Fundamenta Informaticae XXI (2001) 1001–1013
DOI 10.3233/FI-2015-0000
IOS Press

Equivalence Relations Defined by Numbers of Occurrences of Factors

Aleksi Saarela[*]
Department of Mathematics and Statistics
University of Turku, 20014 Turku, Finland
amsaar@utu.fi

Abstract. We study the question of what can be said about a word based on the numbers of occurrences of certain factors in it. We do this by defining a family of equivalence relations that generalize the so called k-abelian equivalence. The characterizations and answers we obtain are linear algebraic. We also use these equivalence relations to help us in solving some problems related to repetitions and palindromes, and to point out that some previous results about Sturmian words and k-abelian equivalence hold in a more general form.

1. Introduction

The motivating question behind this article is the following: If we know the numbers of occurrences of certain factors in a word, then how much do we actually know about that word? As a simple example, suppose that we do not know the word $u \in \{a, b\}^*$, but we know its length $|u|$ and the number of a's $|u|_a$. Then we can of course deduce the number of b's: $|u|_b = |u| - |u|_a$. As another example, suppose that we do not know the word $u \in \{a, b\}^+$, but we know its length $|u|$, first letter $\mathrm{pref}_1(u)$, last letter $\mathrm{suff}_1(u)$, and the number of ab's $|u|_{ab}$. Then we can deduce the number of ba's:

$$|u|_{ba} = |u|_{ab} + [\mathrm{pref}_1(u) = b] - [\mathrm{suff}_1(u) = b].$$

Specifically, we are interested in the following question and its variations. Let Σ be an alphabet, $k \geq 1$, and $S \subseteq \Sigma^{\leq k}$.

[*] Supported by the Academy of Finland under grant 137991

- If we know $|u|_s$ for all $s \in S$ and the prefix and suffix of u of length $k-1$, for which w do we know $|u|_w$?

Instead of the numbers $|u|_s$, we can also use sums like $|u|_s + |u|_t$.

Our considerations are closely related to abelian equivalence, where words u and v are equivalent if $|u|_a = |v|_a$ for every letter a, and to k-abelian equivalence, where u and v are equivalent if $|u|_w = |v|_w$ for every word w such that $|w| \leq k$. Here k is a positive integer, 1-abelian equivalence is the same as abelian equivalence, and as k approaches infinity, k-abelian equivalence approaches the equality relation.

There are many equivalent definitions for k-abelian equivalence. We do not need the condition $|u|_w = |v|_w$ for all $w \in \Sigma^{\leq k}$, only for some w. Especially if we add the condition that u and v must share the same prefix and suffix of length $k-1$, a much smaller subset of $\Sigma^{\leq k}$ is sufficient. One of our goals is to characterize all possible subsets that could be used in the definition, and specifically all minimal subsets.

We will define a new family of equivalence relations, so called (k, S)-equivalences. These equivalences are a generalization of k-abelian equivalence. We prove that they have strong connections to linear algebra, and we obtain linear algebraic characterizations and answers to several questions.

Many of the problems that have been studied for k-abelian equivalence could also be studied for (k, S)-equivalence (for example, we will estimate the number of equivalence classes). However, this is not our main goal. Rather, we use these relations to analyze the questions mentioned above, to help us in solving some counting problems related to repetitions and palindromes, and to point out that some previous results about Sturmian words and k-abelian complexity hold in a more general form.

2. Preliminaries

We use the Iverson bracket notation $[P] = 1$ if P is true and $[P] = 0$ if P is false.

The prefix (suffix) of length n of a word w is denoted by $\mathrm{pref}_n(w)$ (respectively, $\mathrm{suff}_n(w)$). If $n > |w|$, then we define $\mathrm{pref}_n(w) = \mathrm{suff}_n(w) = w$. A word u being a proper prefix (suffix) of w is denoted by $u \lhd w$ (respectively, $w \rhd u$).

The number of occurrences of a factor u in a word w is denoted by $|w|_u$. The empty word is denoted by ε and we define $|w|_\varepsilon = |w| + 1$.

For a finite set of words W, let $\mathcal{S}(W)$ be the set of formal sums

$$\sum_{w \in W} n_w w, \tag{1}$$

where $n_w \in \mathbb{Z}$ for all $w \in W$. We identify a word $u \in W$ with the formal sum

$$\sum_{w \in W} [w = u] w.$$

Then $W \subset \mathcal{S}(W)$. If s is the formal sum (1), we define

$$|u|_s = \sum_{w \in W} n_w |u|_w.$$

Throughout the whole article, Σ will be a fixed finite alphabet and k will be a fixed positive integer. For a set $S \subseteq \mathcal{S}(\Sigma^{\leq k})$, words $u, v \in \Sigma^*$ are called *(k,S)-equivalent* if $|u|_s = |v|_s$ for all $s \in S$, $\mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(v)$, and $\mathrm{suff}_{k-1}(u) = \mathrm{suff}_{k-1}(v)$.

(k, Σ^k)-equivalence is called *k-abelian equivalence*. This concept was introduced by Karhumäki [1]. Lately, there has been a lot of interest and more systematic research on k-abelian equivalence. Many basic properties were proved in [2]. Besides Σ^k, there are many other sets S for which (k,S)-equivalence is k-abelian equivalence. For example, $S = \Sigma^{\leq k}$ or $S = \Sigma^{\leq k} \smallsetminus a\Sigma^* \smallsetminus \Sigma^* a$ (where $a \in \Sigma$) could be used. Proofs can be found in [2] and [3]. In the case of $S = \Sigma^{\leq k}$, the prefix and suffix conditions in the definition are not necessary. In the case of $S = \Sigma^k$, either the prefix or the suffix condition is needed.

Example 2.1. Let $\Sigma = \{a, b\}$. The words $aabab$ and $abaab$ are 2-abelian equivalent, but they are not 3-abelian equivalent.

Example 2.2. Let $\Sigma = \{a, b\}$. The $(1, \{a\})$-equivalence classes are

$$b^*, b^*ab^*, b^*ab^*ab^*, b^*ab^*ab^*ab^*, \dots .$$

Example 2.3. Let $\Sigma = \{a, b, c\}$. Let h be the morphism defined by $h(a) = a$ and $h(b) = h(c) = b$. Then $u, v \in \Sigma^*$ are $(1, \{a, b + c\})$-equivalent if and only if $h(u)$ and $h(v)$ are abelian equivalent.

Example 2.4. Let $\Sigma = \{a, b\}$. The number of different (k,S)-equivalences is infinite, even for a fixed k. For example, $(k, \{a + nb\})$-equivalences are different for different $n \in \mathbb{Z}_+$.

Every (k,S)-equivalence is a congruence, that is, if u and u' are equivalent and v and v' are equivalent, then so are uv and $u'v'$. This is proved in Lemma 2.6.

Lemma 2.5. Let $s \in \mathcal{S}(\Sigma^{\leq k})$ and $u, v \in \Sigma^*$. The number $|uv|_s - |u|_s - |v|_s$ depends only on s, $\mathrm{suff}_{k-1}(u)$, and $\mathrm{pref}_{k-1}(v)$.

Proof:
If $s = \sum_{w \in \Sigma^{\leq k}} n_w w$, then

$$\begin{aligned} |uv|_s = &\sum_{w \in \Sigma^{\leq k}} n_w |uv|_w \\ =&n_\varepsilon(|u|_\varepsilon + |v|_\varepsilon - 1) + \sum_{\substack{w \in \Sigma^{\leq k} \\ w \neq \varepsilon}} n_w(|u|_w + |v|_w + |\mathrm{suff}_{|w|-1}(u)\mathrm{pref}_{|w|-1}(v)|_w) \\ =&|u|_s + |v|_s - n_\varepsilon + \sum_{\substack{w \in \Sigma^{\leq k} \\ w \neq \varepsilon}} n_w |\mathrm{suff}_{|w|-1}(u)\mathrm{pref}_{|w|-1}(v)|_w. \end{aligned}$$

This proves the claim. □

Lemma 2.6. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$. Then (k,S)-equivalence is a congruence.

Proof:
Let u and u' be equivalent and v and v' be equivalent. If $|u|, |u'| \geq k-1$, then

$$\mathrm{pref}_{k-1}(uv) = \mathrm{pref}_{k-1}(u) = \mathrm{pref}_{k-1}(u') = \mathrm{pref}_{k-1}(u'v').$$

If $|u| < k-1$ or $|u'| < k-1$, then $u = u'$ and

$$\mathrm{pref}_{k-1}(uv) = u\mathrm{pref}_{k-1-|u|}(v) = u'\mathrm{pref}_{k-1-|u'|}(v') = \mathrm{pref}_{k-1}(u'v').$$

Similarly, it can be shown that $\mathrm{suff}_{k-1}(uv) = \mathrm{suff}_{k-1}(u'v')$ in all cases.

By Lemma 2.5, $|uv|_s - |u|_s - |v|_s = |u'v'|_s - |u'|_s - |v'|_s$ for all $s \in S$. Because $|u|_s = |u'|_s$ and $|v|_s = |v'|_s$ for all $s \in S$, also $|uv|_s = |u'v'|_s$ for all $s \in S$. This proves that uv and $u'v'$ are equivalent. □

The *closure* of S, denoted by $\overline{S}$, is defined to consist of all $s \in \mathcal{S}(\Sigma^{\leq k})$ such that $|u|_s = |v|_s$ whenever u and v are (k, S)-equivalent. The definition of $\overline{S}$ depends on k, even though k does not appear in the notation. An equivalent definition would be that the closure of S is the maximal set $T \subseteq \mathcal{S}(\Sigma^{\leq k})$ such that (k, S)-equivalence is the same as (k, T)-equivalence. It follows immediately that, for $S_1, S_2 \subseteq \mathcal{S}(\Sigma^{\leq k})$, (k, S_1)-equivalence and (k, S_2)-equivalence are the same if and only if $\overline{S}_1 = \overline{S}_2$.

If $\overline{R} \neq \overline{S}$ for all proper subsets $R \subsetneq S$, then S is *independent*. If also $T \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $\overline{S} = \overline{T}$, then S is a *T-basis*. Every set T has a subset that is a T-basis, but a T-basis need not be a subset of T. We will see later that every T-basis is of the same size.

Example 2.7. Let $\Sigma = \{a, b\}$ and $k = 2$. Then $\{\varepsilon, a\}$ is independent and $b \in \overline{\{\varepsilon, a\}}$, and $\{ab\}$ is independent and $ba \in \overline{\{ab\}}$. This follows from the examples in the first paragraph of the introduction.

Example 2.8. Let $a \in \Sigma$ and $S = \Sigma^{\leq k} \smallsetminus a\Sigma^* \smallsetminus \Sigma^* a$. It was proved in [3] that (k, S)-equivalence is k-abelian equivalence, but (k, R)-equivalence is not k-abelian equivalence for any $R \subsetneq S$. This means that S is a Σ^k-basis.

The above definitions give a convenient way to state our main questions:

1. Given a set S, how big are S-bases? In other words, how many elements do we need to define (k, S)-equivalence?

2. Which sets S are independent? In other words, when does S give a minimal way to define (k, S)-equivalence?

3. Given a set S, what is the set $\overline{S}$? In other words, which numbers $|u|_s$ can we deduce based on the (k, S)-equivalence class of u?

4. For which sets S do we have $\overline{S} = \overline{\Sigma^k}$? In other words, which sets S can we use to define k-abelian equivalence?

We will answer these questions by linear algebra.

3. Connections to Linear Algebra

We will study vectors in the space $\mathbb{Q}^M$, where $M = \#\Sigma^{\leq k}$. The vector space generated by a set of vectors V is denoted by $\mathcal{L}(V)$. The *rank* of V, denoted by $\mathrm{rank}(V)$, is the dimension of $\mathcal{L}(V)$.

Let $w_1, \dots, w_M$ be the words in $\Sigma^{\leq k}$ in radix order (any other order would work as well). The *extended Parikh vector* of a word $u \in \Sigma^*$ is $P_u = (|u|_{w_1}, \dots, |u|_{w_M})$.

We define families of vectors:

$$\begin{aligned}
U_w &= (a_1, \dots, a_M), && \text{where } a_i = [w_i \rhd w] - [w \lhd w_i] && \text{and } w \in \Sigma^{k-1},\\
U'_w &= (a_1, \dots, a_M), && \text{where } a_i = [w_i = w] - [w \lhd w_i \in \Sigma^k] && \text{and } w \in \Sigma^{\leq k-1},\\
V_s &= (a_1, \dots, a_M), && \text{where } s = \sum_{i=1}^{M} a_i w_i \in \mathcal{S}(\Sigma^{\leq k}).
\end{aligned}$$

The reason for these definitions is the following lemma.

Lemma 3.1. Let $u \in \Sigma^*$. Then

$$\begin{aligned}
U_w \cdot P_u &= [u \rhd w] - [w \lhd u] && \text{for } w \in \Sigma^{k-1}, && (2)\\
U'_w \cdot P_u &= |\mathrm{suff}_{k-1}(u)|_w && \text{for } w \in \Sigma^{\leq k-1}, && (3)\\
V_s \cdot P_u &= |u|_s && \text{for } s \in \mathcal{S}(\Sigma^{\leq k}). && (4)
\end{aligned}$$

Proof:
Equations (2) were proved in [2]. A word $w \in \Sigma^{\leq k-1}$ can appear as a factor of u in two ways: As a prefix of some factor $w' \in \Sigma^k$, or as a factor of the suffix of length $k-1$. Thus

$$|u|_w = \sum_{w' \in \Sigma^k} [w \lhd w'] |u|_{w'} + |\mathrm{suff}_{k-1}(u)|_w.$$

Equations (3) follow from this. Equations (4) are clear. □

Let a be the lexicographically smallest letter and

$$\begin{aligned}
\mathcal{U} &= \{U_w \mid w \in \Sigma^{k-1}, w \neq a^{k-1}\} \cup \{U'_w \mid w \in \Sigma^{\leq k-1}\},\\
\mathcal{V}_S &= \{V_s \mid s \in S\} \text{ for } S \subseteq \mathcal{S}(\Sigma^{\leq k}).
\end{aligned}$$

It will be proved in Lemma 3.3 that the set $\mathcal{U}$ is linearly independent. This is the reason why the vector U_w with $w = a^{k-1}$ was excluded from $\mathcal{U}$. Excluding any other U_w would work as well.

For a set $S \subseteq \mathcal{S}(\Sigma^{\leq k})$, its *rank* is defined by

$$\mathrm{rank}(S) = \mathrm{rank}(\mathcal{U} \cup \mathcal{V}_S) - \mathrm{rank}(\mathcal{U}).$$

If one is familiar with matroid theory, it is easy to see that $\mathcal{S}(\Sigma^{\leq k})$ with this rank function is an infinite matroid with finite rank. It will be a consequence of later results that the independent sets and closures of this matroid are exactly the independent sets and closures we have defined.

Example 3.2. Let $\Sigma = \{a, b\}$, $k = 2$, and $S = \{ab\}$. Then

$$\begin{aligned}
\mathcal{U} =& \{U_b, U'_\varepsilon, U'_a, U'_b\} \\
=& \{(0,0,0,0,1,-1,0), (1,0,0,-1,-1,-1,-1), \\
& (0,1,0,-1,-1,0,0), (0,0,1,0,0,-1,-1)\}, \\
\mathcal{V}_S =& \{V_{ab}\} = \{(0,0,0,0,1,0,0)\}, \\
V_{ba} =& (0,0,0,0,0,1,0) \in \mathcal{L}(\mathcal{U} \cup \mathcal{V}_S).
\end{aligned}$$

Lemma 3.3. The set $\mathcal{U}$ is linearly independent.

Proof:
The set $\{U_w \mid w \in \Sigma^{k-1}, w \neq a^{k-1}\}$ was proved to be linearly independent in [2]. Let i be the smallest index such that $w_i \in \Sigma^k$. Then $w_j \in \Sigma^k$ for $j \geq i$ and $w_j \in \Sigma^{\leq k-1}$ for $j < i$. If $U_w = (a_1, \dots, a_M)$, then $a_j = 0$ for $j < i$. If $l < i$ and $U'_{w_l} = (a_1, \dots, a_M)$, then $a_l \neq 0$ and $a_j = 0$ for $j < l$. This proves that the set $\mathcal{U}$ is linearly independent. □

Lemma 3.4. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $s \in \mathcal{S}(\Sigma^{\leq k})$. If $V_s \in \mathcal{L}(\mathcal{U} \cup \mathcal{V}_S)$, then $s \in \overline{S}$.

Proof:
If $V_s \in \mathcal{L}(\mathcal{U} \cup \mathcal{V}_S)$, then

$$V_s = \sum_{w \in \Sigma^{k-1}} a_w U_w + \sum_{w \in \Sigma^{\leq k-1}} b_w U'_w + \sum_{r \in S} c_r V_r,$$

where $a_w, b_w, c_r \in \mathbb{Q}$. Let $u \in \Sigma^*$. It follows from Lemma 3.1 that

$$\begin{aligned}
|u|_s = V_s \cdot P_u &= \sum_{w \in \Sigma^{k-1}} a_w U_w \cdot P_u + \sum_{w \in \Sigma^{\leq k-1}} b_w U'_w \cdot P_u + \sum_{r \in S} c_r V_r \cdot P_u \\
&= \sum_{w \in \Sigma^{k-1}} a_w([u \rhd w] - [w \lhd u]) + \sum_{w \in \Sigma^{\leq k-1}} b_w |\mathrm{suff}_{k-1}(u)|_w + \sum_{r \in S} c_r |u|_r.
\end{aligned}$$

This means that $|u|_s$ depends only on the (k, S)-equivalence class of u, which proves that $s \in \overline{S}$. □

We will say that $\mathcal{U} \cup \mathcal{V}_S$ is *linearly independent as a multiset*, if it is linearly independent and $\mathcal{U} \cap \mathcal{V}_S = \varnothing$. Otherwise, $\mathcal{U} \cup \mathcal{V}_S$ is *linearly dependent as a multiset*.

Lemma 3.5. If $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ is independent, then $\mathcal{U} \cup \mathcal{V}_S$ is linearly independent as a multiset.

Proof:
Let $\mathcal{U} \cup \mathcal{V}_S$ be linearly dependent as a multiset. The set $\mathcal{U}$ is linearly independent by Lemma 3.3, so there exists $s \in S$ such that $V_s \in \mathcal{L}(\mathcal{U} \cup \mathcal{V}_R)$, where $R = S \smallsetminus \{s\}$. By Lemma 3.4, $s \in \overline{R}$, so $\overline{R} = \overline{S}$ and S is not independent. □

The converse of Lemma 3.4 will be proved in Theorem 5.2 and the converse of Lemma 3.5 in Theorem 5.1.

4. Number of Equivalence Classes

For a finite set $A \subset \Sigma^*$, the number of (k, S)-equivalence classes of words in A is denoted by $\mathrm{nec}_{k,S}(A)$.

If $\overline{S} = \mathcal{S}(\Sigma^{\le k})$, then (k, S)-equivalence is k-abelian equivalence and $\mathrm{nec}_{k,S}(\Sigma^n) = \Theta(n^m)$, where $m = \#\Sigma^k - \#\Sigma^{k-1}$. Here, and also later, the hidden constants in the Θ-notation can depend on the size of the alphabet Σ and on the parameter k. In some cases, they can also depend on a set R. A linear algebraic proof was given in [2] and a combinatorial proof in [3]. The following lemma is an immediate consequence.

Lemma 4.1. If $S \subseteq \mathcal{S}(\Sigma^{\le k})$ and $\overline{S} = \mathcal{S}(\Sigma^{\le k})$, then $\mathrm{nec}_{k,S}(\Sigma^{\le n}) = \Theta(n^m)$, where $m = \#\Sigma^k - \#\Sigma^{k-1} + 1$.

The next two lemmas give an upper bound and a lower bound for $\mathrm{nec}_{k,S}(\Sigma^{\le n})$. It will be proved later that these bounds match (up to a constant).

Lemma 4.2. Let $S \subseteq \mathcal{S}(\Sigma^{\le k})$ and let R be an S-basis. Then $\mathrm{nec}_{k,S}(\Sigma^{\le n}) = O(n^{\#R})$.

Proof:
The (k, S)-equivalence class of a word $u \in \Sigma^{\le n}$ is determined by $\mathrm{pref}_{k-1}(u)$, $\mathrm{suff}_{k-1}(u)$, and $|u|_r$ for $r \in R$. The number of different possible values for $\mathrm{pref}_{k-1}(u)$ and $\mathrm{suff}_{k-1}(u)$ is bounded. There is a constant c_R such that $|u|_r < c_R n$ for $r \in R$, so the number of different possible values for each $|u|_r$ is $O(n)$. Multiplying these gives the required bound $O(n^{\#R})$. □

Lemma 4.3. Let $S \subseteq \mathcal{S}(\Sigma^{\le k})$. Then $\mathrm{nec}_{k,S}(\Sigma^{\le n}) = \Omega(n^{\mathrm{rank}(\overline{S})})$.

Proof:
The proof is by reverse induction on the size of the set $\overline{S} \cap \Sigma^{\le k}$. If $\overline{S} \cap \Sigma^{\le k} = \Sigma^{\le k}$, then $\mathcal{V}_{\overline{S}}$ spans the whole space $\mathbb{Q}^{\#\Sigma^{\le k}}$, and by Lemma 3.3, $\mathrm{rank}(\mathcal{U}) = \#\mathcal{U}$, so

$$\begin{aligned}\mathrm{rank}(\overline{S}) &= \mathrm{rank}(\mathcal{U} \cup \mathcal{V}_{\overline{S}}) - \mathrm{rank}(\mathcal{U}) \\ &= \#\Sigma^{\le k} - \#\mathcal{U} \\ &= \#\Sigma^{\le k} - \#\Sigma^{k-1} + 1 - \#\Sigma^{\le k-1} \\ &= \#\Sigma^k - \#\Sigma^{k-1} + 1\end{aligned}$$

and the claim follows from Lemma 4.1.

Assume that $\overline{S} \cap \Sigma^{\le k} \subsetneq \Sigma^{\le k}$ and that the claim holds for every larger set. There exists $t \in \Sigma^{\le k} \setminus \overline{S}$. Let $T = \overline{S} \cup \{t\}$. By Lemma 3.4, $V_t \notin \mathcal{L}(\mathcal{U} \cup \mathcal{V}_{\overline{S}})$, so $\mathrm{rank}(\mathcal{U} \cup \mathcal{V}_T) = \mathrm{rank}(\mathcal{U} \cup \mathcal{V}_{\overline{S}}) + 1$ and $\mathrm{rank}(T) = \mathrm{rank}(\overline{S}) + 1$. By the induction hypothesis, $\mathrm{nec}_{k,T}(\Sigma^{\le n}) = \Omega(n^{\mathrm{rank}(\overline{S})+1})$. On the other hand, $\mathrm{nec}_{k,T}(\Sigma^{\le n}) \le (n + 1)\mathrm{nec}_{k,S}(\Sigma^{\le n})$ because the (k, T)-equivalence class of a word $u \in \Sigma^{\le n}$ is determined by its (k, S)-equivalence class and the number $|u|_t$, which has at most $n + 1$ different possible values. This proves the claim. □

The next theorem links the size of S-bases, the rank of $\mathcal{U} \cup \mathcal{V}_S$, and the number of (k, S)-equivalence classes. It also answers the first one of our main questions: Every S-basis has size $\mathrm{rank}(S)$.

Theorem 4.4. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and let R be an S-basis. Then

$$\#R = \operatorname{rank}(S) \qquad \text{and} \qquad \operatorname{nec}_{k,S}(\Sigma^{\leq n}) = \Theta(n^{\operatorname{rank}(S)}).$$

Proof:
By Lemmas 4.2 and 4.3, $\operatorname{nec}_{k,S}(\Sigma^{\leq n}) = O(n^{\#R})$ and $\operatorname{nec}_{k,S}(\Sigma^{\leq n}) = \Omega(n^{\operatorname{rank}(\overline{S})})$. Thus $\operatorname{rank}(S) \leq \operatorname{rank}(\overline{S}) \leq \#R$. On the other hand, $\mathcal{U} \cup \mathcal{V}_R$ is linearly independent as a multiset by Lemma 3.5, so

$$\begin{aligned}\#R = \#\mathcal{V}_R = \#(\mathcal{U} \cup \mathcal{V}_R) - \#\mathcal{U}& \\ &= \operatorname{rank}(\mathcal{U} \cup \mathcal{V}_R) - \operatorname{rank}(\mathcal{U}) \\ &\leq \operatorname{rank}(\mathcal{U} \cup \mathcal{V}_{\overline{S}}) - \operatorname{rank}(\mathcal{U}) = \operatorname{rank}(\overline{S}).\end{aligned}$$

It follows that $\#R = \operatorname{rank}(R) = \operatorname{rank}(\overline{S})$.

There exists an S-basis $R' \subseteq S$. Then $\operatorname{rank}(R') \leq \operatorname{rank}(S) \leq \operatorname{rank}(\overline{S})$. The above proof holds for R' in place of R, so $\operatorname{rank}(R') = \operatorname{rank}(\overline{S}) = \operatorname{rank}(S)$. This concludes the proof. □

Theorem 4.4 has the following immediate corollary.

Corollary 4.5. Let $S_1, S_2 \subseteq \mathcal{S}(\Sigma^{\leq k})$. If $\overline{S}_1 = \overline{S}_2$, then $\operatorname{rank}(\mathcal{U} \cup \mathcal{V}_{S_1}) = \operatorname{rank}(\mathcal{U} \cup \mathcal{V}_{S_2})$.

5. Answers to the Main Questions

In this section we obtain answers to the three remaining ones of the main questions stated at the end of Section 2.

Theorem 5.1. A set $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ is independent if and only if $\mathcal{U} \cup \mathcal{V}_S$ is linearly independent as a multiset.

Proof:
If S is independent, then the claim follows from Lemma 3.5. If S is not independent, then it has a finite proper subset R such that $\overline{R} = \overline{S}$. By Corollary 4.5, $\operatorname{rank}(\mathcal{U} \cup \mathcal{V}_S) = \operatorname{rank}(\mathcal{U} \cup \mathcal{V}_R)$, but $\#\mathcal{V}_S > \#\mathcal{V}_R$, so $\mathcal{U} \cup \mathcal{V}_S$ cannot be linearly independent as a multiset. □

Theorem 5.2. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $s \in \mathcal{S}(\Sigma^{\leq k})$. Then $s \in \overline{S}$ if and only if $V_s \in \mathcal{L}(\mathcal{U} \cup \mathcal{V}_S)$.

Proof:
If $V_s \in \mathcal{L}(\mathcal{U} \cup \mathcal{V}_S)$, then the claim follows from Lemma 3.4. If $V_s \notin \mathcal{L}(\mathcal{U} \cup \mathcal{V}_S)$, then $\operatorname{rank}(\mathcal{U} \cup \mathcal{V}_{S \cup \{s\}}) > \operatorname{rank}(\mathcal{U} \cup \mathcal{V}_S)$. By Corollary 4.5, $\overline{S \cup \{s\}} \neq \overline{S}$, so $s \notin \overline{S}$. □

Corollary 5.3. Let $S_1, S_2 \subseteq \mathcal{S}(\Sigma^{\leq k})$. Then $\overline{S}_1 = \overline{S}_2$ if and only if

$$\mathcal{L}(\mathcal{U} \cup \mathcal{V}_{S_1}) = \mathcal{L}(\mathcal{U} \cup \mathcal{V}_{S_2}).$$

Corollary 5.4. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$. Then $\overline{S} = \mathcal{S}(\Sigma^{\leq k})$ if and only if

$$\operatorname{rank}(\mathcal{U} \cup \mathcal{V}_S) = \#\Sigma^{\leq k}.$$

Example 5.5. Let $\Sigma = \{a, b\}$. There are 17 different $(2, S)$-equivalences such that $S \subseteq \Sigma^{\leq 2}$. The sets $\overline{S} \cap \Sigma^*$ of different ranks are listed here:

- Rank 0: $\varnothing$.
- Rank 1: $\{\varepsilon\}, \{a\}, \{b\}, \{aa\}, \{bb\}, \{ab, ba\}$.
- Rank 2: $\{\varepsilon, a, b\}$, $\{\varepsilon, aa\}$, $\{\varepsilon, bb\}$, $\{\varepsilon, ab, ba\}$, $\{a, aa, ab, ba\}$, $\{a, bb\}$, $\{b, bb, ab, ba\}$, $\{b, aa\}$, $\{aa, bb\}$.
- Rank 3: $\{\varepsilon, a, b, aa, bb, ab, ba\}$.

Every three-element subset of $\Sigma^{\leq 2}$ that is not a subset of any of the above sets of rank two is a Σ^2-basis. There are 25 such sets, giving 25 equivalent “minimal” definitions for 2-abelian equivalence on a binary alphabet.

6. Counting Powers and Palindromes

In this section, we count (k, S)-equivalence classes containing palindromes and repetitions of certain kind. Counting repetitions is quite easy, while palindromes are more complicated. This provides a nice application for (k, S)-equivalences, since obtaining a result even for just k-abelian equivalence requires the use of other (k, S)-equivalences.

First we try to count the number of squares, cubes, and higher powers from the point of view of (k, S)-equivalence. For an integer $p \geq 1$, there are at least three interesting types of words that could be considered:

- Words of the form u^p, where u can be any word. These are just ordinary pth powers.
- Words of the form $u_1 \cdots u_p$, where $u_1, \ldots, u_p$ are equivalent. For k-abelian equivalence, these are called *k-abelian pth powers*. They were first studied in [4], and the latest avoidability results were proved in [5]. The study of abelian avoidability is older, and major results can be found in [6] and [7].
- Words that are equivalent to pth powers. For k-abelian equivalence, these were called *strongly k-abelian pth powers* in [8] and *weakly k-abelian pth powers* in [9].

We are interested in the number of non-equivalent words of these types. In other words, we will estimate the number of equivalence classes containing a word of one of these types. It does not matter which type we use: An equivalence class contains a pth power if and only if it contains a word that is equivalent to a pth power, and an equivalence class contains $u_1 \cdots u_p$, where $u_1, \ldots, u_p$ are equivalent, if and only if it contains u_1^p.

Theorem 6.1. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $p \geq 1$. The number of (k, S)-equivalence classes containing a pth power of length at most n is $\Theta(\mathrm{nec}_{k,S}(\Sigma^{\leq n}))$.

Proof:
Let $n = pq + r$, where $0 \leq r < p$. Let $u_1, \dots, u_N$ be representatives of (k, S)-equivalence classes of words in $\Sigma^{\leq q}$. Then $u_1^p, \dots, u_N^p$ are pth powers in $\Sigma^{\leq n}$, and every pth power in $\Sigma^{\leq n}$ is equivalent to one of them. Because $N = \mathrm{nec}_{k,S}(\Sigma^{\leq q}) = \Theta(\mathrm{nec}_{k,S}(\Sigma^{\leq n}))$, it remains to be shown that u_i^p and u_j^p are not equivalent for any $i \neq j$.

If u_i^p and u_j^p are equivalent, then $\mathrm{pref}_{k-1}(u_i) = \mathrm{pref}_{k-1}(u_j)$ and also $\mathrm{suff}_{k-1}(u_i) = \mathrm{suff}_{k-1}(u_j)$. By Lemma 2.5, $|u_i^p|_s - p|u_i|_s = |u_j^p|_s - p|u_j|_s$ for all $s \in S$. Thus $|u_i|_s = |v_i|_s$ for all $s \in S$, which means that u_i and u_j are equivalent and $i = j$. □

The reversal of a word w is denoted by w^R. The definition of reversal is extended for elements of $\mathcal{S}(\Sigma^{\leq k})$: If $s = \sum_{w \in \Sigma^{\leq k}} n_w w \in \mathcal{S}(\Sigma^{\leq k})$, then $s^R = \sum_{w \in \Sigma^{\leq k}} n_w w^R$.

A word w is a *palindrome* if $w = w^R$. It is a *(k, S)-palindrome* if w and w^R are (k, S)-equivalent. k-abelian palindromes (that is, (k, Σ^k)-palindromes) were studied in [10].

We try to count the number of palindromes from the point of view of (k, S)-equivalence. As in the case of powers, there are at least three interesting types of words that could be considered:

- Ordinary palindromes.
- (k, S)-palindromes.
- Words that are equivalent to palindromes.

Again, we are interested in the number of non-equivalent words of these types. In other words, we will estimate the number of equivalence classes containing a word of one of these types. Trivially, an equivalence class contains a palindrome if and only if it contains a word that is equivalent to a palindrome. However, an equivalence class that contains a (k, S)-palindrome need not contain a palindrome. For example, $aaba$ is a 2-abelian palindrome, but it is not 2-abelian equivalent to any palindrome. Thus, unlike in the case of powers, we have two different counting problems: How many equivalence classes contain a palindrome, and how many equivalence classes contain a (k, S)-palindrome? It turns out that the answers differ only by a constant factor.

Lemma 6.2. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $T = \{s + s^R \mid s \in S\}$. The number of (k, S)-equivalence classes containing a (k, S)-palindrome of length at most n is $O(n^{\mathrm{rank}(T)})$.

Proof:
Let $u_1, \dots, u_N$ be (k, S)-palindromes of length at most n no two of which are (k, S)-equivalent. If we can prove that no two of them are (k, T)-equivalent, then the claim follows from Theorem 4.4.

If u_i and u_j are (k, T)-equivalent, then $\mathrm{pref}_{k-1}(u_i) = \mathrm{pref}_{k-1}(u_j)$ and $\mathrm{suff}_{k-1}(u_i) = \mathrm{suff}_{k-1}(u_j)$. Further, $|u_i|_{s+s^R} = |u_j|_{s+s^R}$ for all $s \in S$. Because u_i and u_j are (k, S)-palindromes, $|u_i|_s = |u_i|_{s^R}$ and $|u_j|_s = |u_j|_{s^R}$, so it follows that $|u_i|_s = |u_j|_s$. Thus u_i and u_j are (k, S)-equivalent and $i = j$. □

Lemma 6.3. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $T = \{s + s^R \mid s \in S\}$. The number of (k, S)-equivalence classes containing a palindrome of length at most n is $\Omega(n^{\mathrm{rank}(T)})$.

Proof:
Let $n' = \lfloor n/2 \rfloor$. By Theorem 4.4, the number of (k, T)-equivalence classes of words in $\Sigma^{\leq n'}$ is

$\mathrm{nec}_{k,T}(\Sigma^{\leq n'}) = \Theta(n^{\mathrm{rank}(T)})$. Let $\{u_1, \dots, u_N\}$ be a set of representatives of these classes. It has a subset $\{v_1, \dots, v_{N'}\}$ of size $\Theta(n^{\mathrm{rank}(T)})$ such that the words v_i share a common suffix of length $k-1$. Then $v_1 v_1^R, \dots, v_{N'} v_{N'}^R$ are palindromes of length at most n. It remains to be shown that no two of them are (k, S)-equivalent.

If $v_i v_i^R$ and $v_j v_j^R$ are (k, S)-equivalent, then $\mathrm{pref}_{k-1}(v_i) = \mathrm{pref}_{k-1}(v_j)$, and we assumed that $\mathrm{suff}_{k-1}(v_i) = \mathrm{suff}_{k-1}(v_j)$. By Lemma 2.5,

$$|v_i v_i^R|_s - |v_i|_{s+s^R} = |v_j v_j^R|_s - |v_j|_{s+s^R}$$

for all $s \in S$. Thus $|v_i|_{s+s^R} = |v_j|_{s+s^R}$ for all $s \in S$, which means that v_i and v_j are (k, T)-equivalent and $i = j$. □

Theorem 6.4. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$ and $T = \{s + s^R \mid s \in S\}$. The number of (k, S)-equivalence classes containing a palindrome of length at most n is $\Theta(n^{\mathrm{rank}(T)})$. The same is true for (k, S)-palindromes in place of ordinary palindromes.

Proof:
Follows from Lemmas 6.2 and 6.3. □

In the case of k-abelian equivalence, Theorem 6.4 takes the following form.

Theorem 6.5. The number of k-abelian equivalence classes containing a palindrome of length at most n is $\Theta(n^N)$, where

$$N = \frac{1}{2}(m^k - m^{k-1} + m^{\lceil k/2 \rceil} + m^{\lceil (k-1)/2 \rceil})$$

and $m = \#\Sigma$.

Proof:
We can fix a letter $a \in \Sigma$ and use Theorem 6.4 with the Σ^k-basis $S = \Sigma^{\leq k} \setminus a\Sigma^* \setminus \Sigma^* a$. Let $\prec$ be the lexicographical order. We use the following notation:

$$S_= = \{w \in S \mid w = w^R\},$$
$$S_\prec = \{w \in S \mid w \prec w^R\},$$
$$S_\succ = \{w \in S \mid w \succ w^R\}.$$

These sets form a partition of S. The set T in Theorem 6.4 is now

$$\{2w \mid w \in S_=\} \cup \{w + w^R \mid w \in S_\prec\}.$$

If the vectors in $\mathcal{U} \cup \mathcal{V}_T$ would satisfy a nontrivial linear relation, then so would the vectors in $\mathcal{U} \cup \mathcal{V}_S$. By the independence of S and Lemma 5.1, the set T must be independent. Thus $\mathrm{rank}(T) = \#T$ by Theorem 4.4, and we need to show that $\#T = N$.

The size of T is $\#S_= + \#S_\prec$ and $\#S_\prec = \#S_\succ$, so $\#T = (\#S + \#S_=)/2$. We know that $\#S = m^k - m^{k-1} + 1$. For $1 \leq n \leq k$,

$$\#(S_= \cap \Sigma^n) = (m-1)m^{\lceil (n-2)/2 \rceil},$$

so

$$\begin{aligned}\#S_= &= 1 + \sum_{n=1}^{k} (m-1) m^{\lceil (n-2)/2 \rceil} \\ &= 1 + (m-1) \left(\frac{m^{\lceil k/2 \rceil} - 1}{m-1} + \frac{m^{\lceil (k-1)/2 \rceil} - 1}{m-1} \right) \\ &= m^{\lceil k/2 \rceil} + m^{\lceil (k-1)/2 \rceil} - 1.\end{aligned}$$

This gives the required formula for $(\#S + \#S_=)/2$. □

7. Factor Complexity and Sturmian Words

In this section we point out that some (k, S)-equivalences have implicitly appeared before in the study of k-abelian equivalence, and that some results on k-abelian complexity can be just as easily proved in a more general form.

Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$. If $w \in \Sigma^{\omega}$ is an infinite word, we can define its *(k, S)-complexity function* $\mathcal{P}_w^{(k,S)} : \mathbb{Z}_+ \to \mathbb{Z}_+$ by letting $\mathcal{P}_w^{(k,S)}(n)$ be the number of (k, S)-equivalence classes containing a factor of w of length n.

In the case $S = \Sigma$ this gives the definition of *abelian complexity*, which has been studied in many papers, for example [11]. In the case $S = \Sigma^k$ this gives the definition of *k-abelian complexity*, which was first studied in [2]. Examples of other articles on the topic are [3] and [12]. It was proved in [2] that if $\mathcal{P}_w^{(k,\Sigma^k)}(n) < \min\{2k, n+1\}$ for some n, then w is ultimately periodic. Further, if w is aperiodic, then it is Sturmian if and only if $\mathcal{P}_w^{(k,\Sigma^k)}(n) = \min\{2k, n+1\}$ for all n. These results can be seen as k-abelian versions of the classical theorems of Morse and Hedlund [13, 14] and Coven and Hedlund [15].

Actually, the proofs in [2] use (k, Σ)-equivalence. In the article it is denoted by $\mathcal{R}_k$, and $\mathcal{P}_w^{(k,\Sigma)}$ is denoted by $\rho_w^{(k)}$. Theorems 3.2 and 4.1 in [2] give the next result about (k, S)-complexity. Theorem 4.1 is about k-abelian equivalence, but the proof works just as well for (k, Σ)-equivalence.

Theorem 7.1. Let $S \subseteq \mathcal{S}(\Sigma^{\leq k})$. Let $\Sigma \subseteq \overline{S}$ and let w be an infinite word. If $\mathcal{P}_w^{(k,S)}(n) < \min\{2k, n+1\}$ for some n, then w is ultimately periodic. If w is aperiodic, then it is Sturmian if and only if $\mathcal{P}_w^{(k,S)}(n) = \min\{2k, n+1\}$ for all n.

References

[1] Karhumäki J. Generalized Parikh mappings and homomorphisms. Information and Control. 1980;47(3):155–165. doi:10.1016/S0019-9958(80)90493-3.

[2] Karhumäki J, Saarela A, Zamboni LQ. On a generalization of Abelian equivalence and complexity of infinite words. J Combin Theory Ser A. 2013;120(8):2189–2206. doi:10.1016/j.jcta.2013.08.008.

[3] Cassaigne J, Karhumäki J, Saarela A. On growth and fluctuation of k-abelian complexity. In: Proceedings of the 10th CSR. vol. 9139 of LNCS. Springer; 2015. p. 109–122. doi:10.1007/978-3-319-20297-6_8.

[4] Huova M, Karhumäki J, Saarela A, Saari K. Local squares, periodicity and finite automata. In: Calude C, Rozenberg G, Salomaa A, editors. Rainbow of Computer Science. vol. 6570 of LNCS. Springer; 2011. p. 90–101. doi:10.1007/978-3-642-19391-0_7.

[5] Rao M. On some generalizations of abelian power avoidability. Theoret Comput Sci. 2015;601:39–46. doi:10.1016/j.tcs.2015.07.026.

[6] Dekking M. Strongly nonrepetitive sequences and progression-free sets. J Combin Theory Ser A. 1979;27(2):181–185. doi:10.1016/0097-3165(79)90044-X.

[7] Keränen V. Abelian squares are avoidable on 4 letters. In: Proceedings of the 19th ICALP. vol. 623 of LNCS. Springer; 1992. p. 41–52. doi:10.1007/3-540-55719-9_62.

[8] Huova M, Saarela A. Strongly k-abelian repetitions. In: Proceedings of the 9th WORDS. vol. 8079 of LNCS. Springer; 2013. p. 161–168. doi:10.1007/978-3-642-40579-2_18.

[9] Huova M. Combinatorics on words. New aspects on avoidability, defect effect, equations and palindromes; 2014. Doctoral dissertation, University of Turku.

[10] Karhumäki J, Puzynina S. On k-abelian palindromic rich and poor words. In: Proceedings of the 18th DLT. vol. 8633 of LNCS. Springer; 2014. p. 191–202. doi:10.1007/978-3-319-09698-8_17.

[11] Richomme G, Saari K, Zamboni LQ. Abelian complexity of minimal subshifts. J Lond Math Soc (2). 2011;83(1):79–95. doi:10.1112/jlms/jdq063.

[12] Parreau A, Rigo M, Rowland E, Vandomme E. A new approach to the 2-regularity of the l-abelian complexity of 2-automatic sequences. Electron J Combin. 2015;22(1):P1.27.

[13] Morse M, Hedlund GA. Symbolic dynamics. Amer J Math. 1938;60(4):815–866. doi:10.2307/2371264.

[14] Morse M, Hedlund GA. Symbolic dynamics II: Sturmian trajectories. Amer J Math. 1940;62(1):1–42.

[15] Coven EM, Hedlund GA. Sequences with minimal block growth. Math Systems Theory. 1973;7:138–153. doi:10.1007/BF01762232.