Another Generalization of Abelian Equivalence: Binomial Complexity of Infinite Words

M. Rigo[1] and P. Salimov[1,2]*

[1] Dept of Math., University of Liège, Grande traverse 12 (B37), B-4000 Liège, Belgium, M.Rigo@ulg.ac.be

[2] Sobolev Institute of Math., 4 Acad. Koptyug avenue, 630090 Novosibirsk, Russia.

Abstract. The binomial coefficient of two words u and v is the number of times v occurs as a subsequence of u. Based on this classical notion, we introduce the m-binomial equivalence of two words refining the abelian equivalence. The m-binomial complexity of an infinite word x maps an integer n to the number of m-binomial equivalence classes of factors of length n occurring in x. We study the first properties of m-binomial equivalence. We compute the m-binomial complexity of the Sturmian words and of the Thue–Morse word. We also mention the possible avoidance of 2-binomial squares.

1 Introduction

In the literature, many measures of complexity of infinite words have been introduced. One of the most studied is the factor complexity p_x counting the number of distinct blocks of n consecutive letters occurring in an infinite word $x \in A^{\mathbb{N}}$. In particular, Morse–Hedlund theorem gives a characterization of ultimately periodic words in terms of bounded factor complexity. Sturmian words have a null topological entropy and are characterized by the relation $p_x(n) = n + 1$ for all $n \geqslant 0$. Abelian complexity counts the number of distinct Parikh vectors for blocks of n consecutive letters occurring in an infinite word, i.e., factors of length n are counted up to abelian equivalence. Already in 1961, Erdős opened the way to a new research direction by raising the question of avoiding abelian squares in arbitrarily long words [6]. Related to Van der Waerden theorem, we can also mention the arithmetic complexity [1] mapping $n \geqslant 0$ to the number of distinct subwords $x_i x_{i+p} \cdots x_{i+(n-1)p}$ built from n letters arranged in arithmetic progressions in the infinite word x, $i \geqslant 0$, $p \geqslant 1$. In the same direction, one can also consider maximal pattern complexity [7].

As a generalization of abelian complexity, the k-abelian complexity was recently introduced through a hierarchy of equivalence relations, the coarsest being abelian equivalence and refining up to equality. We recall these notions.

* The second author is supported by the Russian President's grant no. MK-4075.2012.1 and Russian Foundation for Basic Research grants no. 12-01-00089 and no. 11-01-00997 and by a University of Liège post-doctoral grant.

Let $k \in \mathbb{N} \cup \{+\infty\}$ and A be a finite alphabet. As usual, $|u|$ denotes the length of u and $|u|_x$ denotes the number of occurrences of the word x as a factor of the word u. Karhumäki *et al.* [8] introduce the notion of *k-abelian equivalence* of finite words as follows. Let u, v be two words over A. We write $u \sim_{\mathrm{ab},k} v$ if and only if $|u|_x = |v|_x$ for all words x of length $|x| \leqslant k$. In particular, $u \sim_{\mathrm{ab},1} v$ means that u and v are *abelian equivalent*, i.e., u is obtained by permuting the letters in v.

The aim of this paper is to introduce and study the first properties of a different family of equivalence relations over A^*, called k-binomial equivalence, where the coarsest relation coincide with the abelian equivalence.

Let $u = u_0 \cdots u_{n-1}$ be a word of length n over A. Let $\ell \leqslant n$. Let $t : \mathbb{N} \to \mathbb{N}$ be an increasing map such that $t(\ell - 1) < n$. Then the word $u_{t(0)} \cdots u_{t(\ell-1)}$ is a *subword* of length ℓ of u. Note that what we call subword is also called scattered subword in the literature. The notion of *binomial coefficient* of two finite words u and v is well-known, $\binom{u}{v}$ is defined as the number of times v occurs as a subword of u. In other words, the binomial coefficient of u and v is the number of times v appears as a subsequence of u. Properties of these coefficients are presented in the chapter of Lothaire's book written by Sakarovitch and Simon [12, Section 6.3]. Let $a, b \in A$, $u, v \in A^*$ and p, q be integers. We set $\delta_{a,b} = 1$ if $a = b$, and $\delta_{a,b} = 0$ otherwise. We just recall that

$$\binom{a^p}{a^q} = \binom{p}{q},\ \binom{u}{\varepsilon} = 1,\ |u| < |v| \Rightarrow \binom{u}{v} = 0,\ \binom{ua}{vb} = \binom{u}{vb} + \delta_{a,b}\binom{u}{v}$$

and the last three relations completely determine the binomial coefficient $\binom{u}{v}$ for all $u, v \in A^*$.

Remark 1. Note that we have to make a distinction between subwords and factors. A factor is a particular subword made of consecutive letters. Factors of u are denoted either by $u_i \cdots u_j$ or $u[i, j]$, $0 \leqslant i \leqslant j < |u|$.

Definition 1. *Let $m \in \mathbb{N} \cup \{+\infty\}$ and u, v be two words over A. We say that u and v are* m-binomially equivalent *if*

$$\binom{u}{x} = \binom{v}{x},\ \forall x \in A^{\leqslant m}.$$

Since the main relation studied in this paper is the m-binomial equivalence, we simply write in that case: $u \sim_m v$.

Since $\binom{u}{a} = |u|_a$ for all $a \in A$, it is clear that two words u and v are abelian equivalent if and only if $u \sim_1 v$. As for abelian equivalence, we have a family of refined relations: for all $u, v \in A^*$, $m \geqslant 0$, $u \sim_{m+1} v \Rightarrow u \sim_m v$.

Example 1. For instance, the four words *ababbba*, *abbabab*, *baabbab* and *babaabb* are 2-binomially equivalent. For any w amongst these words, we have the following coefficients

$$\binom{w}{a} = 3,\ \binom{w}{b} = 4,\ \binom{w}{aa} = 3,\ \binom{w}{ab} = 7,\ \binom{w}{ba} = 5,\ \binom{w}{bb} = 6.$$

But one can check that they are not 3-binomially equivalent, as an example,

$$\binom{ababbba}{aab} = 3 \text{ but } \binom{abbabab}{aab} = 4$$

indeed, for this last binomial coefficient, aab appears as subwords $w_0w_3w_4$, $w_0w_3w_6$, $w_0w_5w_6$ and $w_3w_5w_6$. Considering again the first two words, we find $|ababbba|_{ab} = 2$ and $|abbabab|_{ab} = 3$, showing that these two words are not 2-abelian equivalent. Conversely, the words $abbaba$ and $ababba$ are 2-abelian equivalent but are not 2-binomially equivalent:

$$\binom{abbaba}{ab} = 4 \text{ but } \binom{ababba}{ab} = 5.$$

This paper is organized as follows. In the next section, we present some straightforward properties of binomial coefficients and m-binomial equivalence. In Section 3, we give upper bounds on the number of m-binomial equivalence classes partitioning A^n. Section 3 ends with the introduction of the m-binomial complexity $\mathbf{b}_x^{(m)}$ of an infinite word x. In Section 4, we prove that if x is a Sturmian word then, for any $m \geqslant 2$, $\mathbf{b}_x^{(m)}(n) = n+1$ for all $n \geqslant 0$. In Section 5 we consider the Thue–Morse word t and show that, for all $m \geqslant 1$, there exists a constant C_m such that $\mathbf{b}_t^{(m)}(n) \leqslant C_m$ for all $n \geqslant 0$. For instance, binomial coefficients of t were considered in [3]. Due to space limitations, we only give details for the cases $m = 2, 3$. In the last section, we evoke the problem of avoiding 2-binomial squares.

2 First Properties

We denote by $\mathbf{B}^{(m)}(v)$ the equivalence class of words m-binomially equivalent to v. Binomial coefficients have a nice behavior with respect to the concatenation of words.

Proposition 1. *Let p, s and $e = e_0e_1 \cdots e_{n-1}$ be finite words. We have*

$$\binom{ps}{e} = \sum_{i=0}^{n} \binom{p}{e_0e_1 \cdots e_{i-1}} \binom{s}{e_ie_{i+1} \cdots e_{n-1}}.$$

We can also mention some other basic facts on m-binomial equivalence.

Lemma 1. *Let u, u', v, v' be finite words and $m \geqslant 1$.*

- *If $u \sim_m v$, then $u \sim_\ell v$ for all $\ell \leqslant m$.*
- *If $u \sim_m v$ and $u' \sim_m v'$, then $uu' \sim_m vv'$.*

Proof. Simply note for the second point that, for all $x = x_0 \cdots x_{\ell-1}$ of length $\ell \leqslant m$, $\binom{uu'}{x}$ is equal to

$$\sum_{i=0}^{\ell} \binom{u}{x[0, i-1]} \binom{u'}{x[i, \ell-1]} = \sum_{i=0}^{\ell} \binom{v}{x[0, i-1]} \binom{v'}{x[i, \ell-1]} = \binom{vv'}{x}.$$

Remark 2. Thanks to the above lemma, we can endow the quotient set $A^*/\sim_m$ with a monoid structure using an operation $\circ : A^*/\sim_m \times A^*/\sim_m \to A^*/\sim_m$ defined by $\mathbf{B}^{(m)}(p) \circ \mathbf{B}^{(m)}(q) = \mathbf{B}^{(m)}(r)$ if the concatenation $\mathbf{B}^{(m)}(p).\mathbf{B}^{(m)}(q)$ is a subset of $\mathbf{B}^{(m)}(r)$. In particular, one can take $r = pq$. If a word v is factorized as $v = pus$, then the m-equivalence class $\mathbf{B}^{(m)}(v)$ is completely determined by p, s and $\mathbf{B}^{(m)}(u)$.

3 On the Number of k-Binomial Equivalence Classes

For 2- and 3-abelian equivalence, the number of equivalence classes for words of length n over a binary alphabet are respectively $n^2 - n + 2$ and $\Theta(n^4)$. In general, for k-abelian equivalence, the number of equivalence classes for words of length n over a ℓ-letter alphabet is $\Theta(n^{(\ell-1)\ell^{k-1}})$ [8]. We consider similar results for m-binomial equivalence (proofs can be found in [15]).

Lemma 2. *Let $u \in A^*$, $a \in A$ and $\ell \geqslant 0$. We have*

$$\binom{u}{a^\ell} = \binom{|u|_a}{\ell} \quad \textit{and} \quad \sum_{|v|=\ell} \binom{u}{v} = \binom{|u|}{\ell}.$$

Lemma 3. *Let A be a binary alphabet, we have*

$$\#\left(A^n/\sim_2\right) = \sum_{j=0}^{n} ((n-j)j+1) = \frac{n^3+5n+6}{6}.$$

Proposition 2. *Let $m \geqslant 2$. Let A be a binary alphabet, we have*

$$\#\left(A^n/\sim_m\right) \in \mathcal{O}(n^{2((m-1)2^m+1)}).$$

We denote by $\mathrm{Fac}_x(n)$ the set of factors of length n occurring in x.

Definition 2. *Let $m \geqslant 1$. The m-*binomial complexity *of an infinite word x counts the number of m-binomial equivalence classes of factors of length n occurring in x,*

$$\mathbf{b}_x^{(m)} : \mathbb{N} \to \mathbb{N},\ n \mapsto \#(\mathrm{Fac}_x(n)/\sim_m).$$

Note that $\mathbf{b}_x^{(1)}$ corresponds to the usual abelian complexity denoted by ρ_x^{ab}.

If p_x denotes the usual factor complexity, then for all $m \geqslant 1$, we have

$$\mathbf{b}_x^{(m)}(n) \leqslant \mathbf{b}_x^{(m+1)}(n) \quad \text{and} \quad \rho_x^{\mathrm{ab}}(n) \leqslant \mathbf{b}_x^{(m)}(n) \leqslant p_x(n). \tag{1}$$

4 The m-Binomial Complexity of Sturmian Words

Recall that a *Sturmian word* x is a non-periodic word of minimal (factor) complexity, that is, $p_x(n) = n + 1$ for all $n \geqslant 0$. The following characterization is also useful.

Theorem 1. *[13, Theorem 2.1.5] An infinite word $x \in \{0,1\}^\omega$ is Sturmian if and only if it is aperiodic and balanced, i.e., for all factors u, v of the same length occurring in x, we have $||u|_1 - |v|_1| \leqslant 1$.*

The aim of this section is to compute the m-binomial complexity of a Sturmian word as expressed by Theorem 2. We show that any two distinct factors of length n occurring in a Sturmian words are never m-binomially equivalent. First note that Sturmian words have a constant abelian complexity. Hence, if x is a Sturmian word, then $\mathbf{b}_x^{(1)}(n) = 2$ for all $n \geqslant 1$.

Theorem 2. *Let $m \geqslant 2$. If x is a Sturmian word, then $\mathbf{b}_x^{(m)}(n) = n+1$ for all $n \geqslant 0$.*

Remark 3. If x is a right-infinite word such that $\mathbf{b}_x^{(1)}(n) = 2$ for all $n \geqslant 1$, then x is clearly balanced. If $\mathbf{b}_x^{(2)}(n) = n+1$, for all $n \geqslant 0$, then the factor complexity function p_x is unbounded and x is aperiodic. As a consequence of Theorem 2, an infinite word x is Sturmian if and only if, for all $n \geqslant 1$ and all $m \geqslant 2$, $\mathbf{b}_x^{(1)}(n) = 2$ and $\mathbf{b}_x^{(m)}(n) = n+1$.

Before proceeding to the proof of Theorem 2, we first recall some well-known fact about Sturmian words. One of the two symbols occurring in a Sturmian word x over $\{0,1\}$ is always isolated, for instance, 1 is always followed by 0. In that latter case, there exists a unique $k \geqslant 1$ such that each occurrence of 1 is always followed by either 0^k1 or $0^{k+1}1$ and x is said to be of *type* 0. See for instance [14, Chapter 6]. More precisely, we have the following remarkable fact showing that the recoding of a Sturmian sequence corresponds to another Sturmian sequence. Note that $\sigma : A^\omega \to A^\omega$ is the shift operator mapping $(x_n)_{n\geqslant 0}$ to $(x_{n+1})_{n\geqslant 0}$.

Theorem 3. *Let $x \in \{0,1\}^\omega$ be a Sturmian word of type 0. There exists a unique integer $k \geqslant 1$ and a Sturmian word $y \in \{0,1\}^\omega$ such that $x = \sigma^c(\mu(y))$ for some $c \leqslant k+1$ and where the morphism $\mu : \{0,1\}^* \to \{0,1\}^*$ is defined by $\mu(0) = 0^k1$ and $\mu(1) = 0^{k+1}1$.*

Corollary 1. *Let $x \in \{0,1\}^\omega$ be a Sturmian word of type 0. There exists a unique integer $k \geqslant 1$ such that any factor occurring in x is of the form*

$$0^r 1 0^{k+\epsilon_0} 1 0^{k+\epsilon_1} 1 \cdots 0^{k+\epsilon_{n-1}} 1 0^s \tag{2}$$

where $r, s \leqslant k+1$ and $\epsilon_0\epsilon_1 \cdots \epsilon_{n-1} \in \{0,1\}^*$ is a factor of the Sturmian word y introduced in the above theorem.

Let $\epsilon = \epsilon_0 \cdots \epsilon_{n-1}$ be a word over $\{0,1\}$. For $m \leqslant n-1$, we define

$$S(\epsilon, m) := \sum_{j=0}^{m} (n-j)\epsilon_j \quad \text{and} \quad S(\epsilon) := S(\epsilon, n-1). \tag{3}$$

Remark 4. Let $v = 0^r 1 0^{k+\epsilon_0} 1 0^{k+\epsilon_1} 1 \cdots 0^{k+\epsilon_{n-1}} 1 0^s$ of the form (2), we have

$$\binom{v}{01} = r(n+1) + \sum_{j=0}^{n-1} (k+\epsilon_j)(n-j) = r(n+1) + S(\epsilon_0 \cdots \epsilon_{n-1}) + k\frac{n(n+1)}{2}.$$

We need a technical lemma on the factors of a Sturmian word.

Lemma 4. *Let $n \geqslant 1$. If u and v are two distinct factors of length n occurring in a Sturmian word over $\{0,1\}$, then $S(u) \not\equiv S(v) \pmod{n+1}$.*

Proof. Consider two distinct factors u, v of length n occurring in a Sturmian word y. For $m < n$, we define $\Delta(m) := |u_0 u_1 \cdots u_m|_1 - |v_0 v_1 \cdots v_m|_1$. Due to Theorem 3, we have $|\Delta(m)| \leqslant 1$. Note that, if there exists i such that $\Delta(i) = 1$ then, for all $j > i$, we have $\Delta(j) \geqslant 0$. Otherwise, we would have $|v[i+1,j]|_1 - |u[i+1,j]|_1 > 1$ contradicting the fact that y is balanced. Similarly, for all $j < i$, we also have $\Delta(j) \geqslant 0$.

Since u and v are distinct, replacing u with v if needed, we may assume that there exists a minimal $i \in \{0, \ldots, n-1\}$ such that $\Delta(i) = 1$. From the above discussion and the minimality of i, $\Delta(j) = 0$ for $j < i$ and $\Delta(j) \in \{0,1\}$ for $j > i$.

From (3), for any $j < n$, we have

$$\begin{aligned}
\Delta(j+1) > \Delta(j) &\Rightarrow S(u, j+1) - S(v, j+1) = S(u,j) - S(v,j) + (n-j)\\
\Delta(j+1) = \Delta(j) &\Rightarrow S(u, j+1) - S(v, j+1) = S(u,j) - S(v,j)\\
\Delta(j+1) < \Delta(j) &\Rightarrow S(u, j+1) - S(v, j+1) = S(u,j) - S(v,j) - (n-j).
\end{aligned}$$

In view of these observations, the knowledge of $\Delta(0), \Delta(1), \ldots$ permits to compute $(S(u,j) - S(v,j))_{0 \leqslant j < n}$ and we deduce that $0 < S(u) - S(v) < n+1$ concluding the proof.

Proof (Proof of Theorem 2). Let x be a Sturmian word of type 0 and $m \geqslant 2$. From (1), we have, for all $\ell \geqslant 0$,

$$\mathbf{b}_x^{(2)}(\ell) \leqslant \mathbf{b}_x^{(m)}(\ell) \leqslant p_x(\ell) = \ell + 1.$$

We just need to show that any two distinct factors of length ℓ in x are not 2-binomially equivalent, i.e., $\ell + 1 \leqslant \mathbf{b}_x^{(2)}(\ell)$.

Proceed by contradiction. Assume that x contains two distinct factors u and v that are 2-binomially equivalent. In particular, $\binom{u}{00} = \binom{v}{00}$ and $\binom{u}{11} = \binom{v}{11}$. Hence we get $|u| = |v|$ and $|u|_1 = |v|_1 = n$. From Corollary 1, there exist $k \geqslant 1$ and a Sturmian word y such that

$$u = 0^r 1 0^{k+\epsilon_0} 1 0^{k+\epsilon_1} 1 \cdots 0^{k+\epsilon_{n-1}} 1 0^s, \quad v = 0^{r'} 1 0^{k+\epsilon'_0} 1 0^{k+\epsilon'_1} 1 \cdots 0^{k+\epsilon'_{n-1}} 1 0^{s'}$$

where $\epsilon = \epsilon_0 \epsilon_1 \cdots \epsilon_{n-1}$ and $\epsilon' = \epsilon'_0 \epsilon'_1 \cdots \epsilon'_{n-1}$ are both factors of y.

Since $u \sim_2 v$, it follows $\binom{u}{01} = \binom{v}{01}$. From Remark 4, we get

$$r(n+1) + S(\epsilon) + k\frac{n(n+1)}{2} = r'(n+1) + S(\epsilon') + k\frac{n(n+1)}{2}.$$

Otherwise stated, we get $S(\epsilon) - S(\epsilon') = (r' - r)(n+1)$ contradicting the previous lemma.

5 The Case of the Thue–Morse Word

The *Thue–Morse* word $t = 0110100110010110100101100110100 1\cdots$ is the infinite word $\lim_{n\to\infty} \varphi^n(a)$ where $\varphi : 0 \mapsto 01,\ 1 \mapsto 10$. The factor complexity of the Thue–Morse word is well-known [2, 5]: $p_t(0) = 1$, $p_t(1) = 2$, $p_t(2) = 4$ and

$$p_t(n) = \begin{cases} 4n - 2\cdot 2^m - 4 \text{ if } 2\cdot 2^m < n \leqslant 3\cdot 2^m \\ 2n + 4\cdot 2^m - 2 \text{ if } 3\cdot 2^m < n \leqslant 4\cdot 2^m \end{cases}$$

and the abelian complexity of t is obvious.

Lemma 5. *We have* $\mathbf{b}_t^{(1)}(2n) = 3$ *and* $\mathbf{b}_t^{(1)}(2n+1) = 2$ *for all* $n \geqslant 1$.

The main result of this section is the following one. It is quite in contrast with the Sturmian case because here, the Thue–Morse word exhibits a bounded m-binomial complexity.

Theorem 4. *Let* $m \geqslant 2$. *There exists* $C_m > 0$ *such that the* m*-binomial complexity of the Thue–Morse word satisfies* $\mathbf{b}_t^{(m)}(n) \leqslant C_m$ *for all* $n \geqslant 0$.

For the sake of presentation, we first show that the 2-binomial complexity of the Thue–Morse word is bounded by a constant.

Theorem 5. *There exists* $C_2 > 0$ *such that the* 2*-binomial complexity of the Thue–Morse word satisfies* $\mathbf{b}_t^{(2)}(n) \leqslant C_2$ *for all* $n \geqslant 0$.

Proof. Any factor v of t admits a factorization of the kind $p\varphi(u)s$ with $p, s \in \{0, 1, \varepsilon\}$ and where u is a factor of t. Using Remark 2, it is therefore enough to prove that, for all n,

$$\#\{\mathbf{B}^{(2)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\} \leqslant 9. \tag{4}$$

Recall from the proof of Lemma 3 that the 2-binomial equivalence class of a word v of length $2n$ over a binary alphabet $\{0, 1\}$ is completely determined by its length, $|v|_0$ and $\binom{v}{01}$, i.e.,

$$\begin{aligned} &\#\{\mathbf{B}^{(2)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\} \\ =\ &\#\{(\binom{v}{0}, \binom{v}{1}, \binom{v}{00}, \binom{v}{01}, \binom{v}{10}, \binom{v}{11}) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\} \\ =\ &\#\{(|v|_0, \binom{v}{01}) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi(u)\}. \end{aligned}$$

Fix $n \geqslant 1$. Consider an arbitrary factor $u = u_0 \cdots u_{n-1} \in \mathrm{Fac}_t(n)$ and the corresponding factor $v = \varphi(u) = v_0 \cdots v_{2n-1}$ of t of length $2n$. From Lemma 5, $|v|_0$ takes at most three values (depending on n).

Let us compute the possible values taken by the coefficient $\binom{v}{01}$. Consider an occurrence of 01 as a subword of v, i.e., a pair (i, j), $i < j \leqslant n - 1$, such that $v_i v_j = 01$. There are two possible cases:

– If $i = 2m$ and $j = 2m+1$, for some $m \geqslant 0$, then $u_m = 0$ because $v_{2m}v_{2m+1} = \varphi(u_m)$. There are $|u|_0$ such occurrences.
– Otherwise, we have $i \in \{2m, 2m+1\}$, $j \in \{2m', 2m'+1\}$ with $m' > m$. For all m (resp. m'), exactly one letter of the factor $v_{2m}v_{2m+1} = \varphi(u_m)$ (resp. $v_{2m'}v_{2m'+1} = \varphi(u'_m)$) is 0 and the other one is 1. Hence, for any $i \in \{0, \ldots, n-2\}$, j can take a value of the $n-1-i$ values in $\{i+1, \ldots, n-1\}$.

Summarizing these two cases, we have

$$\binom{v}{01} = |u|_0 + \sum_{i=0}^{n-2}(n-1-i) = |u|_0 + \frac{n(n-1)}{2}.$$

From Lemma 5, $|u|_0$ takes at most three values (depending on n) and therefore the same holds for $\binom{v}{01}$. Hence, the conclusion follows.

We now extend the proof of Theorem 5. The first part is to generalize (4).

Lemma 6. *Let $m, k \geqslant 1$. Assume that there exists D such that, for all n,*

$$\#\{\mathbf{B}^{(m)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^k(u)\} \leqslant D.$$

Then the m-binomial complexity of the Thue–Morse word $\mathbf{b}_t^{(m)}$ is bounded by a constant.

Proof. Let $\ell \geqslant 1$. Let f be a factor of t of length ℓ. This factor is of the form[3] pvs where p (resp. s) is a proper suffix (resp. prefix) of some $\varphi^k(a)$ (resp. $\varphi^k(b)$) where a, b are letters and $v = \varphi^k(u)$ for some factor u of t of length n. In particular, we have $|p|, |q| \leqslant 2^k - 1$. Note that ℓ is of the form $n \cdot 2^k + r$ with $0 \leqslant r \leqslant 2(2^k - 1)$. Hence, for a given f of length ℓ, the corresponding integer n can take at most 2 values which are $\lfloor \ell/2^k \rfloor - 1$ and $\lfloor \ell/2^k \rfloor$. From the assumption, we get

$$\#\{\mathbf{B}^{(m)}(v) \mid \exists u \in \mathrm{Fac}_t(\lfloor \ell/2^k \rfloor - 1) \cup \mathrm{Fac}_t(\lfloor \ell/2^k \rfloor) : v = \varphi^k(u)\} \leqslant 2D.$$

Finally, using Remark 2, we have $\mathbf{B}^{(m)}(f) = \mathbf{B}^{(m)}(p) \circ \mathbf{B}^{(m)}(v) \circ \mathbf{B}^{(m)}(s)$. Since p and s have bounded length, $\mathbf{B}^{(m)}(p)$ and $\mathbf{B}^{(m)}(s)$ take a bounded number of values. Moreover, $\mathbf{B}^{(m)}(v)$ takes at most $2D$ values, hence $\mathbf{b}_t^{(m)}$ is bounded by constant.

From now on, intervals $[r, s]$ (resp. $[r, s)$) will be considered as intervals of integers, i.e., one should understand $[r, s] \cap \mathbb{Z}$ (resp. $[r, s) \cap \mathbb{Z}$).

Aside from the idea of dealing with words of a convenient form, the second key idea of the proof of Theorem 5 is to split the set of occurrences of the subword 01 into two disjoint subsets facilitating the counting. We shall now generalize this idea for m-binomial complexity but some terminology is required. Let v be a word. A subset $T = \{t_1 < t_2 < \ldots < t_n\} \subseteq [0, |v|)$ defines a subword denoted by $v_T = v_{t_1}v_{t_2} \cdots v_{t_n}$.

[3] This is the idea of "de-substitution" where t is factorized into consecutive factors of length 2^k.

Definition 3. *If $\alpha_1, \ldots, \alpha_m$ are non-empty and pairwise disjoint subsets of a set X such that $\cup_i \alpha_i = X$, then $\alpha = \{\alpha_1, \ldots, \alpha_m\}$ is a* partition *of X. Any partition α of a set X is a* refinement *of a partition β of X if every element of α is a subset of some element of β. In that case, α is said to be* finer *than β (equivalently β is* coarser *than α) and we write $\alpha \preceq \beta$. Since $\preceq$ is a partial order, we define a* chain *as a subset of partitions $\beta^{(1)}, \beta^{(2)}, \ldots$ of X satisfying*

$$\beta^{(1)} \preceq \beta^{(2)} \preceq \cdots .$$

*A k-*partition *$\alpha = \{\alpha_1, \ldots, \alpha_m\}$ of the set $[0, mk)$ is a partition into subsets $\alpha_i = [(i-1)k, ik)$ of size k. In particular, a 2^i-partition is a refinement of a 2^j-partition of $[0, 2^k)$, $i < j \leqslant k$.*

Definition 4. *Let X be a set and $T = \{t_1 < t_2 < \ldots < t_n\}$ be a subset of X. A partition $\alpha = \{\alpha_1, \ldots, \alpha_m\}$ of X induces a partition $\alpha_T = \{\gamma_1, \ldots, \gamma_r\}$ of $[1, n]$ defined by*

$$i, j \in \gamma_t \Leftrightarrow \exists s : t_i, t_j \in \alpha_s.$$

Note that for two partitions α, β of X, if $\alpha \preceq \beta$, then $\alpha_T \preceq \beta_T$.

Example 2. Take $X = [0, 7]$ and $T = \{0, 2, 3, 5\}$. Consider the following two partitions of X: $\alpha = \{\{0, 1\}, \{2, 3, 4\}, \{5, 6, 7\}\}$ and $\beta = \{\{0, 1, 2\}, \{3, 4, 5\}, \{6, 7\}\}$. We get $\alpha_T = \{\{1\}, \{2, 3\}, \{4\}\}$ and $\beta_T = \{\{1, 2\}, \{3, 4\}\}$.

Definition 5. *Let $T = \{t_1 < t_2 < \ldots < t_n\}$ and $U = \{u_1 < u_2 < \ldots < u_n\}$ be subsets of X. These subsets are* equidistributed *with respect to a partition α of X if $\alpha_T = \alpha_U$. These subsets are* equidistributed *with respect to a chain $\mathfrak{C}$ of partitions of X if $\alpha_T = \alpha_U$ for all $\alpha \in \mathfrak{C}$. We also say that the subsets are* $\mathfrak{C}$-equidistributed.

Example 3. Consider the chain $\mathfrak{C}$ consisting of the 4-partition $\beta = \{[0, 3], [4, 7]\}$ and the 2-partition $\alpha = \{[0, 1], [2, 3], [4, 5], [6, 7]\}$ of the set $[0, 7]$. The subsets $T = \{0, 5\}$, $U = \{1, 2\}$ and $V = \{3, 4\}$ are equidistributed with respect to the 2-partition ($\alpha_T = \alpha_U = \alpha_V = \{\{1\}, \{2\}\}$), but U is not $\mathfrak{C}$-equidistributed to T (resp. V) because $\beta_T = \beta_V = \{\{1\}, \{2\}\}$ and $\beta_U = \{\{1, 2\}\}$.

Example 4. In the last part of the proof of Theorem 5, we have considered the two possible cases for an occurrence of the subword 01 in v. If $T = \{i, j\}$ is a subset of $[0, |v|)$ and α is the 2-partition of $[0, |v|)$, then these cases correspond exactly to the two possible values $\alpha_T = \{1, 2\}$ or $\alpha_T = \{\{1\}, \{2\}\}$.

Let $\mathfrak{C}$ be a chain $\beta^{(1)} \preceq \beta^{(2)} \preceq \cdots$ of partitions of X and $T = \{t_1, \ldots, t_n\}$ be a subset of X. We use nested brackets to represent the induced chain $\beta_T^{(1)} \preceq \beta_T^{(2)} \preceq \cdots$ of partitions of $[1, n]$. The outer (resp. inner) brackets represent the coarsest (resp. finest) partition of $[1, n]$. As an example $[[t_1 t_2]][[t_3][t_4]]$ represents the partition $\{\{1, 2\}, \{3\}, \{4\}\}$ and the coarser partition $\{\{1, 2\}, \{3, 4\}\}$. To get used to these new definitions, we consider another particular statement. (A precise and formal definition of the bracket notation is given in [15].)

Remark 5. Two subsets T and U of size n of X are equidistributed with respect to a chain $\mathfrak{C}$ of partitions of X if and only if they give rise to the same notation of nested brackets. We call it the *type* of T with respect to $\mathfrak{C}$.

Example 5 (continuing Example 3). Consider the subsets $R = \{0, 1, 4, 7\}$ and $S = \{2, 3, 4, 6\}$ of $[0, 7]$. We have $\alpha_R = \alpha_S = \{\{1, 2\}, \{3\}, \{4\}\}$ and $\beta_R = \beta_S = \{\{1, 2\}, \{3, 4\}\}$. Hence R and S are $\mathfrak{C}$-equidistributed and give both rise to the notation $[[t_1t_2]][[t_3][t_4]]$.

We prove the case of the 3-binomial complexity. The proof of the general case has been treated in [15].

Theorem 6. *There exists $C_3 > 0$ such that the 3-binomial complexity of the Thue–Morse word satisfies $\mathbf{b}_t^{(3)}(n) \leqslant C_3$ for all $n \geqslant 0$.*

Proof. In view of Lemma 6, it is enough to show that there exists a constant D such that, for all n, we have $\#\{\mathbf{B}^{(3)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^2(u)\} \leqslant D$.

Let $n \geqslant 1$. Let $v = \varphi^2(u)$ with $u \in \mathrm{Fac}_t(n)$. In particular, $|v| = 4n$. Consider the chain $\mathfrak{C}$ consisting of the 2-partition and the 4-partition of $[0, 4n)$. Any subset $T = \{t_1 < t_2 < t_3\}$ of $[0, 4n)$ is $\mathfrak{C}$-equidistributed to a subset of one the following types:

- $[t_1][t_2][t_3]$, i.e., the union of the types $[[t_1]][[t_2]][[t_3]]$, $[[t_1][t_2]][[t_3]]$ and $[[t_1]][[t_2][t_3]]$: the 3 elements of T belong to pairwise distinct subsets of the 2-partition of $[0, 4n)$
- $[[t_1t_2][t_3]]$ or $[[t_1][t_2t_3]]$: two elements belong to the same subset of the 2-partition of $[0, 4n)$ and the 3 elements of T belong to the same subset of the 4-partition of $[0, 4n)$.
- $[[t_1t_2]][[t_3]]$ or $[[t_1]][[t_2t_3]]$: two elements belong to the same subset of the 2-partition and to the same subset of the 4-partition of $[0, 4n)$.

Let $e = e_0e_1e_2$ be a word of length 3. We will count the number of occurrences of the subword $e = v_{t_1}v_{t_2}v_{t_3}$ in v depending on the type of $T = \{t_1, t_2, t_3\}$ with respect to $\mathfrak{C}$.

Assume that the type of T is $[t_1][t_2][t_3]$. Each subset S of the 2-partition of $[0, 4n)$ corresponds to a factor $v_S = 01$ or $v_S = 10$ and v contains $2n$ such factors. Hence the number of subwords e occurring in v for this type takes, for a given n, a unique value which is $\binom{2n}{3}$.

Now assume that the type of T is $[[t_1t_2][t_3]]$ (similar arguments apply to $[[t_1][t_2t_3]]$). Each subset S of the 4-partition of $[0, 4n)$ corresponds to a factor v_S which is either $\varphi^2(0) = 0110$ or $\varphi^2(1) = 1001$. Then the number of subwords e occurring in v of this type is

$$\underbrace{\binom{01}{e_0e_1}}_{0 \text{ or } 1}\underbrace{\binom{10}{e_2}}_{1}|u|_0 + \underbrace{\binom{10}{e_0e_1}}_{0 \text{ or } 1}\underbrace{\binom{01}{e_2}}_{1}|u|_1 \in \{0, |u|_0, |u|_1\}.$$

Recall that, for a given $n = |u|$, the pair $(|u|_0, |u|_1)$ can take at most three values (see Lemma 5). The number of subwords e occurring in v of this type takes, for a given n, takes at most 4 values[4].

Now assume that the type of T is $[[t_1t_2]][[t_3]]$ (similar arguments apply to $[[t_1]][[t_2t_3]]$). Each subset S of the 4-partition of $[0, 4n)$ is a union of two sets S', S'' of the 2-partition of $[0, 4n)$ and we have either $v_{S'} = 01, v_{S''} = 10$ or $v_{S'} = 10, v_{S''} = 01$. They are n subsets of size 4 in the 4-partition of $[0, 4n)$ and we have to pick 2 of them. Hence, the number of subwords e occurring in v for this type is

$$\underbrace{(\binom{01}{e_0e_1} + \binom{10}{e_0e_1})}_{0\text{ or }1}\underbrace{(\binom{01}{e_2} + \binom{10}{e_2})}_{2}\binom{n}{2}$$

and this quantity, for a given n, takes at most 2 values.

We have proved that, for all $|e| = 3$ and $v = \varphi^2(u)$ with $u \in \mathrm{Fac}_t(n)$, $\binom{v}{e}$ takes at most $1 + 2 \cdot 4 + 2 \cdot 2 = 13$ values (these values depend on n, but the *number* of values is bounded without any dependence to n). Note that $\mathbf{B}^{(3)}(v)$ is determined from $\mathbf{B}^{(2)}(v)$ and by the values of $\binom{v}{e}$ for the words e of length 3. To conclude the proof, note that $\#\{\mathbf{B}^{(2)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^2(u)\}$ is bounded by $\#\{\mathbf{B}^{(2)}(v) \mid \exists z \in \mathrm{Fac}_t(2n) : v = \varphi(z)\} \leqslant 9$ using (4). Consequently, we have shown that $\#\{\mathbf{B}^{(3)}(v) \mid \exists u \in \mathrm{Fac}_t(n) : v = \varphi^2(u)\} \leqslant 9 \cdot 13^8$ for all $n \geqslant 1$.

Remark 6. By computer experiments, $\mathbf{b}_t^{(2)}(n)$ is equal to 9 if $n \equiv 0 \pmod 4$ and to 8 otherwise, for $10 \leqslant n \leqslant 1000$. Moreover, $\mathbf{b}_t^{(3)}(n)$ is equal to 21 if $n \equiv 0 \pmod 8$ and to 20 otherwise, for $8 \leqslant n \leqslant 500$.

6 A Glimpse at Avoidance

It is obvious that, over a 2-letter alphabet, any word of length $\geqslant 4$ contains a square. On the other hand, there exist square-free infinite ternary words [12]. In the same way, over a 3-letter alphabet, any word of length $\geqslant 8$ contains an abelian square, i.e., a word uu' where $u \sim_1 u'$. But, over a 4-letter alphabet, abelian squares are avoidable, see for instance [10]. So a first natural question in that direction is to determine, whether or not, over a 3-letter alphabet 2-binomial squares can be avoided in arbitrarily long words. Naturally, a 2-*binomial square* is a word of the form uu' where $u \sim_2 u'$. Note that, for abelian equivalence, the longest ternary word which is 2-abelian square-free has length 537 [9].

As an example, $u = 121321231213123132123121312$ is a word of length 27 without 2-binomial squares but this word cannot be extended without getting a 2-binomial square. Indeed, $u1$ (resp. $u3$) ends with a square of length 8 (resp. 26)

Consider the 13-uniform morphism of Leech [11] which is well-known to be square-free, $g : a \mapsto abcbacbcabcba, b \mapsto bcacbacabcacb, c \mapsto cabacbabcabac$. In

[4] A close inspection shows that if $|u| = 2n$, then $|u|_0, |u|_1 \in \{n-1, n, n+1\}$, if $|u| = 2n+1$, then $|u|_0, |u|_1 \in \{n, n+1\}$.

the submitted version of this paper, we conjectured that the infinite square-free word $g^\omega(1)$ avoids 2-binomial squares. For instance, we can prove that

$$u \sim_2 v \Leftrightarrow g(u) \sim_2 g(v).$$

Nevertheless, M. Bennett has recently shown that the factor of length 508 occurring in position 845 is a 2-binomial square [4].

Acknowledgments

The idea of this binomial equivalence came after the meeting "Representing streams" organized at the Lorentz center in December 2012 where Jean-Eric Pin presented a talk, *Noncommutative extension of Mahlers theorem on interpolation series*, involving binomial coefficients on words. Jean-Eric Pin and the first author proposed independently to introduce this new relation.

References

1. S.V. Avgustinovich, D.G. Fon-Der-Flaass, A.E. Frid, Arithmetical complexity of infinite words, in *Words, Languages & Combinatorics III*, M. Ito and T. Imaoka (Eds.), World Scientific Publishing (2003) 51–62.
2. S. Brlek, Enumeration of factors in the Thue-Morse word, *Discrete Appl. Math.* **24** (1989), 83–96.
3. J. Berstel, M. Crochemore, J.-E. Pin, Thue–Morse sequence and p-adic topology for the free monoid, *Disc. Math.* **76** (1989), 89–94.
4. J. Currie, personal communication, 3th June 2013.
5. A. de Luca, S. Varricchio, On the factors of the Thue-Morse word on three symbols, *Inform. Process. Lett.* **27** (1988), 281–285.
6. P. Erdős, Some unsolved problems, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961), 221-254.
7. T. Kamae, L. Zamboni, Sequence entropy and the maximal pattern complexity of infinite words, *Ergodic Theory Dynam. Systems* **22** (2002), 1191–1199.
8. J. Karhumäki, A. Saarela, L. Q. Zamboni, On a generalization of Abelian equivalence and complexity of infinite words, `arXiv:1301.5104`.
9. M. Huova, J. Karhumäki, Observations and problems on k-abelian avoidability, In Combinatorial and Algorithmic Aspects of Sequence Processing (Dagstuhl Seminar 11081), (2011) 2215-2219.
10. V. Keränen, Abelian squares are avoidable on 4 letters, *Lecture Notes in Comput. Sci.* **623** (1992), 41-52.
11. J. Leech, A problem on strings of beads, *Math. Gazette* **41**, 277–278 (1957).
12. M. Lothaire, *Combinatorics on Words*, Cambridge Mathematical Library, Cambridge University Press, (1997).
13. M. Lothaire, *Algebraic Combinatorics on Words*, Encyclopedia of Mathematics and its Applications **90**, Cambridge University Press (2002).
14. N. Pytheas Fogg, *Substitutions in dynamics, arithmetics and combinatorics*, V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel (Eds.), Lecture Notes in Mathematics **1794**, Springer-Verlag, Berlin, 2002.
15. M. Rigo, P. Salimov, Another Generalization of Abelian Equivalence: Binomial Complexity of Infinite Words (long version), preprint (2013), available at `http://hdl.handle.net/2268/149313`