally many times as factors. For $k = 1$ it coincides with the Abelian equivalence. Obviously, if two words u and v are k-Abelian equivalent, in symbols $u \sim_k v$, they are of equal length and, moreover, if they are so for each $k \geq 1$, then they are equal as words.

The k-Abelian equivalence defines in a natural way a complexity measure for languages, as well as for infinite words. Such a research was initiated in [12], and later continued, e.g., in [4]. Other topics of the k-Abelian equivalence such as k-Abelian repetitions, k-Abelian palindromicity and k-Abelian singletons were studied in [8,10,11,14], respectively.

A characterization of the k-Abelian equivalence in terms of rewriting was obtained in [11]. This was based on the so-called k-switching lemma. This turned out to be quite an interesting approach. It allowed, see [2,3], to show that the union of singleton classes is a regular language. Similarly, the set of all lexicographically least (or, equivalently, greatest) representatives constitutes a regular set. All these proofs are constructive, that is, given $k \in \mathbb{N}$, the above regular sets can be algorithmically found, in principle. In practice, this is not true since the size of the automaton grows at least exponentially in k. It follows that the sequence of the number of singletons of length n is a rational, or in fact, an $\mathbb{N}$-rational sequence. Similarly, the sequence of numbers of minimal elements in equivalence classes of words of length n is $\mathbb{N}$-rational, in other words, the number of equivalence classes of words of length n defines an $\mathbb{N}$-rational sequence.

H.-J. Böckenhauer et al. (Eds.): Hromkovič Festschrift, LNCS 11011, pp. 49–62, 2018.
https://doi.org/10.1007/978-3-319-98355-4_4

This latter sequence can be viewed as the complexity function of the k-Abelian equivalence. For a few small values of k and alphabet size m, the sequences are computed in [2], and are included in the On-Line Encyclopedia of Integer Sequences [17] as the sequences A289657 and A289658.

From the above it was concluded in [2] that also the union of two-element equivalence classes constitutes a regular set. This is based on the closure properties of regular sets and the above-mentioned three regular sets: the singletons, the set of minimal, and maximal elements in the equivalence classes. We recall this proof in Sect. 3. Interestingly, this approach does not extend to larger equivalence classes. The reasons for that are elaborated at the beginning of Sect. 3.

In [2], the main usage of the tool of the regular languages was to show that the number of equivalence classes of length n is asymptotic to a certain polynomial. However, by using the ideas presented in [2], we are able to prove our main theorem. For each $r \geq 1$, the union of the k-Abelian equivalence classes of cardinality r is a regular set. This is the main context of this note.

This note is arranged as follows. In Sect. 2 we lay down the basic terminology of combinatorics on words, automata theory, and graph theory needed in the remainder of the text. We also recall relevant results from the literature. In Sect. 3 we prove our main result. We also discuss an approach suggested in [2] and show that it fails. We then conclude with straightforward implications of the main result and further discussion in Sect. 4.

2 Preliminaries and Notation

We set some basic terminology and notions from the literature of combinatorics on words, for more on this topic we refer the reader to [13]. A finite set Σ of symbols is called an *alphabet*. The set of finite sequences, or *words*, over Σ is denoted by Σ^* and the set of non-empty words is denoted by Σ^+. The empty word is denoted by ε. We let $|w|$ denote the length of a word (as a sequence) $w \in \Sigma^*$. By convention, we set $|\varepsilon| = 0$. The set of words of length n over the alphabet Σ is denoted by Σ^n.

We index the letters of a given word starting from 1. For a word $w = a_1 a_2 \cdots a_n \in \Sigma^*$ and indices $1 \leqslant i \leqslant j \leqslant n$, we call the word $a_i \cdots a_j$, denoted by $w[i,j]$, a *factor* of w. For $i > j$ we set $w[i,j] = \varepsilon$. Similarly, for $i < j$ we let $w[i,j)$ denote the factor $a_i \cdots a_{j-1}$, and we set $w[i,j) = \varepsilon$ when $i \geqslant j$. We let $w[i..]$ (resp., $w[..i]$) denote the factor $w[i,n]$ (resp., $w[1,i]$) for brevity. For any i, the factor $w[..i]$ (resp., $w[i..]$) is called a *prefix* (resp., *suffix*) of w. We say that a word $x \in \Sigma^*$ *occurs at position* i *in* w if the word $w[i..]$ has x as a prefix. For $u \in \Sigma^+$ we let $|w|_u$ denote the number of occurrences of u as a factor of w. The set of factors of a word w is denoted by $F(w)$, and we set $F_n(w) = F(w) \cap \Sigma^n$. We call u a *complete return to* x *in* w if $u \in F(w)$ such that $|u|_x = 2$ and x occurs as both a prefix and a suffix of u. The set of complete first returns to x in u is denoted by $\Re_w(x)$.

We also need a few basic properties of *regular languages. Regular expressions* over an alphabet Σ are the finite expressions constructed recursively as follows.

The symbol $\emptyset$, and each $a \in \Sigma \cup \{\varepsilon\}$ are expressions. If E and E' are expressions then so are $E \cdot E'$, $E + E'$, and E^*. Each expression E defines a language, denoted by $L(E)$ as follows: Each $a \in \Sigma \cup \{\varepsilon\}$ defines the singleton language $L(a) = \{a\}$ and $\emptyset$ defines the empty language. For expressions E and E', the expressions $E \cdot E'$, $E + E'$ and E^* define the languages $L(E) \cdot L(E')$, $L(E) \cup L(E')$, and $\bigcup_{n \geq 0} L(E)^n$, respectively.

A *deterministic finite automaton* (DFA) $\mathcal{A}$ over Σ is a tuple (Q, q_0, δ, F), where Q is a finite set of states, q_0 is the initial state, δ is a partial function $\delta \colon Q \times \Sigma \to Q$ called the *transition function*, and $F \subseteq Q$ is the set of final states. Given a word $w = a_1 \cdots a_n$, the automaton operates on w using δ starting from q_0 by the rule $\delta(q, au) = \delta(\delta(q, a), u)$ for all $u \in \Sigma^+$. If $\delta(q_0, w) \in F$ we say that $\mathcal{A}$ *accepts* w. We let $L(\mathcal{A})$ denote the language *recognized* by $\mathcal{A}$; $L(\mathcal{A}) = \{w \in \Sigma^* \mid \mathcal{A} \text{ accepts } w\}$.

The languages defined by regular expressions (or recognized by finite automata) are exactly the regular languages, in fact, these two models are equivalent. Another equivalent model for regular languages considered here are *nondeterministic finite automata* (NFA), in which case the transition function may be multi-valued. We refer to [7] for this knowledge, and on more of equivalent models and closure properties of regular languages (closure under complementation, taking a morphic pre-image, etc.). In addition to classical language-theoretic notions, we use the theory of *languages with multiplicities*. This counts how many times a word occurs in a language. This leads to the theory of $\mathbb{N}$-*rational sets*. Using the terminology of [5,16], a multiset over Σ^* is called $\mathbb{N}$-*rational* if it is obtained from finite multisets by applying finitely many times the rational operations *product*, *union*, and taking *quasi-inverses*, i.e., *iteration* restricted to ε-free languages. Equivalently, an $\mathbb{N}$-rational set equals the set of multiplicities of distinct ways an NFA accepts each word. Further, a unary $\mathbb{N}$-rational set is referred to as an $\mathbb{N}$-*rational sequence*. We refer to [5,16] for more on this topic. For a language $L \subseteq \Sigma^*$, the *generating function* $G_L(x)$ of L is defined as the formal power series $G_L(x) = \sum_{n \geq 0} \#(L \cap \Sigma^n) x^n$. The basic result we need is (see [5,16]):

Proposition 1. *Let L be a regular language. The sequence $(\#(L \cap \Sigma^n))_{n \geq 0}$, is an $\mathbb{N}$-rational sequence. Consequently, the* generating function G_L *is a rational function.*

When speaking of the generating function for a language $L \subseteq \Sigma^*$, we mean the generating function for the function ℓ_L defined by $\ell_L(n) = \#(L \cap \Sigma^n)$.

We now turn to the main notion of this paper, k-Abelian equivalence. We recall some results from the literature needed in the remainder of the paper.

Definition 1. *The words $u, v \in \Sigma^*$ are k-*Abelian equivalent*, $u \sim_k v$ in symbols, if $|u|_x = |v|_x$ for all $x \in \Sigma^+$ with $|x| \leqslant k$.*

The relation $\sim_k$ is clearly an equivalence relation; we let $[u]_k$ denote the k-Abelian equivalence class represented by u.

In [11], k-Abelian equivalence is characterized in terms of rewriting, namely by *k-switching*. For this we define the following. Let $k \geqslant 1$ and let $u \in \Sigma^*$.

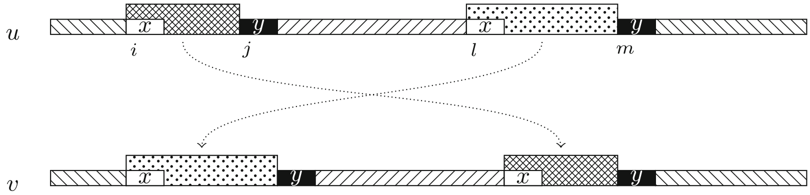

Fig. 1. Illustration of a k-switching. Here $v = S_{k,u}(i,j,l,m)$; the white rectangles symbolize $x \in \Sigma^{k-1}$ and the black rectangles symbolize $y \in \Sigma^{k-1}$.

Suppose that there exist $x, y \in \Sigma^{k-1}$, not necessarily distinct, and indices i, j, l and m, with $1 \leq i < j \leqslant l < m \leq n$, such that x has positions i and l in u, and y has positions j and m in u. In other words, we have

$$u = u[1,i) \cdot u[i,j) \cdot u[j,l) \cdot u[l,m) \cdot u[m..],$$

where the words $u[i..]$ and $u[l..]$ both begin with x, and the words $u[j..]$ and $u[m..]$ both begin with y. Furthermore, $u[i,j)$, $u[l,m) \neq \varepsilon$, but we allow $l = j$, in which case $y = x$ and $u[j,l) = \varepsilon$. We define a *k-switching* on u, denoted by $S_{u,k}(i,j,l,m)$, as

$$S_{u,k}(i,j,l,m) = u[1,i) \cdot u[l,m) \cdot u[j,l) \cdot u[i,j) \cdot u[m..]. \tag{1}$$

An illustration of a k-switching can be found in Fig. 1.

Let us define a relation R_k of Σ^* by uR_kv if and only if v is obtained from u by a k-switching. Now R_k is clearly symmetric, so that the reflexive and transitive closure R_k^* of R_k is an equivalence relation on Σ^*. This gives quite a different characterization of the relation $\sim_k$ in terms of rewriting, see [11] for more details:

Proposition 2. *For $u, v \in \Sigma^*$, we have $u \sim_k v$ if and only if uR_k^*v.*

In order to recall the following result, we need some notation. Let $\lhd$ denote a total order on Σ and the corresponding lexicographic order on Σ^*. We define the language operation $\lhd$-Min$_k$ by $\lhd\text{-Min}_k(L) = \{u \in L \mid w \in L \cap [u]_k \Rightarrow u \trianglelefteqslant w\}$. We shall often omit $\lhd$ from the prefix to avoid cluttering the text, as this does not usually concern us. Whenever $\lhd$ is omitted, the reader should consider some fixed (but arbitrary) total order $\lhd$.

The characterization in Proposition 2 allows to conclude, see [2]:

Proposition 3. *The language $\lhd$-Min$_k(\Sigma^*)$ is regular for any lexicographic order $\lhd$. In other words, the language of lexicographically least representatives with respect to $\lhd$ of the k-Abelian equivalence classes over Σ is regular.*

Example 1. In [2], minimal DFAs are constructed for the languages Min$_k(\Sigma^*)$ for small values of k and small alphabet Σ. The motivation for this is to compute explicit formulae for the number of k-Abelian equivalence classes for a given length. We recall the number of states in the constructed automata. In the case of a binary alphabet, we have minimal DFAs with 10, 49, and 936 states (including the cases for $k = 2$, 3, and 4, respectively). For the ternary alphabet and $k = 2$, the minimal DFA has 66 states. The automata grow quite fast with respect to k, so in practice such automata seem to be intractable to compute.

Another result, more in the spirit of classical formal language theory, is shown in [2] (for a detailed proof, see [3]):

Proposition 4. *Regular languages are closed under the language operation* R_k *defined by* $R_k(L) = \{v \in \Sigma^* \mid \exists u \in L \colon uR_kv\}$.

Note that we overload the symbol R_k; we use R_k either as a relation or as a many-valued function, and it will always be clear from context.

We need some terminology and notation of directed graphs with loops and multiple labeled edges. For a graph $G = (V, E)$, we let $V(G)$ denote the set of vertices, and $E(G) \subseteq V \times V$ the set of edges of G. For an edge $e \in E(G)$ from x to y, we call the vertex x the *tail* of e, denoted by $\text{tail}(e)$, and y the *head* of e, denoted by $\text{head}(e)$. The number of edges from x to y is denoted by $m_G(x, y)$. A sequence $W = (e_i)_{i=1}^t$ of edges satisfying $\text{head}(e_i) = \text{tail}(e_{i+1})$ for each $i \in [1, t-1]$ is called a *walk* (in G). We set $\text{tail}(W) = \text{tail}(e_1)$ and $\text{head}(W) = \text{head}(e_t)$. Further, we let $|W|$ denote the length t of W. We call W a *path* if $\text{tail}(e_i) \neq \text{tail}(e_j)$ when $i \neq j$, and $\text{head}(e_t) \neq \text{tail}(e_i)$ for all $i \in [1, t]$. In other words, a path P does not visit any vertex twice. If the walk $(e_i)_{i=1}^{t-1}$ is a path and e_0 is an edge such that $\text{tail}(e_0) = \text{head}(e_{t-1})$ and $\text{head}(e_0) = \text{tail}(e_1)$, then we call the walk $(e_i)_{i=0}^{t-1}$ a *cycle*. We index the edges of a cycle starting from 0 for notational reasons. We consider also loops as cycles. Finally, W is called an *Eulerian walk* if W traverses each edge of G exactly once.

The concatenation $W \cdot W'$ of walks W and W' satisfying $\text{head}(W) = \text{tail}(W')$ is defined in a natural way. For an empty walk W, we define $W \cdot W' = W' \cdot W = W'$. Note here that a cycle C can be concatenated with itself arbitrarily many times. We say that a walk W is a *repetition of a cycle* if we may write $W = C^r$ (C concatenated r times) for some $r \geq 1$.

In this note we make use of *de Bruijn graphs* (see [1] and references therein) defined as follows. For any $k \geqslant 1$ and alphabet Σ, the *de Bruijn graph* $dB_\Sigma(k)$ *of order* k *over* Σ is defined as a directed graph for which the set of vertices equals Σ^k. For each word $z \in \Sigma^{k+1}$ we have an edge $(z[..k], z[2..])) \in dB(k)$. In other words, we have $(x, y) \in E(dB_\Sigma(k))$ if and only if there exists a letter $a \in \Sigma$ such that the word $xa \in \Sigma^{k+1}$ ends with y. In this case (x, y) is denoted by (x, a), a being the *label* $\text{label}((x, y))$ of the edge. We shall often omit Σ from the subscript, as it is usually clear from the context. For a walk $W = (e_i)_{i=1}^t$ in $dB(k)$, we set $\text{label}W = \text{label}(e_1) \cdots \text{label}(e_t)$.

We note that any word $u = a_1 \cdots a_n$, where $n \geq k$ and $a_i \in \Sigma$ for each $i \in [1, n]$, defines the walk $W_u = (e_i)_{i=1}^{n-k}$ in $dB(k)$. (Here $e_i = (u[i, i+k), a_{i+k})$, $i \in [1, n-k]$.) Conversely, any walk $W_u = ((x_i, a_i))_{i=1}^t$ in $dB(k)$ defines the word $\text{tail}(W_0) \cdot \text{label}(W) = x_1 \cdot a_1 a_2 \cdots a_t \in \Sigma^{k+t}$. Thus a (long enough) word $u \in \Sigma^*$ should be considered as a walk in $dB(k)$ and vice versa.

The authors of [12] observed a connection between k-Abelian equivalence and Eulerian paths in multigraph versions of *de Bruijn graphs* and we overview it here. Let $f \in \mathbb{N}^{\Sigma^k}$ be an arbitrary vector. We define $G_f = (V, E)$ as follows. We set V as the set of words $x \in \Sigma^{k-1}$ such that x is a prefix or a suffix of a word $z \in \Sigma^k$ for which $f[z] > 0$. For each $z \in \Sigma^k$ with $f[z] > 0$, we have the edge $(z[..k-1], z[2..])$ with multiplicity $f[z]$. Now G_f is a subgraph of $dB(k-1)$ equipped with weights on edges. Note that, if $f[z] = |w|_z$ for all $z \in \Sigma^*$ for some $w \in \Sigma^*$ (this being the case we set $f = f_w$), the graph G_f is the *Rauzy graph* of w of order $k-1$ (see [15]) equipped with weights on edges.

In the following, for $u, v \in \Sigma^{k-1}$, we let $\Sigma(u, v) = u\Sigma^* \cap \Sigma^* v$ and $\Sigma(u, v, n) = \Sigma(u, v) \cap \Sigma^n$.

Lemma 1 (Karhumäki et al. [12, Lemma 2.12]**).** *For a vector $f \in \mathbb{N}^{\Sigma^k}$ and words $u, v \in \Sigma^{k-1}$, the following are equivalent:*

1. *There exists a word $w \in \Sigma(u, v)$ such that $f = f_w$;*
2. *G_f has an Eulerian path starting from u and ending at v;*
3. *The underlying graph of G_f is connected, and $d^-(s) = d^+(s)$ for every vertex s, except that if $u \neq v$, then $d^-(u) = d^+(u) - 1$ and $d^-(v) = d^+(v) + 1$.*

The following corollary is immediate, as noted in [11].

Corollary 1. *For a word $w \in \Sigma(u, v)$ and $k \geq 1$, we have that $w' \sim_k w$ if and only if the walk $W_{w'}$ is an Eulerian path from u to v in G_w.*

Example 2. Let $u \in \Sigma^*$ and $x \in F_{k-1}(u)$ such that $|u|_x \geq 3$. We may then write $W_u = W_1 W_2 W_3 W_4$ for some walks W_i with $\text{head}(W_i) = x = \text{tail}(W_{i+1})$ for each $i = 1, \ldots, 3$. Then, by the above corollary, we have $u \sim_k v$, where v is defined by the walk $W_v = W_1 W_3 W_2 W_4$. Indeed, W_v is well-defined due to the choice of the extremal vertices of the walks W_i and the same edges are traversed equally many times as in W_u.

Continuing this line of thought, a formula for computing the size of a k-Abelian equivalence class represented by a given word is obtained in [11]. In the following, a *rooted spanning tree with root v* of a graph G is a spanning tree of G for which all edges are directed towards the root vertex v.

Proposition 5. *Let $k \geq 1$ and $w \in \Sigma(u, v)$ for some $u, v \in \Sigma^{k-1}$. Then*

$$\#[w]_k = \kappa_v \prod_{x \in V(G_w)} \frac{(|w|_x - 1)!}{\prod_{a \in \Sigma} |w|_{xa}!}, \tag{2}$$

where κ_v is the number of rooted spanning trees with root v *in G_w.*

3 The Regularity of Classes of Constant Cardinality

Our main goal is to analyze the language $L_{r,k,\Sigma}$ of words w over Σ satisfying $\#[w]_k = r$, or more formally, $L_{r,k,\Sigma} = \{w \in \Sigma^* \mid \#[w]_k = r\}$. We shall often omit Σ from the subscript when there is no danger of confusion.

The language $L_{1,k}$ consists of words representing singleton classes. They are thus uniquely defined by the frequencies of the factors of length k together with the suffix of length $k-1$. The number of such words of length n is considered extensively in [11]. In [2], $L_{1,k}$ is shown to be regular. Indeed, the complement of $L_{1,k,\Sigma}$ is defined by the regular expression

$$\Sigma^* \left(\sum_{\substack{x,y \in \Sigma^{k-1} \\ a,b \in \Sigma,\ a \neq b}} \Big(\big((xb\Sigma^* \cap \Sigma^* y)\, \Sigma^* \cap \Sigma^* x \big) a\Sigma^* \cap \Sigma^* y \Big) \right) \Sigma^*. \tag{3}$$

We remark that a crude upper bound on the number of states in the minimal DFA recognizing $L_{1,k}$ can be obtained from the above regular expression using well-known conversions between various models of regular languages (see, [6,7]). Indeed, e.g., *Glushkov's algorithm* outputs an equivalent NFA of $n+1$ states, given a regular expression of n occurrences of alphabet symbols. The determinization of an n-state NFA can, in the worst case, give a DFA with 2^n states. Observe that the minimal DFA of a language and its complement have the same number of states.

The first few exact values for $\Sigma = \{a, b\}$ are as follows. The minimal DFAs recognizing $L_{1,k}$ for $\Sigma = \{a, b\}$ contain 15, 87, and 1011 states for $k = 2$, 3, and 4, respectively. These values include the garbage state. The automata for $k = 2$ and 3 are presented in [2] (garbage state omitted), and we propose the value for $k = 4$ without proof. For a ternary Σ, we have 84 states for $k = 2$. We recall the minimal DFA of $L_{1,2,\{a,b\}}$ in Fig. 2. For more on this automaton and for automata for other values of k and alphabets, we refer the reader to [2].

Not only the languages $L_{1,k}$ are regular, but so are the languages $L_{2,k}$ as shown in [2]. The proof goes as follows. Recall that $\lhd$-$\mathrm{Min}_k(\Sigma^*)$ is regular for any lexicographic order $\lhd$. Let $\lhd^R$ be the reversal of $\lhd$, that is, $b \lhd^R a$ if and only if $a \lhd b$. After a brief consideration, it becomes clear that

$$\Sigma^* \setminus R_k^2(\Sigma^* \setminus (\lhd\text{-}\mathrm{Min}_k(\Sigma^*) \cup \lhd^R\text{-}\mathrm{Min}_k(\Sigma^*))) = L_{1,k,\Sigma} \cup L_{2,k,\Sigma}$$

is regular since all the language operations, including R_k, preserve regularity. It follows that $L_{2,k,\Sigma} = \Sigma^* \setminus \big(R_k^2(\Sigma^* \setminus (\lhd\text{-}\mathrm{Min}_k(\Sigma^*) \cup \lhd^R\text{-}\mathrm{Min}_k(\Sigma^*))) \big) \setminus L_{1,k,\Sigma}$ is regular since $L_{1,k,\Sigma}$ is regular.

The main result of this paper is the generalization of the above to all $r \in \mathbb{N}$:

Theorem 1. *For any $k, r \geq 1$ and alphabet Σ, the language $L_{r,k}$ is regular.*

The approach of removing (in a regular way) one element of each class at a time does not seem to extend to the languages $L_{r,k}$, $r \geq 3$. The approach of

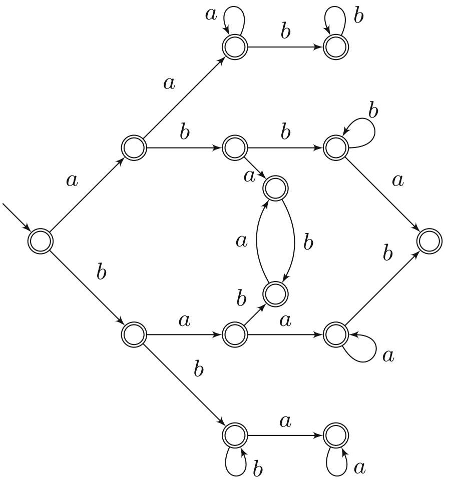

Fig. 2. The minimal DFA recognizing $L_{1,2,\{a,b\}}$. The garbage state is not illustrated. All other states are accepting.

proving the above theorem as suggested in [2, end of Sect. 6], fails as we will shortly show. The method described is of similar flavour as in the proof of the regularity of $L_{2,k,\Sigma}$, namely, by setting $K_{i+1} = K_i \setminus \mathrm{Min}_k(K_i)$, $K_0 = \Sigma^*$, we obtain that $\Sigma^* \setminus R_k^r(K_r) = \cup_{i \le r} L_{r,k}$ for each $r \in \mathbb{N}$. If Min_k preserved regularity, the above theorem would follow since a finite sequence of regularity-preserving operations would be used.

The following example shows that, unfortunately, this approach does not work, as Min_k does not preserve regularity.

Example 3. Let $k \ge 1$ and $L = (ab^k)^* \cup ab^{k-1}b^*(ab^{k-1})^*$. It is straightforward to check, e.g., using Corollary 1, that

$$\mathrm{Min}_k(L) = (ab^k)^* \cup \{ab^{k-1}b^r(ab^{k-1})^s \mid r \neq s+1\}.$$

It follows that $L \setminus \mathrm{Min}_k(L) = \{(ab^{k-1})b^{r+1}(ab^{k-1})^r \mid r \ge 1\}$. The language $h^{-1}(L \setminus \mathrm{Min}_k(L))$, where $a \mapsto ab^{k-1}$, $b \mapsto b$, equals $\{abb^r a^r \mid r \ge 1\}$ which is not regular. Since all other operations preserve regularity, we conclude that Min_k does not preserve regularity.

As a conclusion, to prove Theorem 1 we need a new approach. We do this via a characterization of the lexicographically least representatives of k-Abelian equivalence classes given in [2].

3.1 The Proof of Theorem 1

Before turning to the formal proof, we sketch the main ingredients. Our first observation (Lemma 3) states that there exists a constant $\mathcal{B}_{k,r}$ (depending on k

and r) such that, for any factor x having at least two distinct returns in u, the total number of occurrences of x in $u \in L_{r,k}$ is bounded by $\mathcal{B}_{k,r}$. We then turn to the language of minimal representatives and show that $\mathrm{Min}_k(L_{r,k})$ is regular (Theorem 2). For the proof, we make an observation concerning factors x occurring more than $\mathcal{B}_{k,r}$ times in $u \in \mathrm{Min}_k(L_{r,k})$ (Lemma 4). The main implication is that corresponding to u and a factor x occurring more than $\mathcal{B}_{k,r}$ times, we may construct a regular language $L_x = zy^*z'$ such that $u \in L_x \subseteq \mathrm{Min}_k(L_{r,k})$. In the proof of Theorem 2, this observation is applied to all such factors x to obtain, for each $u \in \mathrm{Min}_k(L_{r,k})$, a regular expression $L_u = z_0 y_1^* z_1 \cdots y_t^* z_t \subseteq \mathrm{Min}_k(L_{r,k})$. This implies that $\mathrm{Min}_k(L_{r,k})$ is a (possibly infinite) union of regular expressions. To conclude the proof we show that there are actually only finitely many distinct regular expressions in the union with the help of Lemma 5. Finally, Theorem 1 follows from Theorem 2 by applying the regularity-preserving language operation R_k on $\mathrm{Min}_k(L_{r,k})$ finitely many times.

We express the above lemmas via de Bruijn graphs and walks within. To this end we first recall some terminology from [2]. We say that a cycle $C = (d_j)_{j=0}^{s-1}$ *occurs along the walk* W if W can be written as the concatenation $W = W_1 \cdot (d_{r+j \pmod s})_{j=0}^{s-1} \cdot W_2$ for some $r \in [0, s-1]$ and some (possibly empty) walks W_1, W_2. We say that W *enters* C *via the vertex* $\mathrm{tail}(d_r)$ if W_1 is either empty or $d_{r-1 \pmod s}$ is not the last edge of W_1. In this case we say that W enters C at position $|W_1| + 1$. We say that W leaves C via the vertex $\mathrm{head}(d_{r+s-1 \pmod s})$ if W_2 is empty or d_r is not the first edge of W_2. In this case we say that W leaves C at position $|W_1 C|$.

Example 4. Consider the de Bruijn graph $dB_{\{a,b\}}(2)$. The walk $W_u = (e_i)_{i=1}^{10}$, defined by the word $u = aaaabaabaaba$, has two distinct cycles occurring along it, namely the loop $C_1 = (aa, a)$ and the cycle $C_2 = ((aa, b), (ab, a), (ba, a))$. The walk W_u enters C_1 at position 1 and leaves C_1 at position 3 (both via the vertex aa), and does not enter C_1 later on. Further, W_u enters the cycle C_2 at position 2 (via the vertex aa) and W leaves the cycle at position 10 (via the vertex ba). We may write $W = C_1^2 \cdot C_2^2 \cdot ((aa, b), (ab, a))$.

On the other hand, the walk W_{uaa} in $dB(2)$ defined by the word uaa is not cycle-deterministic, as we may write $W_{uaa} = W_u \cdot ((ba, a), (aa, a)) = C_1^2 \cdot C_2^3 \cdot C_1$, whence W_{uaa} enters the cycle C_1 at positions 1 and 12. The cycle C_2 is now left at position 11.

We recall a lemma we need in future considerations.

Lemma 2 (Cassaigne et al. [2, Lemma 5.17]**).** *Let W_u be a walk in $dB(k-1)$ defined by $u \in Min_k(\Sigma^*)$. Let $C = (d_j)_{j=0}^{s-1}$ be a cycle occurring along W_u. Suppose further that we may write $W_u = W_0 \cdot C^r \cdot W_1$ for some walks W_1, W_2, $r \geq 2$. Then, for any $t \geqslant 0$, the word u_t corresponding to the walk $W_0 \cdot C^t \cdot W_1$ is in $Min_k(\Sigma^*)$.*

We first need a couple of observations about the properties of lexicographically least representatives of equivalence classes.

Lemma 3. *Let $k, r \geq 1$. There exists an integer $\mathcal{B}_{k,r}$ such that, for each $u \in L_{r,k}$, if $\#\Re_u(x) \geq 2$ for some $x \in F_{k-1}(u)$, then $|u|_x \leq \mathcal{B}_{k,r}$.*

Proof. Let $u \in L_{r,k}$ and assume $\#\Re_u(x) \geq 2$ for some $x \in F_{k-1}(u)$. Let us write W_u in terms of complete first returns of x in u;

$$W_u = W_0 W_1 \cdots W_{|u|_x - 1} W_{|u|_x},$$

where $\mathrm{tail}(W_i) = \mathrm{head}(W_i) = x$ for all $i = 1, \ldots, |u|_x - 1$. Observe now that each walk W_i, $i = 1, \ldots, |u|_x - 1$, corresponds to the complete first return to x in u. Furthermore, W_0 and $W_{|u|_x}$ do not contain the vertex x anywhere else other than what is implied above. Now, for any permutation σ of $[1, |u|_x)$, we have that $u \sim_k v_\sigma$, where v_σ is defined by the walk $W_\sigma = W_0 W_{\sigma(1)} \cdots W_{\sigma(|u|_x - 1)} W_{|u|_x}$ (compare to Example 2). The number of distinct words obtained by this method is $\binom{|u|_x - 1}{m_1, \ldots, m_k} = \frac{(|u|_x - 1)!}{m_1! \cdots m_k!}$, where $k = \#\Re_x(u)$ and $(m_i)_i = (|u|_y)_{y \in \Re_u(x)}$. This is the number of distinct permutations of words $y \in \Re_u(x)$ with multiplicities $|u|_y$, $y \in \Re(x)$. (To see that two words obtained from distinct permutations are distinct, consider their prefixes and recall the definition of a complete first return word.) We now have, by assumption, $k \geq 2$ whence $\binom{|u|_x - 1}{m_1, \ldots, m_k} \geq |u|_x - 1$. This implies $r = \#[u]_k \geq |u|_x - 1$, or in other words, $|u|_x \leq r + 1$. □

Remark 1. For $r \geq 2$, we have $\mathcal{B}_{k,r} \geq 2$. Indeed, we have $u = a^{k+r-2} b a^{k-1} \in L_{r,k}$ and $\#\Re_u(a^{k-1}) = 2$.

In the case of $r = 1$, similar ideas were considered in [11]. Indeed, there it is shown that, for any $u \in L_{1,k,\Sigma}$, we have $\#\Re_u(x) \leq 1$ for all $x \in \Sigma^{k-1}$. In fact a characterization of words in $L_{1,k,\Sigma}$ is obtained in terms of a slight generalization of our notion of return words.

Lemma 4. *Let $u \in Min_k(L_{r,k})$ and assume $|u|_x > \mathcal{B}_{k,r}$. Then we may write $W_u = W \cdot C^{|u|_x - 1} \cdot W'$ for some cycle C with $\mathrm{tail}(C) = x$.*

Proof. Assume that $|u|_x > \mathcal{B}_{k,r}$ for some $x \in \Sigma^{k-1}$. Since $|u|_x > \mathcal{B}_{k,r}$, we have $\Re_u(x) = 1$ by the above lemma. We may write W_u in terms of y, the unique complete first return to x in u; $W_u = W \cdot W_y^{|u|_x - 1} \cdot W'$, where $\mathrm{head}(W) = \mathrm{tail}(W_y) = \mathrm{head}(W_y) = \mathrm{tail}(W') = x$. We claim that W_y is a cycle. Assume the converse; there then exists a vertex z in W_y such that there are two distinct edges both with tail z along W_u. It now follows that $\#\Re_u(z) \geq 2$ and $|u|_z > \mathcal{B}_{k,r}$, a contradiction. □

We now prove another lemma which already hints towards regular properties of our language.

Lemma 5. *Let $r \geq 2$. Let then u_s, for each $s \geq 0$, denote the word defined by the walk $W(s) = W \cdot C^s \cdot W'$ (in $dB(k-1)$) for some cycle C. Then $u_s \in Min_k(L_{r,k})$ for some $s \geq \mathcal{B}_{k,r}$ if and only if $u_s \in Min_k(L_{r,k})$ for all $s \in \mathbb{N}$.*

Proof. The other implication is immediate, so assume $u_s \in \mathrm{Min}_k(L_{r,k})$ for some $s \geq \mathcal{B}_{k,r}$. The fact that $u_s \in \mathrm{Min}_k(\Sigma^*)$ for all $s \in \mathbb{N}$ follows by Lemma 2, so it is enough to show that $u_s \in L_{r,k}$ for all $s \in \mathbb{N}$ (recall that for $r \geq 2$, we have $\mathcal{B}_{k,r} \geq 2$ by Remark 1). Without loss of generality, we may assume that $W(s)$ enters $C = (e_i)_{i=0}^{l-1}$ ($|C| = l \geq 1$) at position $|W| + 1$ and that s is maximal, i.e., $W(s)$ leaves C before position $|W| + (s+1)|C|$ and, further, that $W(s)$ leaves C via vertex $y = \mathrm{tail}(e_o)$, $0 \leq o \leq l-1$.

Observe now that $|u_s|_x \geq s$ for all $x \in V(C)$. It follows by the above lemma that for each vertex $x \in V(C)$ we have $\#\Re_{u_s}(x) = 1$ (if there were another complete first return to x in u for some $x \in V(C)$, we would have $|u_s|_x > \mathcal{B}_{k,r}$, a contradiction). Consequently, by the maximality of s, $|u_s|_{\mathrm{tail}(e_i)} = s + 1$ for all $i \in [0, o]$, and $|u_s|_{\mathrm{tail}(e_i)} = s$ for all $i = (o, l)$. Further, each (except possibly the last) occurrence of x is followed by the same letter a_x in u_s. Moreover, the only vertex $y \in V(C)$ followed by a letter $b \neq a_y$ in u is the vertex $y = \mathrm{tail}(e_o)$ via which $W(s)$ leaves C (the only exception is that W' is a subpath of C).

Consider the graph $G_s = G_{u_s}$ in light of Proposition 5. Let κ_s denote $\kappa_{\mathrm{head}(W(s))}$ for each $s \geq 0$. By the above observations we conclude that any rooted spanning tree with root $\mathrm{head}(W(s))$ of G_s contains one of the (multiple copies of the) edge e_i for each $i = [0, l) \setminus \{o\}$, and the edge $(y, b) \notin E(C)$ (unless $\mathrm{head}(W(s)) = y$ whence no edge from y exists in such a tree). Let us compute κ_s in terms of κ_0 and s. Adding s copies to an edge e_i, $0 \leq i < o$, to G_0 increases the number of trees $(s+1)$-fold, as each tree must contain exactly one copy of this edge and there are $s+1$ to choose from. For the remainder of the vertices $z \in V(C) \setminus y$, any tree in G_s must contain some copy of the path $(e_j)_{j=o+1}^{l-1}$ which connects to a copy of a tree defined by G_0. Given s copies of each edge along this path, there are altogether s^{l-o-1} choices for the path. We conclude that $\kappa_s = \kappa_0 \cdot (s+1)^o s^{l-o-1}$. This may be expressed as $\kappa_s = \kappa_0 \cdot \prod_{x \in V(C) \setminus y} (s + |u_0|_x)$. Further, we observe that, in the product $\prod_{x \in F_{k-1}(u_s)} \frac{(|u_s|_x - 1)!}{\prod_{a \in \Sigma} |u_s|_{|xa|}!}$, the only values that vary according to s are certain values corresponding to the vertices of C. In particular, $|u_s|_x = |u_s|_{xa_x} = s + |u_s|_{x_0} \in \{s, s+1\}$ for each $x \in V(C) \setminus y$, $|u_s|_y = s + 1$, and $|u_s|_{ya_y} = s$. Recall that $|u_s|_{yb} = 0$ or 1 depending on whether $\mathrm{head}(W(s)) = y$ or not. Plugging these values in (2), we find that the following ratio equals 1 for any $s \geq 1$:

$$\frac{\#[u_s]_k}{\#[u_0]_k} = \prod_{x \in V(C) \setminus y} (s + |u_0|_x) \cdot \prod_{x \in V(C) \setminus y} \frac{1}{(s + |u_0|_x)} = 1.$$

Thus, for any choice of s, the obtained word u_s has $\#[u_s] = r$. The claim follows. □

We are now in the position to prove the key result, from which our main result follows.

Theorem 2. *The language $Min_k(L_{r,k})$ is regular.*

Proof. We claim that $\mathrm{Min}_k(L_{r,k})$ is a finite union of languages defined by regular expressions of the form $z_0 y_1^* z_1 \cdots y_t^* z_t$.

Consider now a word $u \in \mathrm{Min}_k(L_{r,k})$ and write

$$W_u = W_0 C_1^{s_1} W_1 \cdots C_t^{s_t} W_t$$

for some paths W_i, $i = 0, \ldots, t$, and some repetitions $C_i^{s_i}$ of cycles C_i, $i = 1, \ldots, t$, such that W_u enters cycle C_i at position $|W_0| + \sum_{j=1}^{i-1} |C_j^{s_j} W_j| + 1$ for all $1 \le i \le t$ and leaves C_i before entering C_{i+1}. Now u may be written as

$$u = \mathrm{tail}(W_0) \cdot \mathrm{label}(W_0 C_1^{s_1} W_1 \cdots C_t^{s_t} W_t) = z_0 y_1^{s_1} z_1 \cdots y_t^{s_t} z_t,$$

where $z_0 = \mathrm{tail}(W_0) \cdot \mathrm{label}(W_0)$, $y_i = \mathrm{label}(C_i)$, and $z_i = \mathrm{label}(W_i)$ for $1 \le i \le t$. The above lemma asserts that, if $s_i \ge \mathcal{B}_{k,r}$, then

$$L(z_0 y_1^{s_1} z_1 \cdots y_i^{*} z_i \cdots y_t^{s_t} z_t) \subseteq L_{r,k}. \tag{4}$$

By repeating the above, we may replace all exponents s_j satisfying $s_j \ge \mathcal{B}_{k,r}$ with $*$ in (4).

Let L be the union of all the languages obtained as above from words $u \in \mathrm{Min}_k(L_{r,k})$ satisfying $|u|_x \le \mathcal{B}_{k,r} + 1$ for all $x \in \Sigma^{k-1}$. These words are bounded in length, so that the union is finite. Clearly $L \subseteq \mathrm{Min}_k(L_{r,k})$ by the above observation. We claim that $\mathrm{Min}_k(L_{r,k}) \subseteq L$.

Indeed, let $u \in \mathrm{Min}_k(L_{r,k})$. If $|u|_x > B_{k,r} + 1$, Lemma 4 ensures that we have $W_u = W_0 W_y^{|u|_x - 1} W_1$ for y being the unique complete first return to x in u. Further W_y is a cycle. If W_u does not enter W_y at position $|W_0| + 1$, we may extend the cycle to the left and right to obtain $W_0' W_{y'}^t W_1'$, where $t \in \{|u|_x - 1, |u|_x\}$. By the above lemma we may reduce the number of repetitions of W_y' to obtain a word u' for which $|u'|_x \le \mathcal{B}_{k,r} + 1$ and u is in the language defined by u' as in (4). If $|u'|_{x'} > \mathcal{B}_{k,r} + 1$ for some $x' \in \Sigma^{k-1}$, we may repeat the above for u' to obtain a word u'' having $|u''|_{x'} \le \mathcal{B}_{k,r} + 1$ and such that u and u' are in the language defined by u'' as in (4). This can be continued until we obtain a word v such that $|v|_x \le \mathcal{B}_{k,r} + 1$ for all $x \in \Sigma^{k-1}$ and u is contained in the language defined by v as in (4). We thus have $u \in L$, which concludes the proof. □

Proof (of Theorem 1). The language $\mathrm{Min}_k(L_{r,k})$ is regular. Since the operation R_k preserves regularity by Proposition 4, by applying finitely many iterations of R_k, we have that $L_{r,k} = R_k^r(\mathrm{Min}_k(L_{r,k}))$ is regular. □

4 Conclusions

In this note, we continued to analyze the structure of the k-Abelian equivalence classes, in particular in the framework of regularity. In [2] we concluded, as a consequence of the k-switching lemma of [11], that the set of singleton classes is a regular language. Therein, this was extended to the union of all two-element classes as well. This was based, on one hand, on the regularity of the union of lexicographically least representatives of the equivalence classes and, on the other hand, on strong closure properties of the family of regular languages.

The approach does not extend, at least immediately, and indeed fails for the first attempt, to the union of larger (fixed-size) classes. To show that also these classes are regular, we developed new techniques. This is the content of this note.

All these regular languages are algorithmically constructable. However, in practice, this can be done only in very restricted cases. Indeed, the analysis of the size of the automata of the mentioned regular languages remains for future research.

Once the regularity of the above languages is established, the well-known techniques in rational power series allow to determine the corresponding enumeration functions. This was discussed in [2]. In the current work, Theorem 2 implies the following result:

Corollary 2. *For each $k \geq 1$, the sequence of numbers of k-Abelian equivalence classes with cardinality r of length n is a rational sequence.*

In this spirit, the first few sequences of minimal elements were shown in the On-Line Encyclopedia of Integer Sequences [17]. In more detail, the sequences

$$\mathcal{P}_{k,m}(n) = \#\{[w]_k \mid |w| = n\}$$

were determined for $k = 2, 3$ in the binary alphabet. Similarly, a first few sequences considered in this note, that is, of the sequences

$$\mathcal{S}_{r,k,m}(n) = \#\{[w]_k \mid |w| = n, \#[w]_k = r\}$$

might be worth including in the encyclopedia.

Acknowledgements. We would like to thank Julien Cassaigne and Svetlana Puzynina for fruitful and insightful discussions on the topic.

References

1. de Bruijn, N.G.: Acknowledgement of priority to C. Flye Sainte-Marie on the counting of circular arrangements of 2^n zeros and ones that show each n-letter word exactly once. Technical report. (EUT report. WSK, Department of Mathematics and Computing Science), vol. 75-WSK-06, Technische Hogeschool Eindhoven, Netherlands (1975)
2. Cassaigne, J., Karhumäki, J., Puzynina, S., Whiteland, M.A.: k-abelian equivalence and rationality. Fundamenta Informaticae **154**(1–4), 65–94 (2017)
3. Cassaigne, J., Karhumäki, J., Puzynina, S., Whiteland, M.A.: k-abelian equivalence and rationality. In: Brlek, S., Reutenauer, C. (eds.) DLT 2016. LNCS, vol. 9840, pp. 77–88. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53132-7_7
4. Cassaigne, J., Karhumäki, J., Saarela, A.: On growth and fluctuation of k-abelian complexity. Eur. J. Comb. **65**(Suppl C), 92–105 (2017)
5. Eilenberg, S.: Automata, Languages, and Machines, vol. A. Academic Press Inc., New York (1974)

6. Gruber, H., Holzer, M.: From finite automata to regular expressions and back - a summary on descriptional complexity. Int. J. Found. Comput. Sci. **26**(08), 1009–1040 (2015)
7. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation, 1st edn. Addison-Wesley Publishing Co., Inc., Boston (1979)
8. Huova, M., Saarela, A.: Strongly k-abelian repetitions. In: Karhumäki, J., Lepistö, A., Zamboni, L. (eds.) WORDS 2013. LNCS, vol. 8079, pp. 161–168. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40579-2_18
9. Karhumäki, J.: Generalized Parikh mappings and homomorphisms. Inf. Control **47**(3), 155–165 (1980)
10. Karhumäki, J., Puzynina, S.: On k-abelian palindromic rich and poor words. In: Shur, A.M., Volkov, M.V. (eds.) DLT 2014. LNCS, vol. 8633, pp. 191–202. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09698-8_17
11. Karhumäki, J., Puzynina, S., Rao, M., Whiteland, M.A.: On cardinalities of k-abelian equivalence classes. Theoret. Comput. Sci. **658**(Part A), 190–204 (2017). Formal Languages and Automata: Models, Methods and Application in honour of the 70th Birthday of Antonio Restivo
12. Karhumäki, J., Saarela, A., Zamboni, L.Q.: On a generalization of abelian equivalence and complexity of infinite words. J. Comb. Theory, Ser. A **120**(8), 2189–2206 (2013)
13. Lothaire, M.: Combinatorics on Words, Encyclopedia of Mathematics and Its Applications. Advanced Book Program, World Science Division, vol. 17. Addison-Wesley, Boston (1983)
14. Rao, M., Rosenfeld, M.: Avoidability of long k-abelian repetitions. Math. Comput. **85**(302), 3051–3060 (2016)
15. Rauzy, G.: Suites á termes dans un alphabet fini. Seminaire de Théorie des Nombres de Bordeaux **12**, 1–16 (1982–1983)
16. Salomaa, A., Soittola, M.: Automata-Theoretic Aspects of Formal Power Series. Texts and Monographs in Computer Science. Springer, New York (1978). https://doi.org/10.1007/978-1-4612-6264-0
17. Sloane, N.J.A.: The on-line encyclopedia of integer sequences. Published electronically at https://oeis.org. Accessed 1 Feb 2018