International Journal of Algebra and Computation
© World Scientific Publishing Company

ON THE BINOMIAL EQUIVALENCE CLASSES OF FINITE WORDS

MARIE LEJEUNE

Dept. of Mathematics, University of Liège,
Allée de la découverte 12 (B37)
Liège, B-4000, Belgium
m.lejeune@uliege.be

MICHEL RIGO

Dept. of Mathematics, University of Liège,
Allée de la découverte 12 (B37)
Liège, B-4000, Belgium
m.rigo@uliege.be

MATTHIEU ROSENFELD

LIS UMR 7020 CNRS / AMU / UTLN,
Aix Marseille Université - Campus de Luminy
163 Avenue de Luminy - case 901, BP 5
13288 Marseille cedex 9
matthieu.rosenfeld@gmail.com

Received (Day Month Year)
Accepted (Day Month Year)

Communicated by M. Volkov

Two finite words u and v are k-binomially equivalent if, for each word x of length at most k, x appears the same number of times as a subsequence (i.e., as a scattered subword) of both u and v. This notion generalizes abelian equivalence. In this paper, we study the equivalence classes induced by the k-binomial equivalence. We provide an algorithm generating the 2-binomial equivalence class of a word. For $k \geq 2$ and alphabet of 3 or more symbols, the language made of lexicographically least elements of every k-binomial equivalence class and the language of singletons, i.e., the words whose k-binomial equivalence class is restricted to a single element, are shown to be non context-free. As a consequence of our discussions, we also prove that the submonoid generated by the generators of the free nil-2 group (also called free nilpotent group of class 2) on m generators is isomorphic to the quotient of the free monoid $\{1, \ldots, m\}^*$ by the 2-binomial equivalence.

Keywords: combinatorics on words; context-free languages; binomial coefficients; k-binomial equivalence; nil-2 group; 2-nilpotent group.

Mathematics Subject Classification 2010: 68R15, 68Q45, 05A05, 20F18

1. Introduction

Let Σ be a totally ordered alphabet of the form $\{1 < \cdots < m\}$. We make use of the same notation $<$ for the induced lexicographic order on Σ^*.

Let $\sim$ be an equivalence relation on Σ^*. The equivalence class of the word w is denoted by $[w]_\sim$. We will be particularly interested in two types of subsets of Σ^* with respect to $\sim$. We let

$$\mathsf{LL}(\sim, \Sigma) = \{w \in \Sigma^* \mid \forall u \in [w]_\sim : w \leq u\}$$

denote the language of lexicographically least elements of every equivalence class for $\sim$. So there is a one-to-one correspondence between $\mathsf{LL}(\sim, \Sigma)$ and $\Sigma^*/\sim$. We let

$$\mathsf{Sing}(\sim, \Sigma) = \{w \in \Sigma^* \mid \#[w]_\sim = 1\}$$

denote the language consisting of the so-called $\sim$-*singletons*, i.e., the elements whose equivalence class is restricted to a single element. Clearly, we have $\mathsf{Sing}(\sim, \Sigma) \subseteq \mathsf{LL}(\sim, \Sigma)$. In the extensively studied context of Parikh matrices (see Section 2), two words are *M-equivalent* if they have the same Parikh matrix. In that setting, singletons are usually called *M-unambiguous words* and have attracted the attention of researchers, see, for instance, [14] and the references therein.

Let $k \geq 1$ be an integer. Let $\sim_{k,ab}$ be the k-abelian equivalence relation introduced by Karhumäki [7]. Two words are *k-abelian equivalent* if they have the same number of factors of length at most k. If $k = 1$, the words are *abelian equivalent*. We denote by $\Psi(u)$ the *Parikh vector* of the finite word u, defined as

$$\Psi(u) = (|u|_1, \ldots, |u|_m)\,,$$

where $|u|_a$ is the number of occurrences of the letter a in u. Two words u and v are abelian equivalent if and only if $\Psi(u) = \Psi(v)$.

The k-abelian equivalence relation has recently received a lot of attention, see, for instance, [9,10]. In particular, the number of k-abelian singletons of length n is studied in [8]. Based on an operation of k-switching, the following result is given in [1].

Theorem 1.1. *Let $k \geq 1$. Let Σ be a m-letter alphabet. For the k-abelian equivalence, the two languages $\mathsf{LL}(\sim_{k,ab}, \Sigma)$ and $\mathsf{Sing}(\sim_{k,ab}, \Sigma)$ are regular.*

As discussed in Section 2, the set of M-unambiguous words over a 2-letter alphabet is also known to be regular. Motivated by these types of results, we will consider another equivalence relation, namely the k-binomial equivalence introduced in [13], and study the corresponding sets LL and Sing.

Definition 1.2. *We let the* binomial coefficient $\binom{u}{v}$ *denote the number of times v appears as a (not necessarily contiguous) subsequence of u. Let $k \geq 1$ be an integer. Two words u and v are* k-binomially equivalent, *denoted $u \sim_k v$, if $\binom{u}{x} = \binom{v}{x}$ for all words x of length at most k.*

We will show that k-abelian and k-binomial equivalences have incomparable properties for the corresponding languages LL and Sing. Both of these equivalence relations refine the classical abelian equivalence and it is interesting to see how they differ. As mentioned by Whiteland in his Ph.D. thesis: "part of the challenges in this case follow from the property that a modification in just one position of a word can have global effects of the distribution of subwords, and thus the structure of the equivalence classes." [16].

This paper is organized as follows. The special case of 2-binomial equivalence over a 2-letter alphabet is presented in Section 2: the corresponding languages are known to be regular. In Section 3, we discuss an algorithm generating the 2-binomial equivalence class of any word over an arbitrary alphabet. Then we prove that the submonoid generated by the generators of the free nil-2 group (also called free nilpotent group of class 2) on m generators is isomorphic to $\{1,\dots,m\}^*/\sim_2$. Section 4 is about the growth rate of $\#(\Sigma^n/\sim_k)$. As a consequence of Sections 3 and 4, the growth function for the submonoid generated by the generators of the free nil-2 group on $m \geq 3$ generators is in $\Theta\left(n^{m^2-1}\right)$. In the last section, contrasting with Theorem 1.1, we show that $\mathsf{LL}(\sim_k, \Sigma)$ and $\mathsf{Sing}(\sim_k, \Sigma)$ are rather complicated languages when $k \geq 2$ and $\#\Sigma \geq 3$: they are not context-free.

2. 2-binomial equivalence over a 2-letter alphabet

Let $\Sigma = \{1, 2\}$ be a 2-letter alphabet. Recall that the *Parikh matrix* associated with a word $w \in \{1,2\}^*$ is the 3×3 matrix given by

$$P(w) = \begin{pmatrix} 1 & |w|_1 & \binom{w}{12} \\ 0 & 1 & |w|_2 \\ 0 & 0 & 1 \end{pmatrix}.$$

See [12] for a longer introduction on Parikh matrices. For $a, b \in \{1,2\}$, $\binom{w}{ab}$ can be deduced from $P(w)$. Indeed, we have $\binom{w}{aa} = \binom{|w|_a}{2}$ and if $a \neq b$,

$$\binom{w}{aa} + \binom{w}{ab} + \binom{w}{ba} + \binom{w}{bb} = \binom{|w|_a + |w|_b}{2}. \tag{2.1}$$

It is thus clear that $w \sim_2 x$ if and only if $P(w) = P(x)$. We can therefore make use of the following theorem of Fossé and Richomme [2]. If two words u and v over an arbitrary alphabet Σ can be factorized as $u = xabybaz$ and $v = xbayabz$ with $a, b \in \Sigma$, we write $u \equiv v$. The reflexive and transitive closure of this relation is denoted by $\equiv^*$.

Theorem 2.1. *Let u, v be two words over $\{1,2\}$. The following assertions are equivalent:*

- *the words u and v have the same Parikh matrix;*
- *the words u and v are 2-binomially equivalent;*
- *$u \equiv^* v$.*

Consequently, the language $\mathsf{Sing}(\sim_2, \{1,2\})$ avoiding two separate occurrences of 12 and 21 (or, 21 and 12) is regular. A regular expression for this language is given by

$$1^*2^* + 2^*1^* + 1^*21^* + 2^*12^* + 1^*212^* + 2^*121^*.$$

A NFA accepting $\mathsf{Sing}(\sim_2, \{1,2\})$ was given in [14].

Remark 2.2. From [13], we know that

$$\#\mathsf{LL}(\sim_2, \{1,2\}) = \#\left(\{1,2\}^n/\sim_2\right) = \frac{n^3+5n+6}{6}.$$

Note that this is exactly the OEIS sequence `A000125` of *cake numbers*, i.e., the maximal number of pieces resulting from n planar cuts through a cube.

Proposition 2.3. *The language* $\mathsf{LL}(\sim_2, \{1,2\})$ *is regular.*

Proof. As a consequence of Theorem 2.1, if a word u belongs to $\mathsf{LL}(\sim_2, \{1,2\})$, it cannot be of the form $x21y12z$ because otherwise, the word $x12y21z$ belongs to the same class and is lexicographically less. Consequently,

$$\mathsf{LL}(\sim_2, \{1,2\}) \subseteq \{1,2\}^* \setminus \{1,2\}^*21\{1,2\}^*12\{1,2\}^*.$$

The reader can check that the language in the r.h.s. has exactly $(n^3+5n+6)/6$ words of length n. We conclude with the previous remark that the two languages are thus equal. □

3. 2-binomial equivalence over an m-letter alphabet

Theorem 2.1 does not hold for ternary or larger alphabets. Indeed, the two words 1223312 and 2311223 are 2-binomially equivalent but both words belong to $\mathsf{Sing}(\equiv, \{1,2,3\})$ which means that $1223312 \not\equiv^* 2311223$. However, we still have that $u \equiv^* v$ implies $u \sim_2 v$. It is therefore meaningful to study $\sim_2$ over larger alphabets and to describe the 2-binomial equivalence classes.

The first few terms of $\left(\#\left(\{1,2,3\}^n/\sim_2\right)\right)_{n\geq 0}$ are given by

$$1, 3, 9, 27, 78, 216, 568, 1410, \ldots.$$

This sequence also appears in the Sloane's encyclopedia as entry `A140348` which is the growth function for the submonoid generated by the generators of the free nil-2 group on three generators (if we stick to the terminology of this entry in the encyclopaedia). In this section, we make explicit the connection between these two notions (see Theorem 3.11).

Recall that the *commutator* of two elements x, y belonging to a multiplicative group $(G, \cdot)$ is $[x,y] = x^{-1}y^{-1}xy$. Hence, the following relations hold

$$xy = yx[x,y] \quad \forall x, y \in G.$$

A group of nilpotency class 2 or, *nil*-2 group for short, is a group G for which the commutators belong to the center $Z(G)$, i.e.,

$$[x, y]z = z[x, y] \quad \forall x, y, z \in G. \tag{3.1}$$

Let $\Sigma = \{1, \dots, m\}$. The free nil-2 group on m generators has thus a presentation

$$N_2(\Sigma) = \langle \Sigma \mid [x, y]z = z[x, y] \ (x, y, z \in \Sigma) \rangle .$$

As an example, making use of these relations, let us show that two elements of the free group on $\{1, 2, 3\}$ are equivalent in $N_2(\{1, 2, 3\})$:

$$12321 = (12[2, 1])[1, 2]321 = 21[1, 2]321 = 213(21[1, 2]) = 21312.$$

Let Σ^{-1} be the alphabet $\{1^{-1}, \dots, m^{-1}\}$ of the inverse letters, that we suppose disjoint from Σ. By abuse of notation, for all $x \in \Sigma$, $(x^{-1})^{-1}$ is the letter x. Since $N_2(\Sigma)$ is the quotient of the free monoid $\left(\Sigma \cup \Sigma^{-1}\right)^*$ under the congruence relations generated by $xx^{-1} = \varepsilon$ and (3.1), we will consider the natural projection denoted by

$$\pi : \left(\Sigma \cup \Sigma^{-1}\right)^* \to N_2(\Sigma). \tag{3.2}$$

In Section 3.1, we provide an algorithmic description of any 2-binomial class. We make use of this description in Section 3.2 to show that the monoid $\Sigma^*/\sim_2$ is isomorphic to the submonoid, generated by Σ, of the nil-2 group $N_2(\Sigma)$.

3.1. *A nice tree generating the $\sim_2$-class of a word*

Let w be a word over Σ and ℓ be the lexicographically least element in its abelian equivalence class, i.e.,

$$\ell = 1^{|w|_1} 2^{|w|_2} \cdots m^{|w|_m}.$$

We describe an algorithm that, given w, produces a finite sequence of words $\mathcal{L}_w = (\ell_i)_i$ starting with $\ell_0 = \ell$, ending with w and such that two consecutive words in the sequence differ only by exchanging two adjacent symbols.

We let $u \wedge v$ denote the longest common prefix of two finite words u and v. The idea is that ℓ_{i+1} is obtained from ℓ_i by a single swap of two adjacent letters $ab \mapsto ba$ with $a < b$, in such a way that $|\ell_i \wedge w| \leq |\ell_{i+1} \wedge w|$.

During step $i \geq 0$, assume that we have already obtained $\ell_0, \dots, \ell_i$. Let $p = \ell_i \wedge w$. If $p = w$, then we are done. Otherwise, there exist $c, d \in \Sigma$, $c \neq d$, $x, y \in \Sigma^*$ such that $\ell_i = pcx, w = pdy$. Since d appears in x, let us consider its leftmost occurrence. There exist $u, v \in \Sigma^*$ such that $cx = cudv$ with $|u|_d = 0$. It can easily be shown by induction that the word cx is the least lexicographic word of its abelian class. It follows that $c < d$ and u only contains letters less than d.

To move the letter d in front of the word cu, perform $|u| + 1$ swaps of the form $c'd \mapsto dc'$ with $c' < d$ (c' is a letter occurring in cu). This defines $\ell_{i+1}, \dots, \ell_{i+|u|+1}$. More precisely, $\ell_{i+j} = pcu_1 \cdots u_{|u|-j} d u_{|u|-j+1} \cdots u_{|u|} v$ for $j \in \{1, \dots, |u|\}$ and

$\ell_{i+|u|+1} = pdcuv$. Note that $|\ell_{i+|u|+1} \wedge w|$ is at least $|\ell_i \wedge w| + 1$. Indeed, with these swaps, the $(|p|+1)^{\text{st}}$ letter of $\ell_{i+|u|+1}$ is now a d, as in w.

Remark 3.1. The letter d has been swapped several times until it reaches a position that corresponds to its position in w. We stress the fact that afterwards, any other letter that will be swapped by the algorithm will not affect the prefix pd. This remark will allow us to define a graph $\mathcal{G}_w$ that will then be restricted to a tree.

To obtain the complete sequence $\mathcal{L}_w$, we iterate the previous process until we reach w.

Example 3.2. Take the word $w = 2113223$. If we apply the above algorithm, we get the sequence

$$\mathcal{L}_w = (\ell_0 = 1122233,\ \ell_1 = 1212233,\ \ell_2 = 2112233,\ \ell_3 = 2112323,\ \ell_4 = 2113223).$$

Using (2.1), the next lemma is obvious.

Lemma 3.3. *Two abelian equivalent words u, v are 2-binomially equivalent if and only if $\binom{u}{ab} = \binom{v}{ab}$ for all $a, b \in \Sigma$ with $a < b$. Let ℓ be a word in $1^*2^* \cdots m^*$. In the set of tuples of size $m(m-1)/2$*

$$\left\{ \left(\binom{w}{12}, \ldots, \binom{w}{1m}, \binom{w}{23}, \ldots, \binom{w}{2m}, \ldots, \binom{w}{(m-1)m} \right) \mid w \in [\ell]_{\sim_1} \right\},$$

the greatest element, for the lexicographic ordering, is achieved for $w = \ell$.

We consider the $m(m-1)/2$ coefficients $\binom{u}{ab}$ with $a < b$. Note that, in the algorithm, if ℓ_{j+1} is obtained from ℓ_j by an exchange of the form $ab \mapsto ba$, all these coefficients remain unchanged except for

$$\binom{\ell_{j+1}}{ab} = \binom{\ell_j}{ab} - 1. \tag{3.3}$$

Corollary 3.4. *When applying the algorithm producing the word w from the word $\ell = 1^{|w|_1} 2^{|w|_2} \cdots m^{|w|_m}$, the total number of exchanges $ab \mapsto ba$, with $a < b$, is given by*

$$\binom{\ell}{ab} - \binom{w}{ab} = \binom{w}{ba}.$$

Consequently two words are 2-binomially equivalent if and only if they are abelian equivalent and the total number of exchanges of each type $ab \mapsto ba$, $a < b$, when applying the algorithm to these two words, is the same. An equivalence class $[w]_{\sim_2}$ is thus completely determined by a word $\ell = 1^{n_1} 2^{n_2} \cdots m^{n_m}$ and the number of different exchanges. We obtain an algorithm generating all words of $[w]_{\sim_2}$.

Definition 3.5. *Given a word w, define a directed graph $\mathcal{G}_w$ whose vertices are the words belonging to the abelian equivalence class of w. There exists an edge from v to v' if and only if v' is obtained by an exchange of the type $ab \mapsto ba$, $a < b$, from v.*

The edge is labeled with the applied exchange. In particular, for any vertex w' in $\mathcal{G}_w$, there is at least a path from $\ell_0 = 1^{|w|_1}2^{|w|_2}\cdots m^{|w|_m}$ to w'.

Definition 3.6. *Let w' be a node of $\mathcal{G}_w$. If there are several paths from ℓ_0 to w' in the graph, then exactly one of them follows the sequence $\mathcal{L}_{w'}$. Thanks to Remark 3.1, restricting $\mathcal{G}_w$ to these paths gives a subgraph having the same set of vertices which is a tree denoted by $\mathcal{T}_w$ and whose root is ℓ_0. In particular, if w'' appears in the sequence $\mathcal{L}_{w'}$, then $\mathcal{L}_{w''}$ is a prefix of $\mathcal{L}_{w'}$.*

Let us recap what we have done so far in the following statement.

Proposition 3.7. *Let w be a finite word. The $\sim_2$-equivalence class of w is composed of all the nodes of $\mathcal{T}_w$ that are at level $\sum_{1\leq a<b\leq m}\binom{w}{ba}$ and such that the path from the root ℓ_0 to such a node is composed of $\binom{w}{ba}$ edges labeled by $ab \mapsto ba$, for all letters $a < b$.*

Recall that our aim is to conveniently find all the words belonging to the $\sim_2$-class of w. Instead of building the full tree $\mathcal{T}_w$, starting from the root we will only build a relevant subtree. Moreover, we conveniently add two pieces of information to help us in the construction. First, if there is an edge from $uabv$ to $ubav$ labeled by $ab \mapsto ba$, then we underline the letter b in the destination node $u\underline{b}av$.

Secondly, a sequence of nodes $w_0 = \ell_0, w_1, \ldots, w_n$ defining a path in $\mathcal{T}_w$ corresponds to the sequence $\mathcal{L}_{w_n}$. Thus, we highlight as in Remark 3.1 the prefix that will not be modified anymore if we extend the path further. This prefix is separated from the remaining part of the word by a vertical line. Two cases can occur, whether or not we continue to swap the same letter:

(1) either $w_i = u|va\underline{b}v'$ and $w_{i+1} = u|v\underline{b}av'$;
(2) or, $w_i = u|x\underline{c}yabv'$ and $w_{i+1} = uxc|y\underline{b}av'$.

Starting from the root $|1^{|w_1|}2^{|w_2|}\cdots m^{|w_m|}$ with no underlined letter, we build the tree level by level, by trying all the possible swaps that may occur on the right hand side of the vertical line. Moreover, if a path from the root to a node has a number of edges labeled by $ab \mapsto ba$ greater than $\binom{w}{ba}$, it is useless to add the children of this node, since they will not lead to any element of $[w]_{\sim_2}$. They are therefore parts of $\mathcal{T}_w$ that will not be explored.

Example 3.8. Let us consider the word $w = 1223312$ on the alphabet $\{1, 2, 3\}$. Its $\sim_2$-equivalence class is $\{1223312, 2311223\}$. It can be read from the tree in Figure 1, which is a subtree of $\mathcal{T}_w$. The edges labeled by $12 \mapsto 21$ (resp., $13 \mapsto 31$, $23 \mapsto 32$) are represented in black (resp., dotted red, dashed gray).

Let us illustrate the differences between $\mathcal{G}_w$ and the tree in Figure 1. For instance, the edges from $12|12\underline{3}23$ to $2|112\underline{3}23$ and from $11223|\underline{3}2$ to $12123|\underline{3}2$ are in the graph $\mathcal{G}_w$ but have been suppressed in the tree $\mathcal{T}_w$. These nodes do not appear as consecutive words in any sequence $\mathcal{L}_{w'}$ for $w' \in [w]_{\sim_1}$.

Moreover, a dashed gray edge from $|1123\underline{2}223$ to $1123|2\underline{3}2$ is in $\mathcal{T}_w$ but has not been computed in our figure since the path leading to this node would have three dashed gray edges but we know that we may only have $\binom{w}{32} = 2$ dashed gray edges on any path from the root.

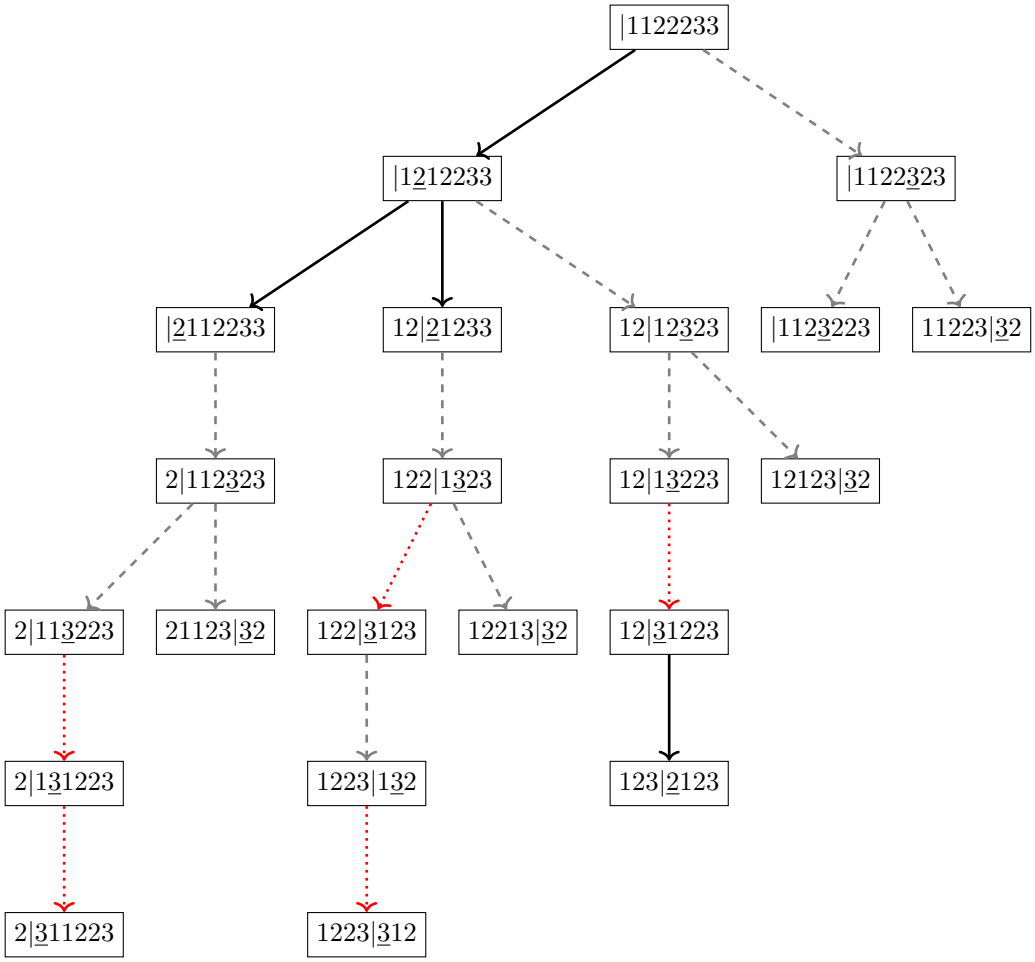

Fig. 1: Generating the $\sim_2$-class of 1223312.

Note that a polynomial time algorithm checking whether or not two words are k-binomially equivalent has been obtained in [3] and is of independent interest.

3.2. *Isomorphism with a nil-2 submonoid*

Since we are dealing with the extended alphabet $\Sigma \cup \Sigma^{-1}$, let us first introduce a convenient variation of binomial coefficients of words taking into account inverse letters.

Definition 3.9. *Let $t \geq 0$ be an integer. For all words u over the alphabet $\Sigma \cup \Sigma^{-1}$ and $v \in \Sigma^t$, let us define*

$$\begin{bmatrix} u \\ v \end{bmatrix} = \sum_{(e_1,\dots,e_t)\in\{-1,1\}^t} \left(\prod_{i=1}^{t} e_i\right) \quad \binom{u}{v_1^{e_1}\cdots v_t^{e_t}},$$

where $\binom{u}{v_1^{e_1}\cdots v_t^{e_t}}$ is the usual binomial coefficient over the alphabet $\Sigma \cup \Sigma^{-1}$. Let $w \in \left(\Sigma \cup \Sigma^{-1}\right)^*$ and denote

$$\Phi(w) = \left(\begin{bmatrix} w \\ 1 \end{bmatrix}, \dots, \begin{bmatrix} w \\ m \end{bmatrix}, \begin{bmatrix} w \\ 12 \end{bmatrix}, \dots, \begin{bmatrix} w \\ m(m-1) \end{bmatrix}\right)^{\intercal} \in \mathbb{Z}^{m^2}$$

where the last $m^2 - m$ components are obtained from all the words consisting of two different letters in Σ, ordered by lexicographical order.

Notice that, if u and v are words over Σ, then

$$\begin{bmatrix} u \\ v \end{bmatrix} = \binom{u}{v}$$

and so, u and v are 2-binomially equivalent if and only if $\Phi(u) = \Phi(v)$.

Example 3.10. Let $\Sigma = \{1, 2, 3\}$ and $w = 123^{-1}231^{-1}$. Applying the previous definition, for all $a \in \Sigma$, we have

$$\begin{bmatrix} w \\ a \end{bmatrix} = \binom{w}{a} - \binom{w}{a^{-1}}.$$

Similarly, for all $a, b \in \Sigma$, we have

$$\begin{bmatrix} w \\ ab \end{bmatrix} = \binom{w}{ab} - \binom{w}{a^{-1}b} - \binom{w}{ab^{-1}} + \binom{w}{a^{-1}b^{-1}}. \tag{3.4}$$

Therefore, computing classical binomial coefficients, we obtain

$$\Phi(w) = (0, 2, 0, 2, 0, -2, 1, 0, -1)^{\intercal}.$$

We are now ready to prove the main result of this section.

Theorem 3.11. *Let $\Sigma = \{1, \dots, m\}$. The monoid $\Sigma^*/\sim_2$ is isomorphic to the submonoid, generated by Σ, of the nil-2 group $N_2(\Sigma)$.*

Proof. Let us recall that π is the natural projection defined in (3.2). We will first show that for any two words w and w' over $\Sigma \cup \Sigma^{-1}$ such that $\pi(w) = \pi(w')$, the

relation $\Phi(w) = \Phi(w')$ holds. Indeed, using (3.4) one can easily check that, for all $a, b \in \Sigma$ and $s, t \in (\Sigma \cup \Sigma^{-1})^*$, we have

$$\begin{bmatrix} st \\ ab \end{bmatrix} = \begin{bmatrix} s \\ ab \end{bmatrix} + \begin{bmatrix} t \\ ab \end{bmatrix} + \begin{bmatrix} s \\ a \end{bmatrix} \begin{bmatrix} t \\ b \end{bmatrix}.$$

Now, one can show that, for all $u, v \in (\Sigma \cup \Sigma^{-1})^*$ and $x, y, z \in \Sigma \cup \Sigma^{-1}$,

$$\Phi(uv) = \Phi(uxx^{-1}v) \quad \text{and} \quad \Phi(u[x,y]zv) = \Phi(uz[x,y]v).$$

For instance, let $a, b \in \Sigma$ with $a \neq b$,

$$\begin{aligned} \begin{bmatrix} uxx^{-1}v \\ ab \end{bmatrix} &= \begin{bmatrix} u \\ ab \end{bmatrix} + \begin{bmatrix} xx^{-1}v \\ ab \end{bmatrix} + \begin{bmatrix} u \\ a \end{bmatrix} \begin{bmatrix} xx^{-1}v \\ b \end{bmatrix} \\ &= \begin{bmatrix} u \\ ab \end{bmatrix} + \underbrace{\begin{bmatrix} xx^{-1} \\ ab \end{bmatrix}}_{=0} + \begin{bmatrix} v \\ ab \end{bmatrix} + \underbrace{\begin{bmatrix} xx^{-1} \\ a \end{bmatrix}}_{=0} \begin{bmatrix} v \\ b \end{bmatrix} + \begin{bmatrix} u \\ a \end{bmatrix} \cdot \left(\underbrace{\begin{bmatrix} xx^{-1} \\ b \end{bmatrix}}_{=0} + \begin{bmatrix} v \\ b \end{bmatrix} \right) \\ &= \begin{bmatrix} uv \\ ab \end{bmatrix}. \end{aligned}$$

This implies that a map Φ_N can be defined on the free nil-2 group (otherwise stated, the diagram depicted in Figure 2 is commutative) by

$$\forall r \in N_2(\Sigma), \quad \Phi_N(r) = \Phi(w) \text{ for any } w \text{ such that } \pi(w) = r.$$

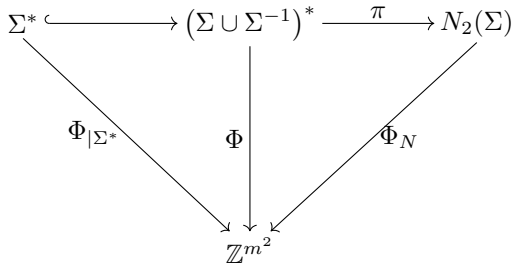

Fig. 2: A commutative diagram (proof of Theorem 3.11).

In particular, if w and w' are words over Σ such that $\pi(w) = \pi(w')$, then we may conclude that $\Phi(w) = \Phi(w')$ meaning that they are 2-binomially equivalent. Otherwise stated, for every $r \in N_2(\Sigma)$, $\pi^{-1}(r) \cap \Sigma^*$ is a subset of an equivalence class for $\sim_2$.

To conclude the proof, we have to show that all the elements of an equivalence class for $\sim_2$ are mapped by π on the same element of $N_2(\Sigma)$. Let $u, v \in \Sigma^*$ be such

that $u \sim_2 v$. Using the algorithm described in Section 3.1, there exists a path in the associated tree from the root $1^{|u|_1}2^{|u|_2}\cdots m^{|u|_m}$ to u and another one to v. By the definition of the commutator, if u is written $pbas$ with $a < b$, then $u = pab[b,a]s$. Moreover, $\pi(u) = \pi(pabs[b,a])$ since the commutators are central in $N_2(\Sigma)$.

Therefore, following backwards the path from u to the root of the tree and recalling that each edge corresponds to an exchange of 2 letters, we obtain

$$\pi(u) = \pi\left(1^{|u|_1}2^{|u|_2}\cdots m^{|u|_m}[2,1]^{\binom{u}{21}}\cdots[m,1]^{\binom{u}{m1}}\cdots[m,m-1]^{\binom{u}{m(m-1)}}\right)$$

and, similarly, following backwards the path from v to the root,

$$\pi(v) = \pi\left(1^{|v|_1}2^{|v|_2}\cdots m^{|v|_m}[2,1]^{\binom{v}{21}}\cdots[m,1]^{\binom{v}{m1}}\cdots[m,m-1]^{\binom{v}{m(m-1)}}\right).$$

But since $u \sim_2 v$, we get $\pi(u) = \pi(v)$. □

4. Order of growth

We first show that the growth of $\#(\Sigma^n/\sim_k)$ is bounded by a polynomial in n. This generalizes a result from [13] for a binary alphabet where it is shown that $\#(\{1,2\}^n/\sim_k) \in \mathcal{O}(n^{2((m-1)2^m+1)})$ for $k \geq 2$. Note that a similar result to Proposition 4.1 was obtained in [11]. That result states that for $\Sigma = \{1,\ldots,m\}$, $m \geq 2$ and $k \geq 2$, we have

$$\#(\Sigma^n/\sim_k) \in \mathcal{O}\left(n^{\frac{m}{(m-1)^2}(1+m^k(km-k-1))}\right).$$

Finally, we obtain better estimates for $\sim_2$.

Proposition 4.1. *Let $\Sigma = \{1,\ldots,m\}$ and $k \geq 1$. We have*

$$\#(\Sigma^n/\sim_k) \in \mathcal{O}\left(n^{k^2m^k}\right)$$

when n tends to infinity.

Proof. For every $u, v \in \Sigma^*$ such that $1 \leq |v| \leq k$ and $|u| = n$, we have

$$0 \leq \binom{u}{v} \leq \binom{|u|}{|v|} \leq n^{|v|} \leq n^k.$$

Therefore, for every v such that $1 \leq |v| \leq k$, we have

$$\#\left\{\binom{u}{v} : |u| = n\right\} \leq n^k + 1.$$

By definition, the $\sim_k$-equivalence class of u is uniquely determined by the values of $\binom{u}{v}$ for all $v \in \Sigma^*$ such that $1 \leq |v| \leq k$. There are

$$\sum_{i=1}^{k} m^i \leq km^k$$

such coefficients and thus,

$$\#(\Sigma^n/\sim_k) \leq \left(n^k+1\right)^{km^k}. \qquad \square$$

We have obtained an upper bound which is far from being optimal but it ensures that the growth is polynomial. However, for $k = 2$, it is possible to obtain the polynomial degree of the growth. We make use of Landau notation: $f \in \Theta(g)$ if there exist constants $A, B > 0$ such that, for all n large enough, $A\,g(n) \leq f(n) \leq B\,g(n)$.

Proposition 4.2. *Let $\Sigma = \{1, \ldots, m\}$ be an alphabet of size $m \geq 2$. We have*

$$\#(\Sigma^n/\sim_2) \in \Theta\left(n^{m^2-1}\right)$$

when n tends to infinity.

Proof. We are first going to show that $\#(\Sigma^n/\sim_2) \in \mathcal{O}\left(n^{m^2-1}\right)$ when n tends to infinity. Let f be the function such that for any $x \in \mathbb{N}^m$,

$$f(x) = \#(\{u \in \Sigma^* : \Psi(u) = x\}/\sim_2).$$

In other words, $f(x)$ counts the number of 2-binomial equivalence classes whose Parikh vector is x. Let $||\cdot||_1 : \mathbb{R}^d \to \mathbb{R}$ be the ℓ_1-norm (i.e., for all vectors v, $||v||_1 = \sum_{i=1}^d |v_i|$). Clearly for all n,

$$\#(\Sigma^n/\sim_2) = \sum_{x \in \mathbb{N}^m, ||x||_1 = n} f(x). \tag{4.1}$$

For any $a, b \in \Sigma$, $a < b$, and $u \in \Sigma^*$, $\binom{u}{ba} = |u|_a|u|_b - \binom{u}{ab}$ and $\binom{u}{aa} = \binom{|u|_a}{2}$. Any word u has its equivalence class uniquely determined by the values of $|u|_a$ for all $a \in \Sigma$ and $\binom{u}{ab}$ for all $a < b \in \Sigma$. Moreover, for all $u \in \Sigma^*$ and $a < b \in \Sigma$, $\binom{u}{ab} \leq |u|_a|u|_b$. We deduce that for all $x = (x_1, \ldots, x_m) \in \mathbb{N}^m$,

$$f(x) \leq \prod_{1 \leq a < b \leq m} x_a x_b \leq \prod_{1 \leq a < b \leq m} ||x||_1^2 \leq ||x||_1^{m(m-1)}.$$

From Equation (4.1), we get that

$$\begin{aligned}\#(\Sigma^n/\sim_2) &\leq \sum_{x \in \mathbb{N}^m, ||x||_1 = n} ||x||_1^{m(m-1)} \\ &\leq n^{m(m-1)} \ \#\left\{x \in \mathbb{N}^m \,:\, ||x||_1 = n\right\}.\end{aligned}$$

Every vector of the set $\{x \in \mathbb{N}^m : ||x||_1 = n\}$ has its components in $\{0, \ldots, n\}$. Moreover, if the first $(m-1)$ components are fixed, the last one is uniquely determined since their sum must equal n. Therefore, there are at most $(n+1)^{m-1}$ such vectors and we get

$$\#(\Sigma^n/\sim_2) \leq n^{m(m-1)}(n+1)^{m-1} \leq (n+1)^{m^2-1}.$$

We conclude that $\#(\Sigma^n/\sim_2) \in \mathcal{O}\left(n^{m^2-1}\right)$ when $n \to +\infty$. It remains to get a convenient lower bound. We are going to give, for each $x \in \mathbb{N}^m$ such that $||x||_1 = n$,

a language $L(x)$ of words taking, for every $a, b \in \Sigma^2$ such that $a < b$, a quadratic number of values for the binomial coefficient $\binom{u}{ab}$. Then, using Equation (4.1), we will obtain a lower bound.

For any $a, b \in \Sigma$, $a \neq b$, and $i, j \in \mathbb{N}$, let

$$L_{a,b,i,j} = \{u \in \{a, b\}^* : |u|_a = i, |u|_b = j\}.$$

Considering all possible letter exchanges as in (3.3) from $a^i b^j$ to $b^j a^i$, the binomial coefficient $\binom{u}{ab}$ decreases by 1 at every step from ij to 0, we thus have

$$\left\{ \binom{u}{ab} : u \in L_{a,b,i,j} \right\} = \{0, 1, \ldots, ij\} \tag{4.2}$$

which is a set of cardinality $ij + 1$. For any $x \in \mathbb{N}^m$, let us consider the following language

$$L(x) = \left(\prod_{a=1}^{m} \prod_{b=a+1}^{m} L_{a,b,\lfloor \frac{x_a}{m-1} \rfloor, \lfloor \frac{x_b}{m-1} \rfloor} \right) \prod_{a=m}^{1} a^{x_a \% (m-1)},$$

where the products must be understood as language concatenations, the indices of the last product are taken in decreasing order, and $x\%y$ is the remainder of the Euclidean division of x by y.

For instance for $m = 3$,

$$L(x) = \left\{ u_{1,2} u_{1,3} u_{2,3} r_3 r_2 r_1 : \forall a, b \in \Sigma, u_{a,b} \in L_{a,b,\lfloor \frac{x_a}{2} \rfloor, \lfloor \frac{x_b}{2} \rfloor}, r_a = a^{x_a \% 2} \right\}.$$

Roughly speaking, for every $a < b$ and $u \in L(x)$, we will show with Equation (4.3) that $\binom{u}{ab}$ mostly depends on $u_{a,b}$ and with Equation (4.4) that this binomial coefficient takes a quadratic number of values (when choosing $u_{a,b}$ accordingly). Furthermore, the role of r_a words is limited to padding. Indeed, observe that for all $u \in L(x)$, $\Psi(u) = x$.

Let $x \in \mathbb{N}^m$ and $u \in L(x)$. Then, by definition, there exist words $u_{1,2}, u_{1,3}, \ldots, u_{m-1,m}$ with, for all $a < b$, $u_{a,b} \in L_{a,b,\lfloor \frac{x_a}{m-1} \rfloor, \lfloor \frac{x_b}{m-1} \rfloor}$, such that

$$u = \left(\prod_{a=1}^{m} \prod_{b=a+1}^{m} u_{a,b} \right) \prod_{a=m}^{1} a^{x_a \% (m-1)}.$$

Let i and j be two integers such that $1 \leq i < j \leq m$ and let us compute the binomial coefficient associated with ij. A subword ij either occurs in a single factor of the above product (the first two terms below), or i and j appear in two different

factors:

$$\binom{u}{ij} = \sum_{a=1}^{m}\sum_{b=a+1}^{m}\binom{u_{a,b}}{ij} + \sum_{a=m}^{1}\binom{a^{x_a\%(m-1)}}{ij} + \sum_{a<b\in\Sigma}|u_{a,b}|_i\left(\sum_{\substack{a'<b'\in\Sigma\\(a',b')>(a,b)}}|u_{a',b'}|_j + \sum_{b\in\Sigma}\left|b^{x_b\%(m-1)}\right|_j\right) + \sum_{a=1}^{m}\sum_{b=1}^{a-1}\left|a^{x_a\%(m-1)}\right|_i\left|b^{x_b\%(m-1)}\right|_j.$$

Observe that by definition of $L(x)$, the second and last terms vanish. Hence,

$$\binom{u}{ij} = \binom{u_{i,j}}{ij} + \underbrace{\sum_{\substack{a<b\in\Sigma\\ a=i \text{ or } b=i}}\left\lfloor\frac{x_i}{m-1}\right\rfloor\left(\left(\sum_{\substack{a'<b'\in\Sigma\\(a',b')>(a,b)\\a'=j \text{ or } b'=j}}\left\lfloor\frac{x_j}{m-1}\right\rfloor\right) + x_j\%(m-1)\right)}_{:=h_{i,j}(x)} \tag{4.3}$$

The second term of the latter expression is uniquely a function of x (there is no dependency on u) while, from (4.2),

$$\left\{\binom{u_{i,j}}{ij} : u_{i,j} \in L_{i,j,\lfloor\frac{x_i}{m-1}\rfloor,\lfloor\frac{x_j}{m-1}\rfloor}\right\} = \left\{0,1,\ldots,\left\lfloor\frac{x_i}{m-1}\right\rfloor\left\lfloor\frac{x_j}{m-1}\right\rfloor+1\right\}.$$

Thus for a fixed x, considering all $u \in L(x)$, $\binom{u}{ij}$ can take

$$\left\lfloor\frac{x_i}{m-1}\right\rfloor\left\lfloor\frac{x_j}{m_1}\right\rfloor + 1 \tag{4.4}$$

different values. Moreover, for all $\left(a_{1,2}, a_{1,3}, \ldots, a_{(m-1),m}\right)$ such that

$$a_{i,j} \in \left\{h_{i,j}(x), \ldots, h_{i,j}(x) + \left\lfloor\frac{x_i}{m-1}\right\rfloor\left\lfloor\frac{x_j}{m-1}\right\rfloor\right\} \qquad \forall i<j,$$

there exists $u \in L(x)$ such that $\binom{u}{ij} = a_{i,j}$ for all $i<j$. We deduce that, for all x,

$$f(x) \geq \prod_{a<b\in\Sigma}\left(\left\lfloor\frac{x_a}{m-1}\right\rfloor\left\lfloor\frac{x_b}{m-1}\right\rfloor+1\right).$$

By Equation (4.1), we finally get the lower bound:

$$\begin{aligned}
\#(\Sigma^n/\sim_2) &\geq \sum_{\substack{x\in\mathbb{N}^m \\ ||x||_1=n}} \prod_{a<b\in\Sigma} \left(\left\lfloor \frac{x_a}{m-1}\right\rfloor \left\lfloor \frac{x_b}{m-1}\right\rfloor + 1\right) \\
&\geq \sum_{\substack{x\in\mathbb{N}^m \\ ||x||_1=n \\ \forall i, x_i \geq \frac{n}{2m}+m}} \prod_{a<b\in\Sigma} \left\lfloor \frac{x_a}{m-1}\right\rfloor \left\lfloor \frac{x_b}{m-1}\right\rfloor \\
&\geq \sum_{\substack{x\in\mathbb{N}^m \\ ||x||_1=n \\ \forall i, x_i \geq \frac{n}{2m}+m}} \prod_{a<b\in\Sigma} \left(\frac{n}{2m(m-1)}\right)^2 \\
&\geq \left(\frac{n}{2m(m-1)}\right)^{m(m-1)} \#\left\{x\in\mathbb{N}^m : ||x||_1 = n \ \wedge\ \forall i, x_i \geq \frac{n}{2m}+m\right\}.
\end{aligned}$$

The latter set contains the set

$$\left\{x\in\mathbb{N}^m : ||x||_1 = n \quad \wedge \quad \forall i, x_i \geq \frac{n}{2m}+m \quad \wedge \quad \forall i < m, x_i \leq \frac{n}{m}\right\}.$$

For n large enough (i.e., $\frac{n}{2m}+m \leq \frac{n}{m}$), the cardinal of this set is

$$\left(\frac{n}{2m} - m + 1\right)^{m-1} \in \Theta(n^{m-1}).$$

Moreover,

$$\left(\frac{n}{2m(m-1)}\right)^{m(m-1)} \in \Theta\left(n^{m(m-1)}\right)$$

and we conclude that $\#(\Sigma^n/\sim_2) \in \Theta\left(n^{m^2-1}\right)$. □

Remark 4.3. Note that even though the growth of $\#(\{1,2,3\}^n/\sim_2)$ is polynomial, this quantity is not a polynomial. It is easy to verify by interpolating the 9 first values. A similar result can be obtained for $\#(\{1,2\}^n/\sim_3)$ whose first values can be found as entry `A258585` in Sloane's encyclopedia.

5. Non context-freeness

In this section, we show that for any alphabet Σ of size at least 3 and for any $k \geq 2$, the languages $\mathsf{LL}(\sim_k, \Sigma)$ and $\mathsf{Sing}(\sim_k, \Sigma)$ are not context-free.

Let $L \subseteq \Sigma^*$ be a language. The *growth function* of L maps the integer n to $\#(L \cap \Sigma^n)$. A language has a *polynomial growth* if there exists a polynomial p such that $\#(L \cap \Sigma^n) \leq p(n)$ for all $n \geq 0$. Recall that a language L is *bounded* if there exist words $w_1, \ldots, w_\ell \in \Sigma^*$ such that $L \subseteq w_1^* w_2^* \cdots w_\ell^*$. Ginsburg and Spanier have obtained many results about bounded context-free languages, see [5]. We will make use of the following result. For relevant bibliographic pointers see, for instance, [4].

Proposition 5.1. *A context-free language is bounded if and only if it has a polynomial growth.*

We easily deduce[a] from the previous section that both languages $\mathsf{LL}(\sim_k, \Sigma)$ and $\mathsf{Sing}(\sim_k, \Sigma)$ have polynomial growth; it is thus enough to show that they are not bounded to infer that they are not context-free. Observe that, in our forthcoming reasonings, we will define particular words $\rho_{p,n}$ over a ternary alphabet (they can trivially be seen as words over a larger alphabet).

Definition 5.2. *Fix a sequence $(s_n)_{n\geq 1}$ of positive integers such that, for all $n \in \mathbb{N}$,*

$$\sqrt{\frac{s_n}{2}} \in \mathbb{N}, \tag{D1}$$

$$s_n > \left(\sqrt{\frac{s_n}{2}} + \sum_{i=1}^{n-1} s_i\right)^2, \tag{D2}$$

$$\sqrt{\frac{s_n}{2}} > \left(\sum_{i=1}^{n-1} s_i\right)\left(\sum_{i=1}^{n-3} s_i\right). \tag{D3}$$

For instance, to get a sequence with those prescribed properties, one can choose

$$s_n = 2 \times 8^{8^n}.$$

For any integers n and p, let us define the word

$$\rho_{p,n} = 1^p 2^{s_{n-1}} 3^{s_{n-2}} 1^{s_{n-3}} \cdots a^{s_1}$$

over $\{1,2,3\}$, where $a \equiv n \pmod 3$.

In Section 5.1, we prove that the $\sim_2$-class of any $\rho_{p,n}$ is a singleton. Then, it is proven in Section 5.2 that $\{\rho_{p,n} \mid p, n \in \mathbb{N}\}$ is not a bounded language. Putting together these results, we get the following.

Theorem 5.3. *For any alphabet Σ of size at least 3 and for any $k \geq 2$, the languages $\mathsf{LL}(\sim_k, \Sigma)$ and $\mathsf{Sing}(\sim_k, \Sigma)$ are not context-free.*

Proof. First note that

$$\{\rho_{p,n} \mid p, n \in \mathbb{N}\} \subseteq \{1,2,3\}^* \subseteq \Sigma^*.$$

Taking into account Corollary 5.6, observe that

$$\{\rho_{p,n} \mid p, n \in \mathbb{N}\} \subseteq \mathsf{Sing}(\sim_k, \Sigma) \subseteq \mathsf{LL}(\sim_k, \Sigma).$$

From Proposition 4.1, the languages $\mathsf{LL}(\sim_k, \Sigma)$, and thus $\mathsf{Sing}(\sim_k, \Sigma)$, have a polynomial growth. From Lemma 5.9, the language $\{\rho_{p,n} \mid p, n \in \mathbb{N}\}$ is not bounded. Therefore, $\mathsf{Sing}(\sim_k, \Sigma)$ and $\mathsf{LL}(\sim_k, \Sigma)$ are not bounded and we conclude from Proposition 5.1. □

[a] $\mathsf{Sing}(\sim_k, \Sigma) \subseteq \mathsf{LL}(\sim_k, \Sigma)$ and $\mathsf{LL}(\sim_k, \Sigma)$ is in one-to-one correspondence with $\Sigma^*/\sim_k$.

Remark 5.4. This result is in fact true for all languages having exactly one representant of each $\sim_k$-class.

5.1. *A family of singletons*

Proposition 5.5. *For any two positive integers n and p and word u, at least one of the following is false:*

- $u \neq \rho_{p,n}$,
- $\Psi(u) = \Psi(\rho_{p,n})$,
- $\binom{u}{12} \geq \binom{\rho_{p,n}}{12}$,
- $\binom{u}{23} \geq \binom{\rho_{p,n}}{23}$,
- $\binom{u}{31} \geq \binom{\rho_{p,n}}{31}$.

As an immediate corollary, we get the following result.

Corollary 5.6. *For any two positive integers n and p and word u such that $u \neq \rho_{p,n}$, we have $u \not\sim_2 \rho_{p,n}$.*

Proof of Proposition 5.5. Let us show the proposition by induction on n. The result clearly holds for $n \leq 3$. Let $n \geq 4$ be an integer such that the result holds for any $i < n$. Now let us proceed by contradiction to show that the result also holds for n.

For the sake of contradiction, let p and u be such that

$$u \neq \rho_{p,n}, \tag{5.1}$$

$$\Psi(u) = \Psi(\rho_{p,n}), \tag{5.2}$$

$$\binom{u}{12} \geq \binom{\rho_{p,n}}{12}, \tag{5.3}$$

$$\binom{u}{23} \geq \binom{\rho_{p,n}}{23}, \tag{5.4}$$

$$\binom{u}{31} \geq \binom{\rho_{p,n}}{31}. \tag{5.5}$$

Let $u = vw$ where v is the prefix of length $p + s_{n-1} - \sqrt{\frac{s_{n-1}}{2}}$ of u. Similarly, let $\rho_{p,n} = v'w'$ where $|v| = |v'|$. In the first part of the proof, we show that $\Psi(v) = \Psi(v')$ and more precisely $|v|_1 = p = |v'|_1$, $|v|_3 = 0 = |v'|_3$. We proceed into three steps.

- Proof of $|v|_1 \geq p$:

For the sake of contradiction, suppose $|v|_1 \leq p-1$. Then

$$\begin{aligned}\binom{u}{12} &= \binom{v}{12} + \binom{w}{12} + |v|_1|w|_2 \\ &\leq |v|_1|v|_2 + |w|_1|w|_2 + |v|_1|w|_2 \\ &\leq |v|_1|u|_2 + |w|_1|w|_2 \\ &\leq (p-1)|u|_2 + |w|^2.\end{aligned}$$

Replacing $|w|$ by its value, we get

$$\binom{u}{12} \leq p|u|_2 + \left(\sqrt{\frac{s_{n-1}}{2}} + \sum_{i=1}^{n-2} s_i\right)^2 - |u|_2. \tag{5.6}$$

By (5.2), $|u|_2 = |\rho_{p,n}|_2$ and condition (D2) implies

$$0 > \left(\sqrt{\frac{s_{n-1}}{2}} + \sum_{i=1}^{n-2} s_i\right)^2 - s_{n-1} \geq \left(\sqrt{\frac{s_{n-1}}{2}} + \sum_{i=1}^{n-2} s_i\right)^2 - |u|_2.$$

Together with (5.6), it gives $\binom{u}{12} < p|\rho_{p,n}|_2 \leq \binom{\rho_{p,n}}{12}$. This is a contradiction with hypothesis (5.3) and we conclude that $|v|_1 \geq p$.

• Proof of $|v|_3 = 0$:
For the sake of contradiction, suppose $|v|_3 \geq 1$.

$$\begin{aligned}\binom{u}{23} &= |u|_2|u|_3 - \binom{u}{32} \\ &= |u|_2|u|_3 - \binom{v}{32} - \binom{w}{32} - |v|_3|w|_2 \\ &\leq |u|_2|u|_3 - |v|_3|w|_2 \\ &\leq |u|_2|u|_3 - |w|_2. \end{aligned} \tag{5.7}$$

Observe that

$$\begin{aligned}|w|_2 &= |u|_2 - |v|_2 \\ &= |u|_2 - |v| + |v|_1 + |v|_3 \\ &> |u|_2 - |v| + |v|_1\end{aligned}$$

and

$$\begin{aligned}|u|_2 - |v| + |v|_1 &= |u|_2 - p - s_{n-1} + \sqrt{\frac{s_{n-1}}{2}} + |v|_1 \\ &\geq (|u|_2 - s_{n-1}) + (|v|_1 - p) + \sqrt{\frac{s_{n-1}}{2}} \\ &\geq \sqrt{\frac{s_{n-1}}{2}}\,.\end{aligned}$$

Moreover, by (5.2), $|u|_2|u|_3 = |\rho_{p,n}|_2|\rho_{p,n}|_3$. We can use these two remarks in inequality (5.7).

$$\begin{aligned}\binom{u}{23} &< |\rho_{p,n}|_2|\rho_{p,n}|_3 - \sqrt{\frac{s_{n-1}}{2}} \\ &< |\rho_{p,n}|_2|\rho_{p,n}|_3 - \left(\sum_{i=1}^{n-2} s_i\right)\left(\sum_{i=1}^{n-4} s_i\right) \text{ (from (D3))} \\ &< |\rho_{p,n}|_2|\rho_{p,n}|_3 - \sum_{i=0}^{\lfloor\frac{n-3}{3}\rfloor}\left(s_{n-2-3i}\sum_{j=i}^{\lfloor\frac{n-5}{3}\rfloor} s_{n-4-3j}\right)\end{aligned}$$

The latter quantity is equal to $|\rho_{p,n}|_2|\rho_{p,n}|_3 - \binom{\rho_{p,n}}{32}$ and thus

$$\binom{u}{23} < \binom{\rho_{p,n}}{23}.$$

This contradicts hypothesis (5.4) and we conclude that $|v|_3 = 0$.

• Proof of $|v|_1 \leq p$:
For the sake of contradiction, suppose $|v|_1 \geq p+1$. Then

$$\binom{u}{13} \geq |v|_1|w|_3 \geq p|w|_3 + |w|_3.$$

Since $|v|_3 = 0$, $|w|_3 = |u|_3 = |\rho_{p,n}|_3 \geq s_{n-2} > \sqrt{\frac{s_{n-2}}{2}}$. From condition (D3), taking into account the structure of $\rho_{p,n}$, we deduce

$$\binom{u}{13} > p|\rho_{p,n}|_3 + \left(\sum_{i=1}^{n-3} s_i\right)\left(\sum_{i=1}^{n-5} s_i\right) \geq \binom{\rho_{p,n}}{13}.$$

This yields a contradiction with hypothesis (5.5). We conclude that $|v|_1 = p$. We thus have

$$\Psi(v') = \Psi(v) \text{ and } \Psi(w') = \Psi(w). \tag{5.8}$$

We will now use Equation (5.8) to find the contradiction; we are going to show that v' and w' are shorter words for which the result doesn't hold. From hypothesis (5.3), we get

$$\begin{aligned}\binom{w}{12} &= \binom{u}{12} - \binom{v}{12} - |v|_1|w|_2 \\ &\geq \binom{\rho_{p,n}}{12} - \binom{v}{12} - |v|_1|w|_2.\end{aligned}$$

Since $\rho_{p,n} = v'w'$, this latter quantity is equal to

$$\binom{v'}{12} + \binom{w'}{12} + |v'|_1|w'|_2 - \binom{v}{12} - |v|_1|w|_2 = \binom{v'}{12} + \binom{w'}{12} - \binom{v}{12}$$

where the last equality is due to (5.8). We thus obtain that

$$\binom{w}{12} \geq \binom{v'}{12} + \binom{w'}{12} - \binom{v}{12}. \tag{5.9}$$

Since v' is of the form $1^\alpha 2^\beta$,

$$\binom{v'}{12} = \max\left\{\binom{x}{12} \,:\, \Psi(x) = \Psi(v')\right\} \geq \binom{v}{12}$$

and we get $\binom{w}{12} \geq \binom{w'}{12}$. With similar arguments and the fact that $|v|_3 = 0 = |v'|_3$, we obtain $\binom{w}{23} \geq \binom{w'}{23}$ and $\binom{w}{31} \geq \binom{w'}{31}$.

Observe that $w \neq w'$. Indeed, it is obvious if $v = v'$ (since $vw \neq v'w'$) and otherwise, $\binom{v'}{12} > \binom{v}{12}$ and from (5.9), we get $\binom{w}{12} > \binom{w'}{12}$.

Let σ be the morphism such that $\sigma(1) = 3; \sigma(2) = 1; \sigma(3) = 2$. Then $\sigma(w') = \rho_{\sqrt{\frac{s_{n-1}}{2}},n-1}$ and since σ is a permutation of the alphabet, we get

- $\sigma(w) \neq \sigma(w')$,
- $\Psi(\sigma(w)) = \Psi(\sigma(w'))$,
- $\binom{\sigma(w)}{12} \geq \binom{\sigma(w')}{12}$,
- $\binom{\sigma(w)}{23} \geq \binom{\sigma(w')}{12}$,
- $\binom{\sigma(w)}{31} \geq \binom{\sigma(w')}{31}$.

This is a contradiction with our induction hypothesis. We deduce that there is no such pair of integers and this concludes the proof of the proposition. □

5.2. *Unboundedness*

It remains us to prove that the language $\{\rho_{p,n} : p, n \in \mathbb{N}\}$ is not bounded. We will make use of the following notation. For all non-empty words $w \in \Sigma^+$, its *letter-factorization* is $(c_1, q_1), \ldots, (c_r, q_r)$, where

$$w = c_1^{q_1} c_2^{q_2} \cdots c_r^{q_r},$$

$r \geq 1$, $c_1, \ldots, c_r$ are letters such that for all i, $c_i \neq c_{i+1}$, and where $q_1, \ldots, q_r$ are positive integers. The *number of blocks* in the word w, denoted by $nb(w)$, is r. It corresponds to the length of the decomposition.

Example 5.7. Let $w = 112333122132$. We have $c_1 = 1$, $c_2 = 2$, $c_3 = 3$, $c_4 = 1$, $c_5 = 2$, $c_6 = 1$, $c_7 = 3$, $c_8 = 2$, and $q_1 = 2$, $q_2 = 1$, $q_3 = 3$, $q_4 = 1$, $q_5 = 2$, $q_6 = q_7 = q_8 = 1$. Moreover, $nb(w) = 8$.

The letter-factorization of a word of the form $\rho_{p,n}$ has particular properties that we record in the following remark.

Remark 5.8. For all $p, n \in \mathbb{N}$, if $(c_1, q_1), \ldots, (c_r, q_r)$ is the letter-factorization of $\rho_{p,n}$, we know that

- for all $i \geq 1$, $c_i \equiv i \pmod 3$, with $c_i \in \{1, 2, 3\}$;

- $q_1 = p$ and for all $i > 1$, $q_i = s_{n-i+1}$;
- $nb(\rho_{p,n}) = n$.

Lemma 5.9. *For all $\ell \in \mathbb{N}$ and words $w_1, \dots, w_\ell \in \Sigma^*$, we have*

$$\{\rho_{p,n} : p, n \in \mathbb{N}\} \not\subset w_1^* \cdots w_\ell^*.$$

Proof. For the sake of contradiction, let us assume that there exist $\ell \in \mathbb{N}$ and words $w_1, \dots, w_\ell \in \Sigma^*$ such that

$$\mathcal{R} := \{1^p 2^{s_{n-1}} 3^{s_{n-2}} \cdots : p \in \mathbb{N}, n \in \mathbb{N}\} \subseteq w_1^* \cdots w_\ell^*.$$

We will first show that, under this assumption, there exist $N \in \mathbb{N}$ and words $z_1, \dots, z_q$ such that, for all i, $nb(z_i) \leq 2$, and the subset

$$\mathcal{R}_N := \{\rho_{p,n} : p \in \mathbb{N}, n \geq N\}$$

of $\mathcal{R}$ is included in $z_1^* \dots z_q^*$.

Let us take the least $i \in \{1, \dots, \ell\}$ such that $nb(w_i) \geq 3$. If such an i does not exist, we can take $N = 0$, $\ell = q$ and $z_i = w_i$ for all i. Otherwise, the letter-factorization of w_i begins with $(a_1, \alpha_1), (a_2, \alpha_2), (a_3, \alpha_3)$. Assume that there exist p, n such that the factorization of $\rho_{p,n}$ in terms of $w_1, \dots, w_\ell$

$$\rho_{p,n} = w_1^{n_1} \cdots w_i^{n_i} \cdots w_\ell^{n_\ell}$$

contains an occurrence of w_i, i.e., $n_i > 0$, (if this is not the case, $\mathcal{R}$ is thus included in $w_1^* \cdots w_{i-1}^* w_{i+1}^* \cdots w_\ell^*$ and we can proceed to the next index such that $nb(w_i) \geq 3$). Because of Remark 5.8, if $nb(w_j) = 2$ then $w_j w_j$ is never a factor of a word in $\mathcal{R}$ (this would mean that two letters out of three are alternating). In that case, we must have $n_j = 1$ in the above factorization. Also, if $nb(w_j) = 1$, then $nb(w_j^{n_j}) = 1$. By definition of i, $nb(w_j) \leq 2$ for all $j < i$. Therefore there exists $\gamma \leq 2i$ such that, if $(c_1, q_1), \dots, (c_r, q_r)$ is the letter-factorization of $\rho_{p,n}$,

- $c_1^{q_1} \cdots c_{\gamma-1}^{q_{\gamma-1}} \in w_1^* \cdots w_{i-1}^* a_1^{\alpha_1}$,
- $c_\gamma = a_2$ and $q_\gamma = \alpha_2$,
- $c_{\gamma+1} = a_3$.

See Figure 3 for an illustration.

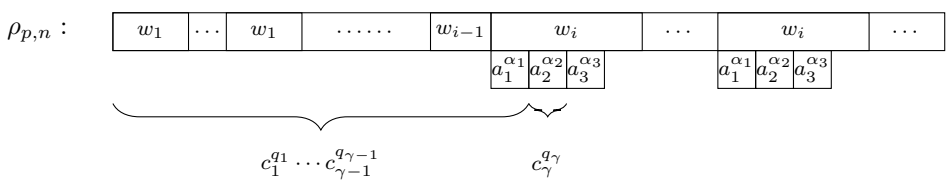

Fig. 3: Decomposition of $\rho_{p,n}$ into blocks

With Remark 5.8, we know that if $\gamma = 2$ then $q_{\gamma-1} = p$ and in all cases, $q_\gamma = s_{n-\gamma+1}$. Therefore, if we take N such that $s_{N-2i+1} > \alpha_2$, the set $\mathcal{R}_N$, which is included in $w_1^* \dots w_\ell^*$ is also included in $w_1^* \dots w_{i-1}^* w_{i+1}^* \dots w_\ell^*$. We can proceed the same way to eliminate other factors w_j with $nb(w_j) \geq 3$ to finally obtain an integer N such that

$$\mathcal{R}_N = \{\rho_{p,n} : p \in \mathbb{N}, n \geq N\}$$

is included in a set of the form $z_1^* \dots z_q^*$ where, for all i, $nb(z_i) \leq 2$.

It remains to show that this observation leads to a contradiction. Let $\rho_{p,n} \in \mathcal{R}_N$. It can be factorized as $z_1^{n_1} \cdots z_q^{n_q}$. We have already observed that if $nb(z_i) = 2$, then $n_i = 1$. Otherwise, $nb(z_i) = 1$ and thus $nb(z_i^{n_i}) = 1$. For this reason, we obtain that for all $n \geq N$,

$$nb(\rho_{p,n}) \leq 2q,$$

which is a contradiction because $nb(\rho_{p,n}) = n$ and this concludes the proof. □

6. Conclusions

As we have seen, there is a simple switch operation given by $\equiv$ that permits us to easily describe the 2-binomial equivalence class of a word over a binary alphabet. One could try to generalize this operation over larger alphabets or for $k \geq 3$, but the question has no clear answer yet.

However, over a larger alphabet, we gave algorithmic and algebraic descriptions of the 2-binomial classes. A natural question is to extend these results for $k \geq 3$.

We proved that $\mathsf{LL}(\sim_k, \Sigma)$ is not context-free if $k \geq 2$ and $\#\Sigma \geq 3$. We know that $\mathsf{LL}(\sim_2, \{1,2\})$ is context-free. However, the question is still open about $\mathsf{LL}(\sim_k, \{1,2\})$ with $k \geq 3$. It seems that a method similar to the one carried in Section 5 could work, but it remains to find an unbounded set of singletons.

When $\mathsf{LL}(\sim_k, \Sigma)$ is not context-free, a measure of descriptional complexity is the so-called automaticity [15]. Let L be a language and C, t be integers. The idea is that we only know the words of L of length at most C. Consider the following approximation of Nerode congruence: for any two words u, v such that $|u|, |v| \leq t$,

$$u \approx_{L,C,t} v \;\Leftrightarrow\; \left(u^{-1}\left(L \cap \Sigma^{\leq C}\right)\right) \cap \Sigma^{\leq C-t} = \left(v^{-1}\left(L \cap \Sigma^{\leq C}\right)\right) \cap \Sigma^{\leq C-t}.$$

The quantity $\#\left(\Sigma^{\leq t}/\approx_{L,C,t}\right)$ gives a lower approximation of the automaticity of L. For $L = \mathsf{LL}(\sim_3, \{1,2\})$, $C = 15$ and $t = 1, 2, \dots, 9$, the first few values are

$$1, 3, 5, 9, 16, 27, 49, 88, 154.$$

For $L = \mathsf{LL}(\sim_2, \{1,2,3\})$, $C = 9$ and $t = 1, 2, \dots, 6$, they are

$$1, 4, 8, 19, 42, 62.$$

Can the automaticity of such languages be characterized or estimated?

Acknowledgments

We thank the anonymous referee. His/her careful reading and comments were useful to improve the general presentation of this paper.

References

[1] J. Cassaigne, J. Karhumäki, S. Puzynina, and M.A. Whiteland, k-abelian equivalence and rationality, *Fund. Infor.* **154** (2017), 65–94.

[2] S. Fossé, and G. Richomme, Some characterizations of Parikh matrix equivalent binary words, *Inform. Process. Lett.* **92** (2004), 77–82.

[3] D. D. Freydenberger, P. Gawrychowski, J. Karhumäki, F. Manea, and W. Rytter, Testing k-binomial equivalence, in *Multidisciplinary Creativity: homage to Gheorghe Paun on his 65th birthday*, 239–248, Ed. Spandugino, Bucharest, Romania (2015).

[4] P. Gawrychowski, D. Krieger, N. Rampersad, and J. Shallit, Finding the growth rate of a regular of context-Free language in polynomial time, In: Ito M., Toyama M. (eds) Developments in Language Theory. DLT 2008, *Lect. Notes in Comp. Sci.* **5257** (2008). Springer, Berlin, Heidelberg

[5] S. Ginsburg, and E. Spanier, Bounded ALGOL-like languages, *Trans. Amer. Math. Soc.* **113** (1964), 333–368.

[6] R. Incitti, The growth function of context-free languages, *Theoret. Comput. Sci.* **255** (2001), 601–605.

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

[8] J. Karhumäki, S. Puzynina, M. Rao, and M. Whiteland, On cardinalities of k-abelian equivalence classes, *Theoret. Comput. Sci.* **658** (2017), 190–204.

[9] J. Karhumäki, A. Saarela, and L. Q. Zamboni, On a generalization of Abelian equivalence and complexity of infinite words, *J. Combin. Theory Ser. A* **120** (2013), 2189–2206.

[10] J. Karhumäki, A. Saarela, and L. Q. Zamboni, Variations of the Morse-Hedlund theorem for k-abelian equivalence, *Lect. Notes in Comput. Sci.* **8633** (2014), 203–214.

[11] M. Lejeune, *Au sujet de la complexité k-binomiale*, Master thesis, University of Liège (2018), `http://hdl.handle.net/2268.2/5007`.

[12] A. Mateescu, A. Salomaa, K. Salomaa, and S. Yu, A sharpening of the Parikh mapping, *RAIRO-Theor. Inf. Appl.* **35** (2001), 551–564.

[13] M. Rigo, and P. Salimov, Another generalization of abelian equivalence: binomial complexity of infinite words, *Theoret. Comput. Sci.* **601** (2015), 47–57.

[14] A. Salomaa, Criteria for the matrix equivalence of words, *Theoret. Comput. Sci.* **411** (2010), 1818–1827.

[15] J. Shallit, and Y. Breitbart, Automaticity. I. Properties of a measure of descriptional complexity, *J. Comput. System Sci.* **53** (1996), no. 1, 10–25.

[16] M. A. Whiteland, *On the k-Abelian Equivalence Relation of Finite Words*, Ph.D. Thesis, TUCS Dissertations **241**, Univ. of Turku (2019).