Computing the k-binomial complexity of the Thue–Morse word

Marie Lejeune[1,2][0000−0001−5620−8052], Julien Leroy[1], and Michel Rigo[1][0000−0001−7463−8507]

[1] Department of Mathematics, University of Liège, Allée de la Découverte 12 (B37), B-4000 Liège, Belgium. {M.Lejeune,J.Leroy,M.Rigo}@uliege.be
[2] The first author is supported by a FNRS fellowship.

Abstract. Two finite words are k-binomially equivalent whenever they share the same subwords, i.e., subsequences, of length at most k with the same multiplicities. This is a refinement of both abelian equivalence and the Simon congruence. The k-binomial complexity of an infinite word $\mathbf{x}$ maps the integer n to the number of classes in the quotient, by this k-binomial equivalence relation, of the set of factors of length n occurring in $\mathbf{x}$. This complexity measure has not been investigated very much. In this paper, we characterize the k-binomial complexity of the Thue–Morse word. The result is striking, compared to more familiar complexity functions. Although the Thue–Morse word is aperiodic, its k-binomial complexity eventually takes only two values. In this paper, we first express the number of occurrences of subwords appearing in iterates of the form $\Psi^\ell(w)$ for an arbitrary morphism Ψ. We also thoroughly describe the factors of the Thue–Morse word by introducing a relevant new equivalence relation.

1 Introduction

The Thue–Morse word $\mathbf{t} = 011010011001\cdots$ is ubiquitous in combinatorics on words [1,20,27]. It is an archetypal example of a 2-automatic sequence: it is the fixed point of the morphism $0 \mapsto 01$, $1 \mapsto 10$. See, for instance, [2]. Its most prominent property is that it avoids overlaps, i.e., it does not contain any factor of the form $auaua$ where u is a word and a a symbol. Consequently it also avoids cubes, i.e., words of the form uuu, and is aperiodic.

Various measures of complexity of infinite words have been considered in the literature. The most usual one is the *factor complexity* that one can, for instance, relate to the topological entropy of a symbolic dynamical system. The factor complexity of an infinite word $\mathbf{x}$ simply counts the number $p_{\mathbf{x}}(n) = \#\operatorname{Fac}_n(\mathbf{x})$ of factors of length n occurring in $\mathbf{x}$. One can also consider other measures such as abelian complexity or k-abelian complexity [10]. For instance, in the sixties, Erdős raised the question whether abelian squares can be avoided by an infinite word over an alphabet of size 4. In an attempt to generalize Parikh's theorem on context-free languages, k-abelian complexity counts the number

of equivalence classes partitioning the set of factors of length n for the so-called k-abelian equivalence. Two finite words u and v are *k-abelian equivalent* if $|u|_x = |v|_x$, for all words x of length at most k, and where $|u|_x$ denotes the number of occurrences of x as a factor of u.

The celebrated theorem of Morse–Hedlund characterizes ultimately periodic words in terms of a bounded factor complexity function; for a reference, see [2,16] or [4, Section 4.3]. Hence, aperiodic words with the lowest factor complexity are exactly the Sturmian words characterized by $p_{\mathbf{x}}(n) = n + 1$. It is also a well-known result of Cobham that a k-automatic sequence has factor complexity in $\mathcal{O}(n)$. The factor complexity of the Thue–Morse word is in $\Theta(n)$ and is recalled in Proposition 6.

For many complexity measures, Sturmian words have the lowest complexity among aperiodic words, and variations of the Morse–Hedlund theorem notably exist for k-abelian complexity [11].

Binomial coefficients of words have been extensively studied [15]: $\binom{u}{x}$ denotes the number of occurrences of x as a subword, i.e., a subsequence, of u. They have been successfully used in several applications: p-adic topology [3], non-commutative extension of Mahler's theorem on interpolation series [19], formal language theory [9], Parikh matrices, and a generalization of Sierpiński's triangle [14].

Binomial complexity of infinite words has been recently investigated [21,23]. The definition is parallel to that of k-abelian complexity. Two finite words u and v are *k-binomially equivalent* if $\binom{u}{x} = \binom{v}{x}$, for all words x of length at most k. This relation is a refinement of abelian equivalence and Simon's congruence. We thus take the quotient of the set of factors of length n by this new equivalence relation. For all $k \geq 2$, Sturmian words have k-binomial complexity that is the same as their factor complexity. However, the Thue–Morse word has bounded k-binomial complexity [23]. So we have a striking difference with the usual complexity measures. This phenomenon therefore has to be closely investigated. In this paper, we compute the exact value of the k-binomial complexity $b_{\mathbf{t},k}(n)$ of the Thue–Morse word $\mathbf{t}$. To achieve this goal, we first obtain general results computing the number of occurrences of a subword in the (iterated) image by a morphism. This discussion is not restricted to the Thue–Morse morphism.

This paper is organized as follows. In Section 2, we recall basic results about binomial coefficients, binomial equivalence and the Thue–Morse word. In Section 3, we give an expression to compute the coefficient $\binom{\Psi(w)}{u}$ for an arbitrary morphism Ψ in terms of binomial coefficients for the preimage w. To that end, we study factorizations of u of the form $u = x\Psi(u')y$.

In the second part of this paper, we specifically study the k-binomial complexity of the Thue–Morse word. For $k = 1$, the abelian complexity of $\mathbf{t}$ is well known and takes only the values 2 and 3. The case $k = 2$ is treated in Section 4. In the last three sections, we consider the general case $k \geq 3$. The precise statement of our main result is given in Theorem 5. The principal tool to get our result is a new equivalence relation discussed in Section 6. This relation is based on particular factorizations of factors occurring in the Thue–Morse word.

Due to space limitations for this 12-page version, we have omitted most of the technical difficulties but tried to convey the main ideas and concepts. The reader can find a comprehensive presentation in [13].

2 Basics

Let $A = \{0, 1\}$. Let $\varphi : A^* \to A^*$ be the classical Thue–Morse morphism defined by $\varphi(0) = 01$ and $\varphi(1) = 10$. The *complement* of a word $u \in A^*$ is the image of u under the involutive morphism mapping 0 to 1 and 1 to 0. It is denoted by $\overline{u}$. The length of the word u is denoted by $|u|$.

2.1 Binomial coefficients and binomial equivalence

The binomial coefficient $\binom{u}{v}$ of two finite words u and v is the number of times v occurs as a subsequence of u (meaning as a "scattered" subword). As an example, we consider two particular words over $\{0, 1\}$ and

$$\binom{101001}{101} = 6 \ .$$

For more on these binomial coefficients, see, for instance, [15, Chap. 6]. In particular, $\binom{u}{\varepsilon} = 1$. In this paper, a *factor* of a word is made of consecutive letters. However this is not necessarily the case for a *subword* of a word.

Definition 1 (Binomial equivalence). *Let $k \in \mathbb{N}$ and u, v be two words over A. We let $A^{\leq k}$ denote the set of words of length at most k over A. We say that u and v are* k-binomially equivalent *if*

$$\binom{u}{x} = \binom{v}{x}, \ \forall x \in A^{\leq k} \ .$$

We simply write $u \sim_k v$ if u and v are k-binomially equivalent. The word u is obtained as a permutation of the letters in v if and only if $u \sim_1 v$. In that case, we say that u and v are abelian equivalent*. Note that, for all $k \geq 1$, if $u \sim_{k+1} v$, then $u \sim_k v$.*

Example 2. The four words 0101110, 0110101, 1001101 and 1010011 are 2-binomially equivalent. Let u be any of these four words. We have

$$\binom{u}{0} = 3, \ \binom{u}{1} = 4, \ \binom{u}{00} = 3, \ \binom{u}{01} = 7, \ \binom{u}{10} = 5, \ \binom{u}{11} = 6 \ .$$

For instance, the word 0001111 is abelian equivalent to 0101110 but these two words are not 2-binomially equivalent. To see this, simply compute the number of occurrences of the subword 10 in each.

Many classical questions in combinatorics on words can be considered in this binomial context [22,24]. Avoiding binomial squares and cubes is considered in [21]. The problem of testing whether two words are k-binomially equivalent or not is discussed in [7]. In particular, one can introduce the k-binomial complexity function.

Definition 3 (Binomial complexity). *Let* $\mathbf{x}$ *be an infinite word. The* k*-binomial complexity function of* $\mathbf{x}$ *is defined as*

$$b_{\mathbf{x},k} : \mathbb{N} \to \mathbb{N},\ n \mapsto \#\left(\mathrm{Fac}_n(\mathbf{x})/\sim_k\right)$$

where $\mathrm{Fac}_n(\mathbf{x})$ *is the set of factors of length* n *occurring in* $\mathbf{x}$.

2.2 Context of this Paper

The Thue–Morse word denoted by $\mathbf{t}$ is the fixed point starting with 0 of the morphism φ. In [23, Thm. 13], it is shown that $\mathbf{t}$ has a bounded k-binomial complexity. Actually, this behavior occurs for all morphisms where images of letters are permutations of the same word.

Theorem 4. *[23] Let* $k \geq 1$. *There exists* $C_k > 0$ *such that the* k*-binomial complexity of the Thue–Morse word satisfies* $b_{\mathbf{t},k}(n) \leq C_k$ *for all* $n \geq 0$.

Our contribution is the exact characterization of $b_{\mathbf{t},k}(n)$.

Theorem 5. *Let* k *be a positive integer. For all* $n \leq 2^k - 1$, *we have*

$$b_{\mathbf{t},k}(n) = p_{\mathbf{t}}(n).$$

For all $n \geq 2^k$, *we have*

$$b_{\mathbf{t},k}(n) = \begin{cases} 3 \cdot 2^k - 3, & \text{if } n \equiv 0 \pmod{2^k}; \\ 3 \cdot 2^k - 4, & \text{otherwise.} \end{cases}$$

Observe that $3 \cdot 2^k - 4$ is exactly the number of words of length $2^k - 1$ in $\mathbf{t}$, for $k \neq 2$. Indeed, the factor complexity of $\mathbf{t}$ is well known [4, Corollary 4.10.7].

Proposition 6. *[4,5,6] The factor complexity* $p_{\mathbf{t}}$ *of the Thue–Morse word is given by* $p_{\mathbf{t}}(0) = 1$, $p_{\mathbf{t}}(1) = 2$, $p_{\mathbf{t}}(2) = 4$ *and for* $n \geq 3$,

$$p_{\mathbf{t}}(n) = \begin{cases} 4n - 2 \cdot 2^m - 4, & \text{if } 2 \cdot 2^m < n \leq 3 \cdot 2^m; \\ 2n + 4 \cdot 2^m - 2, & \text{if } 3 \cdot 2^m < n \leq 4 \cdot 2^m. \end{cases}$$

There are 2 factors of length $1 = 2^1 - 1$ and 6 factors of length $3 = 2^2 - 1$. The number of factors of $\mathbf{t}$ of length $2^k - 1$ for $k \geq 3$ is given by $2(2^k - 1) + 4 \cdot 2^{k-2} - 2 = 3 \cdot 2^k - 4$,

$$(p_{\mathbf{t}}(2^k - 1))_{k \geq 0} = 1, 2, 6, 20, 44, 92, 188, 380, 764, 1532, \ldots$$

which is exactly one of two values stated in our main result, Theorem 5.

3 Occurrences of Subwords in Images by φ

The aim of this section is to obtain an expression for coefficients of the form $\binom{\varphi(w)}{u}$. Even though we are mainly interested in the Thue–Morse word, our observations can be applied to any non-erasing morphism as summarized by Theorem 15.

A *multiset* is just a set where elements can be repeated with a (finite) integer multiplicity. If x belongs to a multiset M, its multiplicity is denoted by $m_M(x)$ or simply $m(x)$. If $x \notin M$, then $m_M(x) = 0$. If we enumerate the elements of a multiset, we adopt the convention to write multiplicities with indices. The *multiset sum* $M \uplus N$ of two multisets M, N is the union of the two multisets and the multiplicity of an element is equal to the sum of the respective multiplicities.

Let us start with an introductory example. We hope that this example will forge the intuition of the reader about the general scheme.

Example 7. We want to compute

$$\binom{\varphi(01100)}{u} \quad \text{with } u = 011.$$

The word $w = \varphi(01100)$ belongs to $\{01, 10\}^*$. It can be factorized with consecutive blocks $b_1 b_2 \cdots b_5$ of length 2. To count the number of occurrences of the subword u in the image by φ of a word, two cases need to be taken into account:

- the three symbols of u appear in pairwise distinct 2-blocks of w (each 2-block contains both 0 and 1 exactly once), and there are
$$\binom{|w|/2}{|u|} = \binom{5}{3}$$
such choices, or;
- the prefix 01 of u is one of the 2-blocks b_i of w and the last symbol of u appears in subsequent distinct 2-block b_j, $j > i$. Since $\varphi(0) = 01$, we have to count the number of occurrences of the subword $0z$, for all words z of length 1, in the preimage of w. There are
$$\sum_{z \in A} \binom{01100}{0z} = 4 + 1 = 5$$
such choices.

The general scheme behind this computation is expressed by Theorem 12 given below. The reader can already feel that we need to take into account particular factorizations of u with respect to occurrences of a factor $\varphi(0)$ or $\varphi(1)$. The two cases discussed in Example 7 correspond to the following factorizations of u:

$$011, \ \varphi(0)1.$$

We thus introduce the notion of a φ-factorization.

Definition 8 (φ-factorization). *If a word $u \in A^*$ contains a factor* 01 *or* 10*, then it can be factorized as*

$$u = w_0\, \varphi(a_1)\, w_1 \cdots w_{k-1}\, \varphi(a_k)\, w_k \tag{1}$$

for some $k \geq 1$, $a_1, \ldots, a_k \in A$ and $w_0, \ldots, w_k \in A^*$ (some of these words are possibly empty). We call this factorization, a φ-factorization *of u. It is coded by the k-tuple of positions where the $\varphi(a_i)$'s occurs:*

$$\kappa = (|w_0|, |w_0\varphi(a_1)w_1|, |w_0\varphi(a_1)w_1\varphi(a_2)w_2|, \ldots, |w_0\varphi(a_1)w_1\varphi(a_2)w_2 \cdots w_{k-1}|)\,.$$

The set of all the φ-factorizations of u is denoted by φ-Fac(u)*.*

Since $|\varphi(a)| = 2$, for all $a \in A$, observe that if $(i_1, \ldots, i_k)$ codes a φ-factorization, then $i_{j+1} - i_j \geq 2$ for all j. Note that u starts with a prefix 01 or 10 if and only if there are φ-factorizations of u coded by tuples starting with 0.

We define a map f from A^* to the set of finite multisets of words over A^*. This map is defined as follows.

Definition 9. *If $u \in 0^* \cup 1^*$, then $f(u) = \emptyset$ (the meaning for this choice will be clear with Theorem 12). If u is not of this form, it contains a factor* 01 *or* 10*. With every φ-factorization $\kappa \in \varphi$-Fac(u) of u of the form* (1)

$$u = w_0\, \varphi(a_1)\, w_1 \cdots w_{k-1}\, \varphi(a_k)\, w_k$$

for some $k \geq 1$, $a_1, \ldots, a_k \in A$ and $w_0, \ldots, w_k \in A^*$, we define the language

$$\mathcal{L}(u, \kappa) := A^{|w_0|}\, a_1\, A^{|w_1|} \cdots A^{|w_{k-1}|} a_k A^{|w_k|}$$

of words of length $|u| - k$ (there are $2^{|u|-2k}$ of these words[3]). Such a language is considered as a multiset whose elements have multiplicities equal to 1*. Now, $f(u)$ is defined as the multiset sum (i.e., we sum the multiplicities) of the above languages for all φ-factorizations of u, i.e.,*

$$f(u) := \biguplus_{\kappa \in \varphi\text{-Fac}(u)} \mathcal{L}(u, \kappa)\,.$$

Definition 10. *Now that f is defined over A^*, we can extend it to any finite multiset M of words over A. It is the multiset sum of the $f(v)$'s, for all $v \in M$, repeated with their multiplicities.*

Remark 11. If u does not belong to $0^* \cup 1^*$, then $f^{|u|-2}(u)$ contains only elements in $\{0, 1, 00, 01, 10, 11\}$ and $f^{|u|-1}(u)$ contains only elements in $\{0, 1\}$. For $n \geq |u|$, $f^n(u)$ is empty.

[3] We have all the words of length $|u| - k$ where in k positions the occurring symbol is given.

Recall that $f(u)$ is a multiset. Hence $m_{f(u)}(v)$ denotes the multiplicity of v as element of $f(u)$.

Theorem 12. *With the above notation, for all words u, w, we have*

$$\binom{\varphi(w)}{u} = \binom{|w|}{|u|} + \sum_{\substack{\kappa \in \varphi\text{-Fac}(u) \\ v \in \mathcal{L}(u,\kappa)}} \binom{w}{v} = \binom{|w|}{|u|} + \sum_{v \in f(u)} m_{f(u)}(v) \binom{w}{v} .$$

We can then establish the following result.

Corollary 13. *Let $k \geq 1$. For all words u, v, we have*

$$u \sim_k v \Rightarrow \varphi(u) \sim_{k+1} \varphi(v) .$$

In particular, $\varphi^k(0) \sim_k \varphi^k(1)$ for all $k \geq 1$.

Theorem 12 can be extended to iterates of φ.

Corollary 14. *With the above notation, for $\ell \geq 1$ and all words u, w, we have*

$$\binom{\varphi^\ell(w)}{u} = \sum_{i=0}^{\ell-1} \sum_{v \in f^i(u)} m_{f^i(u)}(v) \binom{|\varphi^{\ell-i-1}(w)|}{|v|} + \sum_{x \in f^\ell(u)} m_{f^\ell(u)}(x) \binom{w}{x} .$$

The reader should be convinced that the following general statement holds.

Theorem 15. *Let $\Psi : A^* \to B^*$ be a non-erasing morphism and $u \in B^+$, $w \in A^+$ be two words. We have*

$$\binom{\Psi(w)}{u} = \sum_{k=1}^{|u|} \sum_{\substack{u_1,\ldots,u_k \in B^+ \\ u = u_1 \cdots u_k}} \sum_{a_1,\ldots,a_k \in A} \binom{\Psi(a_1)}{u_1} \cdots \binom{\Psi(a_k)}{u_k} \binom{w}{a_1 \cdots a_k}.$$

The word u occurs as a subword of $\Psi(w)$ if and only if there exists $k \geq 1$ such that u can be factorized into $u_1 \cdots u_k$ where, for all i, u_i is a non-empty subword occurring in $\Psi(a_i)$ for some letter a_i and such that $a_1 \cdots a_k$ is a subword of w.

4 Computing $b_{\mathbf{t},2}(n)$

In this section we compute the value of $b_{\mathbf{t},2}(n)$. First of all, the next proposition ensures us that all the words we will consider in the proof of Theorem 17 really appear as factors of $\mathbf{t}$.

Proposition 16 (folklore). *Let $k, m \in \mathbb{N}$ and $a, b \in \{0, 1\}$. Let p_u be a suffix of $\varphi^k(a)$ and s_u be a prefix of $\varphi^k(b)$. There exists $z \in \{0, 1\}^m$ such that $p_u \varphi^k(z) s_u$ is a factor of $\mathbf{t}$.*

Using this result, we can compute the values of $b_{\mathbf{t},2}$.

Theorem 17. *[12, Thm. 3.3.6] We have* $b_{\mathbf{t},2}(0) = 1$, $b_{\mathbf{t},2}(1) = 2$, $b_{\mathbf{t},2}(2) = 4$, $b_{\mathbf{t},2}(3) = 6$ *and for all* $n \geq 4$,

$$b_{\mathbf{t},2}(n) = \begin{cases} 9, \text{ if } n \equiv 0 \pmod 4; \\ 8, \text{ otherwise.} \end{cases}$$

Proof. Assume $n \geq 4$.

We have to consider four cases depending on the value of $\lambda \in \{0,1,2,3\}$ such that $\lambda = n \bmod 4$. For every one of them, we want to compute

$$b_{\mathbf{t},2}(n) = \#\left\{ \left(\binom{u}{0}, \binom{u}{01} \right) \in \mathbb{N} \times \mathbb{N} : u \in \mathrm{Fac}_n(\mathbf{t}) \right\}.$$

Since $\mathbf{t}$ is the fixed point of the morphism φ, we know that every factor u of length n of $\mathbf{t}$ can be written $p_u\varphi^2(z)s_u$ for some $z \in A^*$ and p_u (resp., s_u) suffix (resp., prefix) of a word in $\{\varphi^2(0), \varphi^2(1)\}$. From the previous proposition, we also know that every word of that form occurs at least once in $\mathbf{t}$. Moreover, we have $|p_u| + |s_u| \in \{\lambda, \lambda + 4\}$ and, as a consequence, $|z| = \left\lfloor \frac{n}{4} \right\rfloor = \frac{n-\lambda}{4}$ or $|z| = \left\lfloor \frac{n}{4} \right\rfloor - 1$. Set $\ell = \frac{n-\lambda}{4}$.

Let us first consider the case $\lambda = 0$. We have

$$\begin{aligned} \mathrm{Fac}_n(\mathbf{t}) = \{&\varphi^2(az), 0\varphi^2(z)011, 0\varphi^2(z)100, 1\varphi^2(z)011, 1\varphi^2(z)100, \\ &01\varphi^2(z)01, 01\varphi^2(z)10, 10\varphi^2(z)01, 10\varphi^2(z)10, \\ &110\varphi^2(z)0, 110\varphi^2(z)1, 001\varphi^2(z)0, 001\varphi^2(z)1 : z \in A^{\ell-1}, a \in A, az \in \mathrm{Fac}(\mathbf{t})\}. \end{aligned}$$

Let us illustrate the computation of $\left(\binom{u}{0}, \binom{u}{01} \right)$ on $u = 0\varphi^2(z)011 \in \mathrm{Fac}_n(\mathbf{t})$. Firstly,

$$\binom{u}{0} = \binom{0}{0} + \binom{\varphi^2(z)}{0} + \binom{011}{0} = 2 + 2|z| = 2\ell$$

since $|z| = \ell - 1$. Similarly, we have

$$\begin{aligned} \binom{u}{01} &= \binom{0}{01} + \binom{\varphi^2(z)}{01} + \binom{011}{01} + \binom{0}{0}\binom{\varphi^2(z)}{1} + \binom{0}{0}\binom{011}{1} + \binom{\varphi^2(z)}{0}\binom{011}{1} \\ &= \binom{|\varphi(z)|}{2} + \binom{\varphi(z)}{0} + 2 + |\varphi(z)| + 2 + 2|\varphi(z)| \\ &= |z|(2|z| - 1) + |z| + 6|z| + 4 = 2\ell^2 + 2\ell. \end{aligned}$$

All the computations are summarized in the table below. We give the form of the factors and respective values for the pairs $\left(\binom{u}{0}, \binom{u}{01} \right)$.

Case	$\varphi^2(az)$ $01\varphi^2(z)10$ $10\varphi^2(z)01$	$0\varphi^2(z)011$ $001\varphi^2(z)1$	$1\varphi^2(z)100$ $110\varphi^2(z)0$	$0\varphi^2(z)100$	$001\varphi^2(z)0$
$\binom{u}{0}$	2ℓ	2ℓ	2ℓ	$2\ell + 1$	$2\ell + 1$
$\binom{u}{01}$	$2\ell^2$	$2\ell^2 + 2\ell$	$2\ell^2 - 2\ell$	$2\ell^2 - 1$	$2\ell^2$

Case	$1\varphi^2(z)011$	$110\varphi^2(z)1$	$01\varphi^2(z)01$	$10\varphi^2(z)10$	
$\binom{u}{0}$	$2\ell - 1$	$2\ell - 1$	2ℓ	2ℓ	
$\binom{u}{01}$	$2\ell^2$	$2\ell^2 + 1$	$2\ell^2 + 1$	$2\ell^2 - 1$	

This is thus clear that if $n \equiv 0 \pmod 4$, we have $b_{\mathbf{t},2}(n) = 9$.

The same type of computations can be carried out in cases where $\lambda \neq 0$, and give 8 equivalence classes. The obtained values can be found in [13].

5 How to Cut Factors of the Thue–Morse Word

Computing $b_{\mathbf{t},k}(n)$, for all $k \geq 3$, will require much more knowledge about the factors of $\mathbf{t}$. This section is concerned about particular factorizations of factors occurring in $\mathbf{t}$. Similar ideas first appeared in [25,26].

Since $\mathbf{t}$ is a fixed point of φ, it is very often convenient to view $\mathbf{t}$ as a concatenation of blocks belonging to $\{\varphi^k(0), \varphi^k(1)\}$. Hence, we first define a function bar_k that roughly plays the role of a ruler marking the positions where a new block of length 2^k occurs (these positions are called *cutting bars of order k*). For all $k \geq 1$, let us consider the function $\mathrm{bar}_k : \mathbb{N} \to \mathbb{N}$ defined by

$$\mathrm{bar}_k(n) = |\varphi^k(\mathbf{t}_{[0,n)})| = n \cdot 2^k,$$

where $\mathbf{t}_{[0,n)}$ is the prefix of length n of $\mathbf{t}$.

Given a factor u of $\mathbf{t}$, we are interested in the relative positions of $\mathrm{bar}_k(\mathbb{N})$ in u: we look at all the occurrences of u in $\mathbf{t}$ and see what configurations can be achieved, that is how an interval I such that $\mathbf{t}_I = u$ can intersect $\mathrm{bar}_k(\mathbb{N})$.

Definition 18 (Cutting set). *For all $k \geq 1$, we define the set $\mathrm{Cut}_k(u)$ of non-empty sets of relative positions of cutting bars*

$$\mathrm{Cut}_k(u) := \left\{ \big([i, i+|u|] \cap \mathrm{bar}_k(\mathbb{N})\big) - i \mid i \in \mathbb{N}, u = \mathbf{t}_{[i,i+|u|)} \right\}.$$

A cutting set of order k is an element of $\mathrm{Cut}_k(u)$. Observe that we consider the closed interval $[i, i+|u|]$ because we are also interested in knowing if the end of u coincide with a cutting bar.

Example 19. The word $u = 01001$ is the factor $\mathbf{t}_{[3,8)}$ so the set $\{1, 3, 5\}$ which is equal to $([3, 8] \cap 2\mathbb{N}) - 3$ is a cutting set of order 1 of u. Observing that the factor 00 can only occur as a factor of $\varphi(10)$, one easily deduces that it is the unique cutting set of order 1 of u. On the opposite, we have $010 = \mathbf{t}_{[3,6)} = \mathbf{t}_{[10,13)}$, so that $\mathrm{Cut}_1(010)$ contains both $\{1, 3\}$ and $\{0, 2\}$.

Remark 20. Let u be a factor of $\mathbf{t}$. Observe that, for all $\ell \geq 1$, $\mathrm{Cut}_\ell(u) \neq \emptyset$. It results from the following three observations.

Obviously, $\mathrm{bar}_k(\mathbb{N}) \subset \mathrm{bar}_{k-1}(\mathbb{N})$ and thus if $\mathrm{Cut}_k(u)$ is non-empty, then the same holds for $\mathrm{Cut}_{k-1}(u)$. Next notice that if $\mathrm{Cut}_k(u)$ contains a singleton, then $\mathrm{Cut}_{k+1}(u)$ contains a singleton. Finally, there exists a unique k such that $2^{k-1} \leq |u| \leq 2^k - 1$. There also exists i such that $u = \mathbf{t}_{[i,i+|u|)}$. Simply notice that either $[i, i+|u|] \cap \mathrm{bar}_k(\mathbb{N})$ is a singleton or, $[i, i+|u|] \cap \mathrm{bar}_{k-1}(\mathbb{N})$ is a singleton.

Observe that for any word u and any set $C \in \mathrm{Cut}_k(u)$, there is a unique integer $r \in \{0, 1, \ldots, 2^k - 1\}$ such that $C \subset 2^k\mathbb{N} + r$.

Lemma 21. *Let k be a positive integer and u be a factor of $\mathbf{t}$. Let C be a set $\{i_1 < i_2 < \cdots < i_n\}$ in $\mathrm{Cut}_k(u)$. There is a unique factor v of $\mathbf{t}$ of length $n-1$ such that $u = p\varphi^k(v)s$, with $|p| = i_1$. Furthermore, if $i_1 > 0$ (resp., $i_n < |u|$), there is a unique letter a such that p (resp., s) is a proper suffix (resp., prefix) of $\varphi^k(a)$.*

Definition 22 (Factorization of order k). *Let u be a factor of $\mathbf{t}$ and C a cutting set in $\mathrm{Cut}_k(u)$. By Lemma 21, we can associate with C a unique pair $(p, s) \in A^* \times A^*$ and a unique triple $(a, v, b) \in (A \cup \{\varepsilon\}) \times A^* \times (A \cup \{\varepsilon\})$ such that $u = p\varphi^k(v)s$, where either $a = p = \varepsilon$ (resp., $b = s = \varepsilon$), or $a \neq \varepsilon$ and p is a proper suffix of $\varphi^k(a)$ (resp., $b \neq \varepsilon$ and s is a proper prefix of $\varphi^k(b)$). In particular, we have $a = p = \varepsilon$ exactly when $\min(C) = 0$ and $b = s = \varepsilon$ exactly when $\max(C) = |u|$. The triple (a, v, b) is called the* desubstitution of u associated with C *and the pair (p, s) is called the* factorization of u associated with C. *If $C \in \mathrm{Cut}_k(u)$, then (a, v, b) and (p, s) are respectively desubstitutions and factorizations of* order k.

Pursuing the reasoning of Example 19, one could easily show that for any factor u of $\mathbf{t}$ of length at least 4, $\mathrm{Cut}_1(u)$ contains a single set. Furthermore, the substitution φ being primitive and $\mathbf{t}$ being aperiodic, Mossé's recognizability theorem ensures that the substitution φ^k is *bilaterally recognizable* [17,18] for all $k \geq 1$, i.e., any sufficiently long factor u of $\mathbf{t}$ can be uniquely desubstituted by φ^k (up to a prefix and a suffix of bounded length). In the case of the Thue–Morse substitution, we can make this result more precise. Similar results are considered in [8] where the term (maximal extensible) reading frames is used.

Lemma 23. *Let $k \geq 3$ be an integer and u be a factor of $\mathbf{t}$ of length at least $2^k - 1$. Then $\mathrm{Cut}_k(u)$ is a not a singleton if and only if u is a factor of $\varphi^{k-1}(010)$ or of $\varphi^{k-1}(101)$, in which case we have $\mathrm{Cut}_k(u) = \{C_1, C_2\}$ and $|\min C_1 - \min C_2| = 2^{k-1}$. In this case, let (p_1, s_1), (p_2, s_2) be the two factorizations of order k respectively associated with $C_1, C_2 \in \mathrm{Cut}_k(u)$. Without loss of generality, assume that $|p_1| < |p_2|$. Then, there exists $a \in A$ such that either*

$$|p_1| + |s_1| = |p_2| + |s_2| \text{ and } (p_2, \varphi^{k-1}(a)s_2) = (p_1\varphi^{k-1}(a), s_1)$$

or,

$$||p_1| + |s_1| - (|p_2| + |s_2|)| = 2^k \text{ and } (p_2, s_2) = (p_1\varphi^{k-1}(\bar{a}), \varphi^{k-1}(a)s_1).$$

Example 24. Let us consider $u = 101001011$. It is a factor of $\varphi^2(010)$. We have $\mathrm{Cut}_3(u) = \{\{2\}, \{6\}\}$, which means that $(p_1, s_1) = (10, 1001011)$ and $(p_2, s_2) = (101001, 011)$ are two factorisations of u of order 3. By taking $a = 1$, we have $(p_2, \varphi^2(a)s_2) = (101001, 1001011) = (p_1\varphi^2(a), s_1)$ as claimed in the previous lemma.

6 Types Associated with a Factor

Remark 25. All the following constructions rely on Lemma 23. Thus, in the remaining of this paper, we will always assume that $k \geq 3$.

Lemma 23 ensures us that whenever a word has two cutting sets, then their associated factorizations are strongly related. We will now show that whenever two factors u, v of the same length of $\mathbf{t}$ admit factorizations of order k that are similarly related, then these two words are k-binomially equivalent.

To this aim, we introduce an equivalence relation $\equiv_k$ on the set of pairs $(x, y) \in A^{<2^k} \times A^{<2^k}$. The core result of this section is given by Theorem 31 stating that two words are k-binomially equivalent if and only if their factorizations of order k are equivalent for this new relation $\equiv_k$. So, the computation of $b_{\mathbf{t},k}(n)$ amounts to determining the number of equivalence classes for $\equiv_k$ among the factorizations of order k for words in $\mathrm{Fac}_n(\mathbf{t})$.

Definition 26. *Two pairs (p_1, s_1) and (p_2, s_2) of $A^{<2^k} \times A^{<2^k}$ are equivalent for $\equiv_k$ whenever there exists $a \in A$ such that one of the following situations occurs:*

1. $|p_1| + |s_1| = |p_2| + |s_2|$ *and*
   (a) $(p_1, s_1) = (p_2, s_2)$;
   (b) $(p_1, \varphi^{k-1}(a)s_1) = (p_2\varphi^{k-1}(a), s_2)$;
   (c) $(p_2, \varphi^{k-1}(a)s_2) = (p_1\varphi^{k-1}(a), s_1)$;
   (d) $(p_1, s_1) = (s_2, p_2) = (\varphi^{k-1}(a), \varphi^{k-1}(\bar{a}))$;
2. $\big||p_1| + |s_1| - (|p_2| + |s_2|)\big| = 2^k$ *and*
   (a) $(p_1, s_1) = (p_2\varphi^{k-1}(a), \varphi^{k-1}(\bar{a})s_2)$;
   (b) $(p_2, s_2) = (p_1\varphi^{k-1}(a), \varphi^{k-1}(\bar{a})s_1)$.

Remark 27. Note that if $(p_1, s_1) \equiv_k (p_2, s_2)$, then either $|p_1| = |p_2|$ or, $||p_1| - |p_2|| = 2^{k-1}$. So $(p_1, s_1) \equiv_k (p_2, s_2)$ implies that $|p_1| \equiv |p_2| \pmod{2^{k-1}}$.

The next result is a direct consequence of Lemma 23.

Corollary 28. *If a factor of $\mathbf{t}$ has two distinct factorizations of order k, then these two are equivalent for $\equiv_k$.*

Definition 29 (Type of order k). *Given a factor u of $\mathbf{t}$ of length at least $2^k - 1$, the* type of order k *of u is the equivalence class of a factorization of order k of u. We also let (p_u, s_u) denote the factorization of order k of u for which $|p_u|$ is minimal (we assume that k is understood from the context). Therefore, two words u and v have the same type of order k if and only if $(p_u, s_u) \equiv_k (p_v, s_v)$.*

Example 30. Continuing Example 24, the word u has two factorizations of order 3 that verify case 1.(c) in Definition 26. Thus, $(10, 1001011) \equiv_3 (101001, 011)$ and the type of order 3 of u is $\{(10, 1001011), (101001, 011)\}$.

Theorem 31. *Let u, v be factors of $\mathbf{t}$ of length $n \geq 2^k - 1$. We have*

$$u \sim_k v \Leftrightarrow (p_u, s_u) \equiv_k (p_v, s_v).$$

The proof that the condition is sufficient easily follows from Corollary 13 and [13, Lemma 31].

The proof that the condition is necessary is done in the extended version of this paper [13]. First, we consider the case of words u, v that do not have any non-empty common prefix or suffix and split the result into two lemmas: either $|p_u| \not\equiv |p_v| \pmod{2^{k-1}}$ or, $|p_u| \equiv |p_v| \pmod{2^{k-1}}$. We then add a lemma that permits us to deal with factors having some common prefix or suffix.

7 k-binomial Complexity of the Thue–Morse Word

The first results of this section deal with small factors.

Proposition 32. *Let u, v be two different factors of $\mathbf{t}$ of length $n \leq 2^k - 1$, which do not have any common prefix or suffix. We have $u \not\sim_k v$.*

Corollary 33. *Let $k \geq 3$. For all $n \leq 2^k - 1$, we have $b_{\mathbf{t},k}(n) = p_{\mathbf{t}}(n)$.*

Proof. Let us take two different factors u and v of the same length $n \leq 2^k - 1$. If u and v do not share any common prefix or suffix, $u \not\sim_k v$ by the previous proposition. Otherwise, there exist words x, y, u', v' such that $u = xu'y$ and $v = xv'y$, where u' and v' do not share any common prefix or suffix. We apply the previous proposition to u', v' and conclude because $u' \not\sim_k v'$ implies $u \not\sim_k v$ [13, Lemma 10].

Due to Theorem 31, the k-binomial complexity of $\mathbf{t}$ can be computed from

$$b_{\mathbf{t},k}(n) = \#\left(\mathrm{Fac}_n(\mathbf{t})/\sim_k\right) = \#\left(\{(p_u, s_u) : u \in \mathrm{Fac}_n(\mathbf{t})\}/\equiv_k\right) .$$

The last theorem provides this quantity. The idea of the proof is just to enumerate all the possible factorizations and count them. The proof can be found in the extended version [13].

Theorem 34. *For all $k \geq 3$, $n \geq 2^k$, we have*

$$\#\left(\{(p_u, s_u) : u \in \mathrm{Fac}_n(\mathbf{t})\}/\equiv_k\right) = \begin{cases} 3 \cdot 2^k - 3, \text{ if } n \equiv 0 \pmod{2^k}; \\ 3 \cdot 2^k - 4, \text{ otherwise.} \end{cases}$$

As a consequence of Corollary 33, Theorem 31 and Theorem 34, we get the expected result stated in Theorem 5.

8 Acknowledgments

We would like to thank Jeffrey Shallit for his participation in the statement of the initial problem. A conjecture about $b_{\mathbf{t},k}(n)$ was made when he was visiting the last author.

References

1. J.-P. Allouche, J. Shallit, The ubiquitous Prouhet-Thue-Morse sequence, *Sequences and their applications* (Singapore, 1998), 1–16, Springer Ser. Discrete Math. Theor. Comput. Sci., Springer, London, 1999.
2. J.-P. Allouche, J. Shallit, *Automatic sequences. Theory, applications, generalizations*, Cambridge University Press (2003).
3. J. Berstel, M. Crochemore, J.-E. Pin, Thue-Morse sequence and p-adic topology for the free monoid, *Discrete Math.* **76** (1989), 89–94.
4. V. Berthé, M. Rigo (Eds.), *Combinatorics, Automata and Number Theory*, Encyclopedia Math. Appl., **135**, Cambridge Univ. Press (2010).
5. S. Brlek, Enumeration of factors in the Thue-Morse word, *Discrete Appl. Math.* **24** (1989), 83–96.
6. A. de Luca, S. Varricchio, Some combinatorial properties of the Thue–Morse sequence and a problem in semigroups, *Theoret. Comput. Sci.* **63** (1989), 333–348.
7. D. D. Freydenberger, P. Gawrychowski, J. Karhumäki, F. Manea, W. Rytter, Testing k-binomial equivalence, arXiv:1509.00622.
8. F. Greinecker, On the 2-abelian complexity of the Thue–Morse word, *Theoret. Comput. Sci.* **593** (2015), 88–105.
9. P. Karandikar, M. Kufleitner, Ph. Schnoebelen, On the index of Simon's congruence for piecewise testability, *Inform. Process. Lett.* **115** (2015), 515–519.
10. J. Karhumäki, A. Saarela, L. Q. Zamboni, On a generalization of Abelian equivalence and complexity of infinite words, *J. Combin. Theory Ser. A* **120** (2013), 2189–2206.
11. J. Karhumäki, A. Saarela, L. Q. Zamboni, Variations of the Morse-Hedlund theorem for k-abelian equivalence, *Lect. Notes in Comput. Sci.* **8633** (2014), 203–214.
12. M. Lejeune, *Au sujet de la complexité k-binomiale*, Master thesis, University of Liège (2018), `http://hdl.handle.net/2268.2/5007`.
13. M. Lejeune, J. Leroy, M. Rigo, Computing the k-binomial complexity of the Thue–Morse word (2018), `arXiv:1812.07330`, 34 pages.
14. J. Leroy, M. Rigo, M. Stipulanti, Generalized Pascal triangle for binomial coefficients of words, *Adv. in Appl. Math.* **80** (2016), 24–47.
15. M. Lothaire, *Combinatorics on words*, Cambridge Mathematical Library. Cambridge University Press (1997).
16. M. Morse, G. A. Hedlund, Symbolic dynamics II. Sturmian trajectories, *Amer. J. Math.* **62** (1940), 1–42.
17. B. Mossé, Puissances de mots et reconnaissabilité des points fixes d'une substitution, *Theoret. Comput. Sci.* **99** (1992), 327–334.
18. B. Mossé, Reconnaissabilité des substitutions et complexité des suites automatiques, *Bull. Soc. Math. France* **124** (1996), 329–346.
19. J.-É. Pin, P. V. Silva, A noncommutative extension of Mahler's theorem on interpolation series, *European J. Combin.* **36** (2014), 564–578.
20. N. Pytheas Fogg, *Substitutions in Dynamics, Arithmetics and Combinatorics*, Lect. Notes in Math. **1794**, V. Berthé et al. Eds, Springer (2002).
21. M. Rao, M. Rigo, P. Salimov, Avoiding 2-binomial squares and cubes, *Theoret. Comput. Sci.* **572** (2015), 83–91.
22. M. Rigo, *Formal Languages, Automata and Numeration Systems 1, Introduction to Combinatorics on Words*, Network and Telecommunications series, ISTE-Wiley (2014).

23. M. Rigo, P. Salimov, Another generalization of abelian equivalence: binomial complexity of infinite words, *Theoret. Comput. Sci.* **601** (2015), 47–57.
24. M. Rigo, Relations on words, *Indag. Math. (N.S.)* **28** (2017), 183–204.
25. A. M. Shur, The structure of the set of cube-free words over a two-letter alphabet, *Izv. Math.* **64** (2000), 847–871
26. A. M. Shur, Combinatorial complexity of rational languages. (Russian) *Diskretn. Anal. Issled. Oper. Ser. 1* **12** (2005), 78–99.
27. A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl.* **10**, Christiana (1912).