On winning shifts of marked uniform substitutions

Jarkko Peltomäki and Ville Salo
r@turambar.org, vosalo@utu.fi

Turku Centre for Computer Science TUCS, Turku, Finland
University of Turku, Department of Mathematics and Statistics, Turku, Finland

Abstract

The second author introduced with I. Törmä a two-player word-building game [Playing with Subshifts, Fund. Inform. 132 (2014), 131–152]. The game has a predetermined (possibly finite) choice sequence $\alpha_1, \alpha_2, \ldots$ of integers such that on round n the player A chooses a subset S_n of size α_n of some fixed finite alphabet and the player B picks a letter from the set S_n. The outcome is determined by whether the word obtained by concatenating the letters B picked lies in a prescribed target set X (a win for player A) or not (a win for player B). Typically, we consider X to be a subshift. The winning shift $W(X)$ of a subshift X is defined as the set of choice sequences for which A has a winning strategy when the target set is the language of X. The winning shift $W(X)$ mirrors some properties of X. For instance, $W(X)$ and X have the same entropy. Virtually nothing is known about the structure of the winning shifts of subshifts common in combinatorics on words. In this paper, we study the winning shifts of subshifts generated by marked uniform substitutions, and show that these winning shifts, viewed as subshifts, also have a substitutive structure. Particularly, we give an explicit description of the winning shift for the generalized Thue-Morse substitutions. It is known that $W(X)$ and X have the same factor complexity. As an example application, we exploit this connection to give a simple derivation of the first difference and factor complexity functions of subshifts generated by marked substitutions. We describe these functions in particular detail for the generalized Thue-Morse substitutions.

Keywords: two-player game, winning shift, marked substitution, factor complexity, generalized Thue-Morse word

1 Introduction

In the paper [15], the second author introduced with I. Törmä a two-player word-building game. The two players, Alice and Bob, agree on a finite alphabet S, a target set X of words over S, game length $n \in \mathbb{N} \cup \{\mathbb{N}\}$, and a choice sequence $\alpha_1\alpha_2 \cdots \alpha_n$ (a word) of integers in $\{1, 2, \ldots, |S|\}^n$. On the round j of the game, $1 \leq j \leq n$, Alice first chooses a subset S_j of S of size α_j and then Bob picks a letter a_j from the subset S_j. During the game, Alice and Bob thus together build the word $a_1a_2 \cdots a_n$ (finite or infinite). If this built word is in the target set X, then Alice wins, otherwise Bob does. In other words, Alice aims to build a valid word of X while her adversary Bob attempts to introduce a forbidden word.

In studying games of this sort, it would be typical to fix a choice sequence and see what conditions on X guarantee the existence of a winning strategy for one of the players. The work of [15] adopts the opposite point of view: fix a set X and see for which choice sequences Alice has a winning strategy. This set of choice sequences, dubbed as the winning set $W(X)$ of X, turns out to be a very interesting object. First of all, if X is a subshift, then $W(X)$, now called the winning shift of X, is also a subshift, and the set of factors of $W(X)$ of length k is exactly the winning set of factors of X of length k. Actually the winning set $W(X)$ inherits many properties of X. For instance, if X is a regular language, so is $W(X)$, and if X computable, then so is $W(X)$. The most interesting result, which sparked the research in this paper, is the fact that the sets X and $W(X)$ have the same cardinality so, for a subshift X, the winning shift $W(X)$ has the same entropy and factor complexity function as X. Now the winning set $W(X)$ is in a sense simpler than X because it is downward closed: if any letter of a choice sequence in $W(X)$ is downgraded to a smaller letter, then the resulting word is still in $W(X)$. The winning set $W(X)$ is thus a rearrangement of X to a downward closed set. Indeed, the winning set can be significantly simpler: for instance, the winning set of a Sturmian subshift is the subshift over $\{1, 2\}$ whose words contain at most one letter 2.

Descriptions of the winning shifts for particular subshifts remain largely unknown. In this work, we provide such descriptions for the winning shifts of subshifts generated by marked uniform substitutions. A marked substitution is a substitution such that all images of letters begin with distinct letters and end with distinct letters. We prove that all long enough choice sequences in such a winning shift are obtained from a few core choice sequences by substitution (Theorem 4.9). Let us make this more precise. Let $\tau \colon S^* \to S^*$ be a marked uniform substitution of length M, and let w be a short choice sequence in the language of the winning shift $W(\tau)$ of the subshift generated by τ. Write $w = \diamond ua$ for letters $\diamond$ and a. Then $z\sigma(u)a$ is in the language of $W(\tau)$; here σ is the substitution defined by $\sigma(k) = k1^{M-1}$ and the word z is in the winning set of certain suffixes of the τ-images of a subset of S of size $\diamond$. All long enough choice sequences in the language of $W(\tau)$ are essentially obtained in this way. In general, the short choice sequences and possible words z can be very complex and they elude any simple description, but they can be efficiently computed. This together with Theorem 4.9 allows us to rapidly compute the language of the winning shift $W(\tau)$. If we make additional assumptions on τ, then the situation can be simplified. For instance, if τ is permutive (letters at a fixed position of the τ-images form a permutation of the alphabet S), then z is simply of the form $\diamond 1^i$ for some i such that $0 \leq i < M$ (Proposition 4.10). This class of permutive uniform substitutions includes the generalized Thue-Morse substitutions. For them, we compute all involved parameters and give full description of the whole winning shift (Section 5).

The structure of the winning shift of a marked uniform substitution is quite easy to comprehend, and we apply our results to give a simple derivation of the first difference function of such a substitution (Theorem 4.12). This function can in turn be used to derive the factor complexity function. A. Frid has derived these functions previously with other methods [7]; see also [12]. Our arguments and Frid's arguments, which by the way apply in a more general setting, in the end reduce to the same fundamental observations, but the high-level view is completely different. We prefer gaming and feel that analyzing Alice and Bob's match is fresh and, more importantly, fun. The aim of this paper is to describe the winning shift; the connection to factor complexity is more of a motive for the study, a curiosity. We do, however, derive the factor complexity function in full detail for the generalized Thue-Morse words, just as we describe their winning shifts completely (Section 5). These complexity functions have been derived in full generality previously by Š. Starosta in [14] using an intriguing connection to so-called G-rich words. Results in specialized cases were known before Starosta, see [5, 6, 16]. A short version of this paper with results applying only to the generalized Thue-Morse words was presented in the proceedings of RuFiDiM IV [13].

The paper is organized as follows. In the next section, we give the necessary definitions and results needed. After this in Section 3, we outline the structure of the winning shift of the Thue-

Morse substitution and use it as a motivating example to introduce our ideas. Section 4 contains the main results. We show that generally short choice sequences can be substituted to obtain longer choice sequences, but the additional assumption of markedness is needed for desubstitution. We end Section 4 by deriving a recurrence for the first difference function of a marked uniform substitution. The final section is devoted to the generalized Thue-Morse substitutions. We completely describe their winning shifts and, as an application, derive formulas for their factor complexity functions.

2 Notation and Preliminary Results

2.1 Standard Definitions

Here we briefly define word-combinatorial notions; further details are found in, e.g., [10]. An alphabet S is a nonempty finite set of letters, and we denote by S^* the set of finite words over S. The set of words over S of length n is denoted by S^n, and by $S^{\leq n}$ we denote the set of words over S with length at most n. Infinite words over S are sequences in $S^{\mathbb{N}}$. The length of a finite word w is denoted by $|w|$, and the empty word ε is the unique word of length 0. Suppose that w is a word (finite or infinite) such that $w = uzv$ for some words u, z, and v. Then we say that z is a factor of w. If $u = \varepsilon$ (respectively $v = \varepsilon$), then we call the factor z a prefix (respectively suffix) of w. If $u = \varepsilon$ and $z \neq w$, then z is a proper prefix of w; similarly we define a proper suffix of w. We say that z occurs at position $|u|$ of w; the position $|u|$ is an occurrence of the factor z. Thus we index letters from 0. The word $\partial_{i,j}(w)$, where $i + j \leq |w|$, is obtained from the word w by deleting i letters from the beginning and j letters from the end. An infinite word is ultimately periodic if it is of the form $uvvv\cdots$; otherwise it is aperiodic.

A *subshift* X is a subset of $S^{\mathbb{N}}$ defined by some set F of forbidden words:

$$X = \{w \in S^{\mathbb{N}} : \text{no word of } F \text{ occurs in w}\}.$$

We denote by $\mathcal{L}_X(n)$ the set of words of length n occurring in words of X and define the *language* $\mathcal{L}(X)$ of X as the set $\bigcup_{n \in \mathbb{N}} \mathcal{L}_X(n)$. The subshift X is uniquely defined by its language. The function f defined by letting $f(n) = |\mathcal{L}_X(n)|$ is called the *factor complexity function of* X (we assume that X is known from context), and it counts the number of words of length n in the language of X. We define the *first difference function* Δ by setting $\Delta(n) = f(n) - f(n-1)$ and $\Delta(0) = 1$. This function measures the growth of the factor complexity function.

2.2 Substitutions

A function $\tau \colon S^* \to S^*$ is a called a *substitution* if $\tau(uv) = \tau(u)\tau(v)$ for all $u, v \in S^*$. In this paper, we typically select $S = \{0, 1, \ldots, |S| - 1\}$. If $\tau(s)$ has the same length for every $s \in S$, then we say that τ is *uniform*. In this paper, we assume that for uniform substitutions we have $|\tau(s)| \geq 2$ for all $s \in S$. We call the images of letters, the words $\tau(s)$, τ-*images*. If $\tau(s)$ begins with s and $\lim_{n\to\infty} |\tau^n(s)| = \infty$ for a letter s, then the infinite word obtained by repeatedly applying τ to s, denoted by $\tau^\omega(s)$, is a *fixed point* of the substitution τ. Consider the language $\mathcal{L}$ defined as the set

$$\bigcup_{s \in S} \{w \in S^* : w \text{ occurs in } \tau^n(s) \text{ for some } n \geq 0\}$$

consisting of the factors of the words obtainable by applying τ repeatedly to the letters of S. Let

$$\mathcal{L}(\tau) = \{w \in \mathcal{L} : \text{there exists arbitrarily long words } u \text{ and } v \text{ such that } uwv \in \mathcal{L}\}.$$

The subshift generated by τ is simply the subshift with the language $\mathcal{L}(\tau)$ (i.e., we forbid the complement of $\mathcal{L}(\tau)$). The substitution τ is *primitive* if there is an integer n such that $\tau^n(s)$ contains all letters of S for every $s \in S$. The substitution τ is *aperiodic* if the subshift generated by τ does not contain ultimately periodic infinite words. We assume that all substitutions considered are aperiodic.

We call a substitution τ *left-marked* if all of its τ-images begin with distinct letters. In other words, there exists a permutation $\pi\colon S \to S$ such that $\tau(k) = \pi(k)w_k$ for $k \in S$. Analogously we define *right-marked* substitutions. If a substitution is left-marked and right-marked, then it is simply called a *marked* substitution. Observe also that marked substitutions have an obvious but important property: if a single letter of a τ-image is changed, then the resulting word is no longer a valid τ-image. A substitution is *permutive* if there exists permutations π_1, π_2, ..., π_M from S to S such that $\tau(k) = \pi_1(k)\pi_2(k)\cdots\pi_M(k)$ for $k \in S$. A permutive substitution is uniform and marked.

We say that a word w in $\mathcal{L}(\tau)$ admits an interpretation $(a_0 \cdots a_{n+1}, i, j)$ for letters a_0, ..., a_{n+1} by τ if $w = \partial_{i,j}(\tau(a_0 \cdots a_{n+1}))$, $0 \leq i < |\tau(a_0)|$, $0 \leq j < |\tau(a_{n+1})|$, and $a_0 \cdots a_{n+1} \in \mathcal{L}(\tau)$. The word $a_0 \cdots a_{n+1}$ is called an *ancestor* of the word w. We say that (u_1, u_2) is a *synchronization point* of w (for τ) if $w = u_1u_2$ and whenever $v_1wv_2 = \tau(z)$ for some $z \in \mathcal{L}(\tau)$ and some words v_1 and v_2, then $v_1u_1 = \tau(t_1)$ and $u_2v_2 = \tau(t_2)$ for some words t_1 and t_2 such that $z = t_1t_2$. We say that τ has *synchronization delay* L if every word in $\mathcal{L}(\tau)$ of length at least L has at least one synchronization point and L is minimal. Observe that if τ is marked, then all words in $\mathcal{L}(\tau)$ of length at least L have a unique ancestor. We assume that all substitutions considered in this paper have a synchronization delay. It follows from a theorem of Mossé [11, Corollaire 3.2.] that the synchronization delay of a uniform, primitive, and aperiodic substitution always exists.[1]

Let τ be a uniform substitution of length M with synchronization delay L. Let w in $\mathcal{L}(\tau)$ be a word such that $|w| \geq L$. Suppose that w has an ancestor z, so that $w = \partial_{i,j}(\tau(z))$ with $0 \leq i, j < M$. While w might have several ancestors, the uniformity of τ and the fact that w has at least one synchronization point ensure that the numbers i and j are independent of the chosen ancestor z. In fact, the positions i and j mark a synchronization point of w. All in all, the number i uniquely identifies the positions of w where the τ-images of the letters of any ancestor of w begin at, and we say that w has *decomposition* i mod M.

2.3 Word Games

Next we define precisely the word game in which two players, Alice and Bob, build a finite or infinite word. A *word game* is a quadruple (S, n, X, α), where S is an alphabet, $n \in \mathbb{N} \cup \{\mathbb{N}\}$, the *target set* X is a subset of S^n, and the *choice sequence* α is a word of length n (an infinite word if $n = \mathbb{N}$) over the alphabet $\{1, 2, \ldots, |S|\}$. We may allow the target set X to contain words of distinct lengths by using $X \cap S^n$ in place of X; this will always be clear from context.

Denote by G the word game (S, n, X, α) with $n \in \mathbb{N}$, and write $\alpha = \alpha_1 \cdots \alpha_n$ for letters α_i. During the round i, $1 \leq i \leq n$, of this game, first Alice chooses a subset S_i of S of size α_i. Then Bob picks a letter a_i from the set S_i. After n rounds, Alice and Bob have together built a word $a_1a_2 \cdots a_n$. If $a_1a_2 \cdots a_n \in X$, then Alice wins the game G and otherwise Bob does. An example is provided at the beginning of Section 3, and more examples are found in [15]. The notions presented in this paragraph extend to the case $n = \mathbb{N}$ in a natural way.

Alice's *strategy* for G is a function $s\colon S^{\leq i} \to 2^S$ that specifies which subset of size α_{i+1} she should choose next given the word of length i constructed so far. Similarly we define Bob's strategy as a partial function $s\colon S^{\leq i} \times 2^S \to S$ specifying which letter Bob should pick given the word constructed so far and the subset chosen by Alice. Let s_A and s_B respectively be Alice's

[1] Mossé's Theorem applies to any primitive and aperiodic substitution.

strategy and Bob's strategy for the game G. The *play* $p(G, s_A, s_B)$ of the strategy pair (s_A, s_B) is the word $a_1a_2\cdots a_n$ defined inductively by $a_{i+1} = s_B(a_1\cdots a_i, s_A(a_1\cdots a_i))$ with $a_1\cdots a_0 = \varepsilon$ (if $n = \mathbb{N}$, then the play $a_1a_2\cdots$ is simply infinite). We say that Alice's strategy s is *winning* if $p(G, s, s_B) \in X$ for all Bob's strategies s_B (Alice wins no matter how Bob plays). Analogously Bob's strategy s is winning if $p(G, s_A, s) \notin X$ for all Alice's strategies s_A. If $n \in \mathbb{N}$ or X is a closed set in the product topology of $S^{\mathbb{N}}$ (in particular, if X is a subshift), then a winning strategy always exists for one of the players [8]. In this paper, we consider Bob's strategies only indirectly. Thus whenever we talk about a winning strategy we mean that it is Alice's winning strategy. Similarly by a *winning play* we mean a play by a strategy pair (s_A, s_B) where s_A is Alice's winning strategy.

As mentioned in the introduction, we are interested in the choice sequences for which Alice has a winning strategy. Given a subset X of S^n, where $n \in \mathbb{N} \cup \{\mathbb{N}\}$, we define the *winning set* $W(X)$ of X as the set

$$\{\alpha \in \{1, \ldots, |S|\}^n : \text{Alice has a winning strategy for the word game } (S, n, X, \alpha)\}.$$

Notice that in general Alice has several winning strategies for a choice sequence in $W(X)$ We often omit the alphabet S, it will be clear from the context. For a language $X \subseteq S^*$, we set

$$W(X) = \bigcup_{n\in\mathbb{N}} W(X \cap S^n)$$

and call also this set the winning set of X. If $n = \mathbb{N}$ and X is a subshift, then we call $W(X)$ the *winning shift* of X; if the subshift X is generated by a substitution τ, then we denote its winning shift by $W(\tau)$. Indeed, in [15, Proposition 3.4], the following result was obtained.

Proposition 2.1. *If X is a subshift, then $W(X)$ is a subshift and $\mathcal{L}(W(X)) = W(\mathcal{L}(X))$.*

We abuse notation and write $W(X)$ for $\mathcal{L}(W(X))$, it is always clear from context whether we consider finite words or infinite words. In addition, we have the following observation.

Lemma 2.2. *Let X and Y be sets containing words of equal length. If $X \subseteq Y$, then $W(X) \subseteq W(Y)$.*

Proof. Alice's winning strategy for a word game with target set X and choice sequence in $W(X)$ is sufficient as it is for her to win in the game with the same choice sequence and target set Y. □

We endow the alphabet $\{1, \ldots, |S|\}$ with the natural order $1 < 2 < \ldots < |S|$. Suppose that u and v are words over this alphabet (finite or infinite), and write $u = u_0\cdots u_{n-1}$ and $v = v_0\cdots v_{m-1}$ for letters u_i, v_i. Then we write $u \leq v$ if and only if $n = m$ and $u_i \leq v_i$ for $i = 0, \ldots, n-1$. The winning set $W(X)$ is *downward closed* with respect to this partial ordering: if $u \leq v$ and $v \in W(X)$, then $u \in W(X)$. This is simply because *downgrading* a letter from the choice sequence only makes Bob's chances of winning slimmer.

Observe that the winning strategies for finite choice sequences ending with the letter 1 are just trivial extensions of winning strategies of shorter choice sequences ending with a letter greater than 1. Thus we say that a finite choice sequence is *reducible* if it ends with 1 and *irreducible* otherwise. The infinite words of the winning shift $W(X)$ are obtainable from irreducible choice sequences by appending infinitely many letters 1 and by taking closure. A rule of thumb for the rest of the paper is that to describe the structure of the winning sets it is enough to study only irreducible choice sequences.

Finally, we need the next proposition [15, Proposition 5.7] that motivates the presented results.

Proposition 2.3. *If $n \in \mathbb{N}$ and $X \subseteq S^n$, then $|W(X)| = |X|$.*

We note that a subset W of $\{0, 1\}^n$ can be interpreted as a family of subsets of $\{1, 2, \ldots, n\}$ (a so-called set system) by considering a word $w \in \{0, 1\}^n$ as the characteristic function of a subset. Proposition 2.3 has been proven in relation to set systems in [3].[2]

[2]Formally, the result of [3] corresponds to the binary case of Proposition 2.3. Their *order-shattered sets* for the set system

n		n		n	
1	◇	9	◇11111112	17	◇1111111111111112
2	◇2	10	◇111111112 ◇211111112	18	◇11111111111111112 ◇21111111111111112
3	◇12	11	◇1111111112 ◇1211111112	19	◇111111111111111112 ◇121111111111111112
4	◇112 ◇212	12	◇11111111112 ◇11211111112	20	◇1111111111111111112 ◇1121111111111111112
5	◇1112	13	◇111111111112 ◇111211111112	21	◇11111111111111111112 ◇11121111111111111112
6	◇11112 ◇21112	14	◇1111111111112	22	◇111111111111111111112 ◇111121111111111111112
7	◇111112 ◇121112	15	◇11111111111112	23	◇1111111111111111111112 ◇1111121111111111111112
8	◇1111112	16	◇111111111111112	24	◇11111111111111111111112 ◇11111121111111111111112

Table 1: The irreducible choice sequences of the winning shift of the Thue-Morse substitution for lengths 1 to 24. The letter ◇ can be substituted by both of the letters 1 and 2.

3 The Motivating Example of the Thue-Morse Substitution

In this section, we consider the winning shift of the Thue-Morse substitution. Through examples, we describe the substitutive structure of this winning shift and outline how it can be used to compute the factor complexity of the subshift generated by the Thue-Morse substitution. Our claims are rigorously derived in the subsequent sections in a more general setting.

Let τ be the Thue-Morse substitution: $\tau(0) = 01$, $\tau(1) = 10$. The substitution τ is uniform, primitive, and marked, and it is readily proven that it is aperiodic. With an exhaustive search, it is easily established that its synchronization delay is 4 (see also Lemma 5.2). The fixed point

$$\tau^{\omega}(0) = 0110100110010110100101100110100110010110011010010110100 1\cdots$$

is the famous Thue-Morse word, which is overlap-free (i.e., it does not contain a factor of the form $auaua$ for a word u and a letter a). For more details on the substitution τ, see for example [9, Section 2.2].

In Table 1, we list irreducible choice sequences of $W(\tau)$ for lengths 1 to 24.[3] For the choice sequence 2212, Alice has the following winning strategy:

$$\begin{aligned} \varepsilon &\mapsto \{0,1\}, \\ 0,1 &\mapsto \{0,1\}, \\ 00,10 &\mapsto \{1\}, \\ 01,11 &\mapsto \{0\}, \\ 001,101,010,110 &\mapsto \{0,1\}, \end{aligned}$$

the other arguments being irrelevant. This strategy is depicted in Figure 1 as a *strategy tree*; this tree representation is used throughout this paper. Whenever Alice has more than one choice

whose characteristic functions are $X \subseteq \{0,1\}^n$ correspond to the choice sequences in $W(X^R)^R$, where R is word reversal, that is, their games are played from right to left.

[3]Here we indeed abuse notation, and we should write $\mathcal{L}(W(\tau))$ for $W(\tau)$. Remember that reducible choice sequences of length n are obtained by padding shorter irreducible choice sequences with the letter 1.

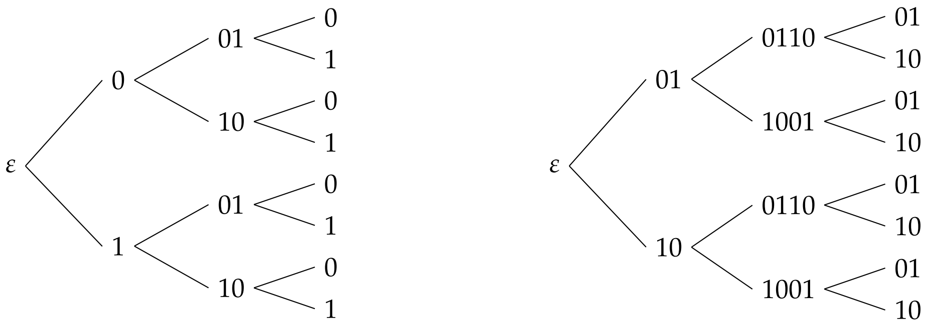

Figure 1: Winning strategies for Alice for the choice sequences 2212 and 21211121 in the case of the Thue-Morse substitution.

according to her strategy, the tree branches to several nodes that correspond to Alice's possible choices of letters. We omit edges from the tree when there are no branchings.

Table 1 contains many patterns. By Proposition 2.3, the number of irreducible choice sequences of length n is counted by the first difference function $\Delta(n)$. Based on the data, it seems that $\Delta(n) \in \{2, 4\}$ for all $n \geq 1$ and $\Delta(n) = 4$ only if $n = 2^k + \ell + 1$ for $k \geq 1$ and $1 \leq \ell \leq 2^{k-1}$. This is of course readily observed when looking at the factor complexity function; here we see much more: the rule described next confirms the preceding observations.

We observe that a choice sequence α in the winning shift always seems to contain at most three occurrences of 2. Moreover, if α contains exactly three occurrences of 2, then the distance between the two final occurrences is $2^k - 1$ for some $k \geq 1$, and the middle occurrence is preceded by at most 2^{k-1} occurrences of the letter 1. The rule seems to be the following. If $n = 3 \cdot 2^k + 2$, then the only irreducible choice sequence of length n (up to the difference at the very beginning) is $\diamond 1^{3 \cdot 2^k} 2$. Then the number of 1s increases until there are $2^{k+2} - 1$ of them. Next a third occurrence of 2 can be introduced: the choice sequences of length $2^{k+2} + 2$ are $\diamond 2 1^{2^{k+2}-1} 2$ and $\diamond 1^{2^{k+2}} 2$ (the former choice sequence downgraded). Then the number of 1s before the second to last occurrence of 2 starts to grow one by one until the choice sequences considered are of length $3 \cdot 2^{k+1} + 1$, and then the pattern repeats. The observed rule suggests that irreducible choice sequences of $W(\tau)$ of lengths $2^k + 2$ to $3 \cdot 2^k + 1$ are related to irreducible choice sequences of lengths $2^{k+1} + 2$ to $3 \cdot 2^{k+1} + 1$. Indeed, these choice sequences look identical: the latter ones are just "blown up" by a factor of 2. Since the substitution τ also "blows up" words by a factor of 2, we proceed to look at τ-images of the strategy trees of short choice sequences.

Consider the strategy tree for the choice sequence 2212 depicted in Figure 1. Substitute all letters of this tree with τ while preserving the branch structure to obtain the right tree of Figure 1. The obtained strategy tree gives a winning strategy for Alice in a word game with choice sequence 21211121. Let us next give an intuitive explanation for the strategy from Alice's point of view. Alice can beat Bob in the word game with choice sequence 21211121 by imagining that she plays the word game with choice sequence 2212, for which she has a winning strategy. On her first turn, Alice lets Bob choose between 0 and 1. Since Alice wins this game of length 1, Alice can also win the game of length 2 with choice sequence 21 played on the τ-images $\tau(0)$ and $\tau(1)$ (choice sequence 11 is also possible but less interesting). Continuing, Alice lets Bob again choose between 0 and 1. The win on this play of length 2 ensures Alice winning the game of length 4 with choice sequence 2121 played on the τ-images $\tau(00)$, $\tau(01)$, $\tau(10)$, $\tau(11)$. Next, Alice gives Bob only one choice to ensure a win, so Bob, having no options, loses in the game of length 6 with choice sequence 212111 played on the respective τ-images. Overall, we see that the short winning strategy for the choice sequence 2212 enables Alice to always win the game with choice sequence 21211121. This longer choice sequence is constructed in such a way that all occasions of

Bob having a real choice (branches of the strategy tree) correspond to Bob having a choice of two letters in the shorter game with choice sequence 2212; Alice just imagines playing a short game with choice sequence 2212 filling the suffixes of the τ-images by not letting Bob choose. Alice's method can indeed be viewed as a branch-preserving substitution of the strategy tree.

The method described above does not explain if it is possible for Alice to obtain a winning strategy for, e.g., the choice sequence 2211121 from some shorter winning strategy. Let us see how she could do this. Alice again imagines playing the winning strategy of the word game with choice sequence 2212 using her winning strategy of Figure 1. Now, however, during the first turn Alice lets Bob pick a suffix of length 1 of the τ-images of the letters 0 and 1 (which Bob is allowed to play on the first turn of the shorter game). Continuing as above, the played word will be a suffix of a word played in the word game with choice sequence 21211121 and a suffix of a τ-image of a word played in the word game with choice sequence 2212. Therefore also $2211121 \in W(\tau)$. Similarly the play on the τ-images does not have to complete the final image, the play can be restricted to a proper prefix of the τ-images. In this particular case of the Thue-Morse substitution, it is easy to be convinced that all long enough winning strategies are obtainable by substitution by working out some example desubstitutions on strategy trees.

In the next section, we will prove that the above methods always produce longer winning strategies from short winning strategies, even in the case of a general uniform substitution. We will show that not all long enough winning strategies are necessarily obtainable from short ones by substitution, but we will show that this holds for marked uniform substitutions. In essence, Alice can derive winning strategies for all long enough choice sequences in $W(\tau)$ from a few core strategies. Moreover, we are able to deduce the first difference function of a marked uniform substitution, which makes it possible to derive a formula for the factor complexity function.

Knowing that winning strategies are obtained by substitution is not enough to give a complete description of the winning shift $W(\tau)$. There is typically some ambiguity on short prefixes of words in $W(\tau)$ due to the fact that they are related to the winning sets of word games played on suffixes of τ-images. The winning sets of proper suffixes of τ-images of a marked substitution can be very complicated—nothing general can be stated about their form. Thus at the end of Section 4, we introduce additional assumptions that simplify these winning sets. We show that the winning sets of proper suffixes of the τ-images of permutive uniform substitutions are trivial, so that $W(\tau)$ admits a complete description. In this case, it can be shown that also the winning shift $W(\tau)$, not only the winning strategies, has a substitutive structure.

Let us conclude this section by describing the substitutive structure of $W(\tau)$ in our example case of the Thue-Morse substitution. Let σ be a substitution defined by $\sigma(1) = 11$ and $\sigma(2) = 21$, and let $\diamond w2$ be an irreducible choice sequence in $W(\tau)$ for a letter $\diamond$. The result is that the words $\diamond\sigma(w)2$ and $\sigma(\diamond w)2$ are in $W(\tau)$ and that all irreducible choice sequences of length at least 5 are obtained in this manner. Thus in our particular example it is sufficient to know all irreducible choice sequences of $W(\tau)$ of length at most 4 to completely describe $W(\tau)$.

4 Main Results

In general, for a uniform substitution $\tau\colon S^* \to S^*$, substituting short winning strategies yields longer winning strategies in a manner similar to what was outlined in the previous section. To figure out the longer choice sequence obtained from a substituted short winning strategy, we need to identify the positions of the τ-images $\tau(A)$ of a subset A of S where Bob can make choices without compromising the chances of Alice winning; in other words, we need to identify the winning set of $\tau(A)$. Notice that in general we obtain many possible choice sequences as the winning set of $\tau(A)$ might contain several words. We also want to consider the winning sets of prefixes and suffixes of these τ-images since we want to include plays where in the beginning Bob plays a proper suffix of a τ-image of a letter and in the end he plays a proper prefix of a τ-image of a letter, just like in the examples of the previous section. Throughout this section, we assume that $\tau\colon S^* \to S^*$ is a uniform and aperiodic substitution of length M with synchronization delay L.

Before formalizing the ideas in the following lemma, we introduce some notation. Let s be Alice's strategy for a word game G. We define its *language* $\mathcal{L}(s)$ to consist of all possible plays with this strategy, that is, it is the set containing all words $p(G, s, s_B)$ for Bob's strategies s_B. Here, we let $\mathcal{L}_s(n)$ denote the set $\mathrm{pref}_n(\mathcal{L}(s))$, that is, $\mathcal{L}_s(n)$ contains the words that are playable in n rounds when Alice uses the strategy s.

Lemma 4.1. *Let s be Alice's winning strategy for a word game $(S, n, \mathcal{L}(\tau), \alpha)$ with $n \geq 2$. Then*

$$W(\mathrm{suff}_i(\tau(s(\varepsilon)))) \cdot \prod_{k=1}^{n-2} \bigcap_{a \in \mathcal{L}_s(k)} W(\tau(s(a))) \cdot \bigcap_{a \in \mathcal{L}_s(n-1)} W(\mathrm{pref}_j(\tau(s(a)))) \subseteq W(\tau)$$

for all integers i and j such that $1 \leq i, j \leq M$.

Proof. Let β be in the set on the left side of the inclusion in the statement of the lemma. Notice that this set is indeed nonempty as the intersected sets all contain the word 1^M or 1^j. We can factorize β as $\beta_0\beta_1 \cdots \beta_{n-1}$, where $|\beta_0| = i$, $|\beta_{n-1}| = j$, and $|\beta_k| = M$ for $1 \leq k < n-1$. We define a strategy s' for Alice for the word game $(S, i + (n-2)M + j, \mathcal{L}(\tau), \beta)$ as follows:

- first Alice plays according to a winning strategy for the game $(S, i, \mathrm{suff}_i(\tau(s(\varepsilon))), \beta_0)$ (such a strategy exist as β_0 was chosen to be in the winning set of $\mathrm{suff}_i(\tau(s(\varepsilon)))$);
- after $i + rM$ rounds have been played, Alice plays according to a winning strategy for the game $(S, M, \tau(s(a)), \beta_{r+1})$, where a is a word in $\mathcal{L}_s(r+1)$ such that $\tau(a)$ has the word of length $i + rM$ played so far as a suffix (the winning strategy exists because β_{r+1} is in $W(\tau(s(a)))$ for all $a \in \mathcal{L}_s(r+1)$);
- finally, after $i + (n-2)M$ rounds, Alice plays according to a winning strategy for the game $(S, j, \mathrm{pref}_j(\tau(s(a))), \beta_{n-1})$, where a is a word in $\mathcal{L}_s(n-1)$ such that $\tau(a)$ has the word of length $i + (n-2)M$ played so far as a suffix (again, the winning strategy exists because β_{n-1} is in $W(\mathrm{pref}_j(\tau(s(a))))$ for all $a \in \mathcal{L}_s(n-1)$).

The described procedure clearly defines a strategy for Alice. What is left is to prove that the strategy s' is a winning strategy for Alice in order to conclude that $\beta \in W(\tau)$.

We show that Bob cannot produce a forbidden word during any round. During the first i rounds Alice plays according to a winning strategy for the word game $(S, i, \mathrm{suff}_i(\tau(s(\varepsilon))), \beta_0)$, so a forbidden word cannot be produced. Suppose then that $i + rM$ rounds have been played without producing a forbidden word. The word played so far is a suffix of length $i + rM$ of the word $\tau(a_0 a_1 \cdots a_r)$ where $a_0 a_1 \cdots a_r \in \mathcal{L}(\tau)$. Alice plays next according to a winning strategy for the word game $(S, M, \tau(s(a_0 a_1 \cdots a_r)), \beta_{r+1})$, so the word played during the first $i + (r+1)M$ rounds is a suffix of length $i + (r+1)M$ of the word $\tau(a_0 a_1 \cdots a_{r+1})$ for some $a_{r+1} \in s(a_0 a_1 \cdots a_r)$. Since s is a winning strategy, we see that $a_0 a_1 \cdots a_{r+1} \in \mathcal{L}(\tau)$, so also $\tau(a_0 a_1 \cdots a_{r+1}) \in \mathcal{L}(\tau)$. This means that no forbidden words are produced during the first $i + (r+1)M$ rounds. Similarly we see that no forbidden word is produced during the final j rounds. We conclude that s' is a winning strategy for Alice. $\square$

Example 4.2. In general, not all choice sequences in $W(\tau)$ are obtainable from shorter ones as in Lemma 4.1. Consider for instance the left-marked substitution

$$\tau\colon \begin{array}{l} 0 \mapsto 001 \\ 1 \mapsto 120 \\ 2 \mapsto 201 \end{array}$$

Figure 2: Example of a long strategy that cannot be desubstituted into a short strategy.

with synchronization delay 5.[4] Its fixed point is

$$001001120001001120120201001001001120001001120120201001120201001201001 \cdots .$$

The left strategy of Figure 2 is a winning strategy of Alice for the choice sequence 12111111111112. Let us show that this strategy is not obtainable from a shorter strategy by substitution. If it would be the case then, by desubstituting the words on the four paths of the strategy tree, we would obtain a winning strategy for Alice. This desubstituted strategy is depicted on the right in Figure 2. The letter $\diamond$ stands for one of the letters 0, 1, and 2; as τ is not right-marked, it is not immediately obvious what $\diamond$ should be. Consider the words $\diamond 01201$ and $\diamond 01202$ corresponding to the two top paths of this desubstituted tree. It is straightforward to see that in $\mathcal{L}(\tau)$ the factor 01201 is extended to the left only by the letter 0, but 01202 is not extended to the left by 0. This means that there is no choice for $\diamond$, so no desubstituted strategy is winning for Alice. Observe that this happens essentially due to the fact that the τ-images of 0 and 2 have a common suffix of length 2. Notice also that the right tree of Figure 2 corresponds by its branch structure to the choice sequence 21112, which can checked not to be in $W(\tau)$.

Next we turn our attention to substitutions whose winning shifts consist essentially only of choice sequences as in Lemma 4.1. We begin with a definition.

Definition 4.3. Let α in $W(\tau)$ be a (finite) choice sequence such that $|\alpha| > L$. If the winning strategies of α are obtainable from the winning strategies of shorter choice sequences in $W(\tau)$ by substitution as in Lemma 4.1, then we call α *substitutive*.

Our first step towards desubstituting long enough winning strategies is to consider left-marked substitutions for which we can prove the following lemma.

Lemma 4.4. *Suppose that τ is left-marked. Let α in $W(\tau)$ be an irreducible choice sequence such that $|\alpha| > L$. Then all winning plays of the game $(S, |\alpha|, \mathcal{L}(\tau), \alpha)$ have decomposition $|\alpha| - 1 \bmod M$.*

Proof. Let s be any winning strategy for Alice for the word game $(S, |\alpha|, \mathcal{L}(\tau), \alpha)$. We will prove that the last branching at the end of the strategy tree of s marks a synchronization point of any winning play, that is, we claim that all winning plays by Alice with strategy s have decomposition $|\alpha| - 1 \bmod M$. Let ua be a word in $\mathcal{L}(s)$ for some word u and letter a, and suppose that ua has decomposition $i \bmod M$ (the decomposition is well-defined as $|ua| \geq L$). Let r be the largest integer such that $rM < |\alpha| - i$. Consider the suffix v of u of length $|\alpha| - rM - i - 1$, so that va is a prefix of $\tau(c)$ for some $c \in S$. Since α is irreducible, the word ub is a winning play for some letter b such that $a \neq b$. Now the word ub must also have decomposition $i \bmod M$ as otherwise deleting the last letter from the words ua and ub would yield two different decompositions $\bmod M$ for the word u contradicting the assumption $|u| \geq L$. Thus by repeating the preceding arguments, we see that vb is a prefix of $\tau(d)$ for some $d \in S$. Since τ is left-marked, the only option is that v is empty. Consequently, we have $i \equiv |\alpha| - 1 \pmod{M}$. Since s was an arbitrary winning strategy, the claim follows. $\square$

[4]This is quite tedious to find by hand, we used a computer.

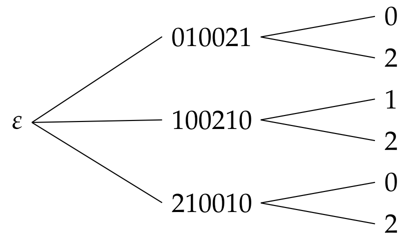

Figure 3: Example of a long winning strategy whose plays have different decompositions $\bmod 3$.

Example 4.5. Continuing Example 4.2, consider the winning plays of the word game with choice sequence 12111111111112, depicted on the left in Figure 2. All four possible plays 10011202010011, 10011202010012, 11200010011200, and 11200010011201 indeed have decomposition $14 - 1 \equiv 1 \pmod 3$.

Lemma 4.4 lets us define the notion of decomposition $\bmod M$ for long enough irreducible choice sequences.

Definition 4.6. Suppose that τ is left-marked, and let α be an irreducible choice sequence in $W(\tau)$ such that $|\alpha| > L$. We say that α has *decomposition i mod M* where i is the unique number such that all winning plays of the game $(S, |\alpha|, \mathcal{L}(\tau), \alpha)$ have decomposition i mod M.

Example 4.7. Let us show that without assuming that τ is left-marked, the claim of Lemma 4.4 is not always true. Consider the primitive substitution

$$\tau\colon \begin{array}{l} 0 \mapsto 021 \\ 1 \mapsto 010 \\ 2 \mapsto 210, \end{array}$$

which is not left-marked, nor are any of its conjugates since τ does not have any.[5] The substitution τ has synchronization delay 6, and its fixed point is

$$021210010210010021021010021210010021021010021021210010021210010021010 \cdots .$$

The strategy tree of Figure 3 shows that $3111112 \in W(\tau)$. Now not all plays with this winning strategy have the same decomposition $\bmod 3$ because in the τ-images only two distinct letters may occur at a fixed position. In fact, we conjecture something stronger: $31^n2 \in W(\tau)$ for infinitely many n.

Before we begin desubstituting long strategies, we prove the following lemma, which gives a description of the form of the choice sequences in $W(\tau)$. Let $\sigma_i\colon \{1, 2, \ldots, |S|\}^* \to \{1, 2, \ldots, |S|\}^*$ be the substitution defined by $\sigma_i(k) = k1^{i-1}$ for $k \in \{1, 2, \ldots, |S|\}$.

Lemma 4.8. *Suppose that τ is left-marked. If α is an irreducible choice sequence in $W(\tau)$ such that $|\alpha| = rM + i + 1 > L$ with $0 \le i < M$ (α has decomposition i* mod *M), then $\partial_{i,1}(\alpha) \in \sigma_M(\{1, 2, \ldots, |S|\}^r)$.*

Proof. If $r = 0$, then there is nothing to prove, so we assume that $r > 0$. Consider the positions $|\alpha| - M - 1$, $|\alpha| - (M-1) - 1$, $\ldots$, $|\alpha| - 2$ of α. Among these positions only the position $|\alpha| - M - 1$ may contain a letter that is greater than 1. Otherwise in some play Bob could make a choice inside a τ-image; recall that the decomposition $\bmod M$ of the plays is fixed before the

[5] The conjugate of a substitution τ is the substitution obtained by cyclically shifting the common prefix of the τ-images. A substitution and its conjugate have the same language.

game even starts, see Lemma 4.4. This is impossible as τ is left-marked. Thus the letters at positions $|\alpha| - M - 1$ to $|\alpha| - 2$ spell out a word of the form $k1^{M-1}$ with $k \in \{1, 2, \ldots, |S|\}$. Thus by repeating this argument $r - 1$ more times, the claim follows. $\square$

Next we consider only marked substitutions and show that then desubstitution is possible.

Theorem 4.9. *Suppose that τ is marked. Let α in $W(\tau)$ be an irreducible choice sequence such that $|\alpha| > L$. Then α is substitutive and if α has decomposition $i \bmod M$, then there exists an irreducible choice sequence $a_0 a_1 \cdots a_{n-1}$ in $W(\tau)$ with winning strategy s such that*

$$\alpha \in W(\mathrm{suff}_i(\tau(s(\varepsilon))))\sigma_M(a_1 \cdots a_{n-2})a_{n-1}$$

where n is the largest integer such that $(n-2)M < |\alpha| - i$.

Proof. Let α in $W(\tau)$ be an irreducible choice sequence having decomposition $i \bmod M$ such that $|\alpha| > L$. Let s be Alice's winning strategy for the word game with choice sequence α. By definition, the strategy tree of s branches at positions where α contains a letter that is greater than 1. Let us show how to perform a branch-preserving desubstitution on s to obtain a shorter winning strategy s'.

Consider first a leaf of the strategy tree of s. Since $i \equiv |\alpha| - 1 \pmod{M}$ by Lemma 4.4, the last letter a of the play corresponding to this leaf is the first letter of some τ-image. Since τ is left-marked, there is a unique letter b such that $\tau(b)$ begins with a. We replace the leaf corresponding to a with a leaf corresponding to b.

Next we show how to desubstitute the factors between two branchings in the middle of the strategy tree s. Say j and k are consecutive positions of α containing letters that are greater than 1 such that $k > j \geq i$. By Lemma 4.8, the factor of α starting at position j and ending at position $k - 1$ is of the form $\ell 1^{tM-1}$ for some $\ell \in \{2, \ldots, |S|\}$. Let w be any winning play with the strategy s. Since w has decomposition $i \bmod M$, it follows that the factor of w starting at position j and ending at position $k - 1$ is a τ-image of some shorter word in $\mathcal{L}(\tau)$. This means that after i rounds have been played, any time Alice's strategy branches, Bob has just completed a τ-image on his previous turn. This means that it is possible to do a branch-preserving desubstitution on the subtrees of length $|\alpha| - i$ of the strategy tree of s: the factor played between two branchings is a τ-image of a shorter word in $\mathcal{L}(\tau)$ and can be directly desubstituted (since τ is injective). If there are no branchings before the final branching, then we can directly desubstitute the factor of any play starting at position i and ending at position $|\alpha| - 2$ (which could be empty).

Now if $i = 0$, then we have desubstituted the whole strategy tree of s, and we are done. Suppose that $i > 0$. As τ is right-marked, the letter at position $i - 1$ of w uniquely identifies a letter a in S such that the prefix of w of length i is a suffix of $\tau(a)$. We modify s by replacing the first i choices by a single choice of a on a path corresponding to the play w. In other words, we let $a \in s'(\varepsilon)$ and set $s'(a)$ to contain the desubstituted subtree obtained above for the suffix of w of length $|\alpha| - i$. Now s' is a strategy and it has the same branch structure as s save for the initial part of i rounds. By construction, all plays by s' are ancestors of the plays with the winning strategy s, so s' must also be a winning strategy. The strategy s is clearly obtained from the strategy s' by substitution as in Lemma 4.1. Therefore α is substitutive. The desubstitution process described clearly indicates that α has the claimed form. $\square$

The essential message of Theorem 4.9 is that knowing all winning strategies for irreducible choice sequences in $W(\tau)$ up to length L is enough to derive winning strategies for all irreducible choice sequences—Alice does not need to learn much to beat Bob. Notice also that we can effectively enumerate $W(\tau)$ when τ is marked, the sets $W(\mathrm{pref}_i(\tau(s(\varepsilon))))$ in the statement of Theorem 4.9 are easily found by exhaustive search.

Notice that substituting a strategy tree by σ_M preserves its branch structure. Conversely, desubstituting, as in Theorem 4.9, preserves most of the branch structure. Indeed, supposing that τ is marked, then the subtree of the winning strategy of a word in $W(\tau)$, as in the third paragraph of the proof of Theorem 4.9, has the same branch structure as the desubstituted subtree. The initial part of the tree comes from a winning set played on suffixes of τ-images. As there are finitely many of these, we conclude that there can be only finitely many different branch structures in the winning trees associated to the winning shift $W(\tau)$. This means that in any choice sequence the number of letters greater than 1 is bounded. In essence, Bob can almost never make a difference: on most turns, he has no options but to play what Alice wants. Compared to real life games, this makes our game somewhat degenerate. We emphasize that a priori it is not clear if Bob gets to play often or not.

Observe that substituting two short winning strategies for two distinct choice sequences of the same length could yield the same longer choice sequence. For instance, if $2u$ and $3u$ are in $W(\tau)$, then cutting a branch of length $|u|+1$ from the winning strategy s for the choice sequence $3u$ yields a winning strategy s' for the choice sequence $2u$. It follows that $W(\mathrm{suff}_i(\tau(s'(\varepsilon)))) \subseteq W(\mathrm{suff}_i(\tau(s(\varepsilon))))$, so all choice sequences obtained by substituting the winning strategy s' are already obtained by substituting the winning strategy s. This is further elaborated in the proof of Theorem 4.12. Moreover, it is possible that by substituting two distinct winning strategies for a fixed choice sequence produces distinct long choice sequences.

Notice that the prefix of α of length i, as in the statement of Theorem 4.9, can be very complicated: we only assume that τ is aperiodic and marked and that it has synchronization delay, so the interior parts of the τ-images can be chosen almost arbitrarily. To simplify the situation, assume that τ is permutive. It is now clear that the suffix games related to the τ-images are trivial: $W(\mathrm{suff}_i(\tau(A))) = \{k1^{i-1}, (k-1)1^{i-1}, \dots, 1^i\}$, where A is a subset of S of k elements. To put it in other words: $W(\mathrm{suff}_i(\tau(A))) = \sigma_i(\{k, k-1, \dots, 1\})$. Thus by Theorem 4.9, we see that the winning shift $W(\tau)$ has the following substitutive structure.

Proposition 4.10. *Suppose that τ is permutive. If α in $W(\tau)$ is an irreducible choice sequence such that $\alpha = \diamond wa$ with letters $\diamond$ and a, then $\sigma_i(\diamond)\sigma_M(w)a$ is in $W(\tau)$ for $1 \le i \le M$, and all choice sequences α of length at least $L+1$ are obtained in this way.*

Since σ_i is injective, the relation of the preceding proposition is a bijection from irreducible choice sequences of length $|\alpha|$ to irreducible choice sequences of length $i + (|\alpha| - 2)M + 1$. Such a bijection exists also in the case where τ is only marked as we shall see next in Theorem 4.12. For its proof, we need the following lemma.

Lemma 4.11. *Let $ku \in W(X)$ for a set X, a letter k, and a word u, and suppose that k is maximal (for u). Then there exists a unique subset A of S of size k such that $s(\varepsilon) \subseteq A$ for all Alice's winning strategies s for a choice sequence tu with $0 \le t \le k$.*

Proof. Let s and s' be two different winning strategies for the choice sequence ku. If $s(\varepsilon) \neq s'(\varepsilon)$, then there would be a letter in, say, $s(\varepsilon) \setminus s'(\varepsilon)$. By removing the subtree of length n associated to this letter from the strategy tree of s and attaching it to the strategy tree of s', we obtain a new strategy. This new strategy clearly is a winning strategy for Alice for the choice sequence $(k+1)u$ contradicting the maximality of the letter k. Thus the set $s(\varepsilon)$ is the same for all Alice's winning strategies s for the choice sequence ku, and we may denote it by A.

Consider then a choice sequence tu with $t < k$, and let e be Alice's arbitrary winning strategy for it. It must be that $e(\varepsilon) \subseteq A$ as otherwise there would be a letter in $e(\varepsilon) \setminus A$, and we could attach the subtree associated to it to the strategy tree of Alice's winning strategy for the choice sequence ku, like above, to obtain a contradiction with the maximality of the letter k. $\square$

The next theorem states the same result as [7, Corollary 3]. For the statement, we define K to be the least integer such that $MK + 1 \geq L$.

Theorem 4.12. *Assume that τ is marked. Suppose that $n \geq K + 2$, and write $n = M^k r + \ell + 1$ with $k \geq 0$, $r \in \{K, K+1, \ldots, KM-1\}$, and $\ell \in \{1, \ldots, M^k\}$. Then $\Delta(n) = \Delta(r + 2)$*

Proof. Consider irreducible choice sequences in $W(\tau)$ of length n ending with a word u of length $n-1$. Let k be the largest letter such that $ku \in W(\tau)$. When a winning strategy for the choice sequence ku is substituted, as in Lemma 4.1, we obtain a winning strategy for an irreducible choice sequence of length $n(i)$, where $n(i) = i + (n-2)M + 1$ with $1 \leq i \leq M$. Moreover, the final $(n-2)M + 1$ letters of such a choice sequence are independent of the prefix k by Theorem 4.9. Further, as $n(i) > L$, Theorem 4.9 implies that all irreducible choice sequences of length $n(i)$ are obtained by substitution. Now there are a total of k irreducible choice sequences of length n with suffix u, so if we show that a total of k distinct irreducible choice sequences of length $n(i)$ are obtainable from them by substitution, then we have shown that there are equally many irreducible choice sequences of length n and $n(i)$.

Let A be as in Lemma 4.11. Consider a choice sequence tu, $0 \leq t \leq k$, with winning strategy s. The choice sequences of length $n(i)$ obtained from tu by substitution are determined by the words in $W(\mathrm{suff}_i(\tau(s(\varepsilon))))$. Lemmas 4.11 and 2.2 imply that $W(\mathrm{suff}_i(\tau(s(\varepsilon)))) \subseteq W(\mathrm{suff}_i(\tau(A)))$, so what is relevant is the size of $W(\mathrm{suff}_i(\tau(A)))$. Lemma 4.11 and Proposition 2.3 show that the size of $W(\mathrm{suff}_i(\tau(A)))$ is k. Therefore a total of k irreducible choice sequences of length $n(i)$ are obtainable from choice sequences with suffix u. As mentioned in the previous paragraph, we have proved that $\Delta(n) = \Delta(n(i))$. The claim follows by a straightforward computation. $\square$

Example 4.13. Theorem 4.12 is not true if τ is only left-marked. Consider for instance the substitution τ of Example 4.2 Now $14 = 3 \cdot 4 + 1 + 1$, so Theorem 4.12 would predict that $\Delta(14) = \Delta(6)$. However, by a direct computation, it can be seen that in this case $\Delta(14) = 5$ but $\Delta(6) = 4$.

Theorem 4.12 can be used to derive the factor complexity function f of a marked uniform substitution τ because $f(n) = 1 + \sum_{i=1}^{n} \Delta(i)$. As the precise details in finding the exact formula do not involve word games, we omit the details and refer the reader to [7, Theorem 2].

Notice also that Theorem 4.12 proves that the first difference function is a M-automatic sequence, so the factor complexity function is a M-regular sequence; see [2]. This holds for arbitrary uniform substitution.

5 Winning Shifts of Generalized Thue-Morse Words

In this section, we describe the winning shifts of generalized Thue-Morse words and, using our results, derive the known formulas for their factor complexity functions. For more on generalized Thue-Morse words, see e.g. [1]. Our notation largely follows [14].

Let $s_b(n)$ denote the sum of digits in the base-b representation of the integer n. For $b \geq 2$ and $m \geq 1$, the generalized Thue-Morse word $\mathbf{t}_{b,m}$ is defined as the infinite word whose letter at position n equals $s_b(n) \bmod m$. It is straightforward to prove that $\mathbf{t}_{b,m}$ is the fixed point, beginning with the letter 0, of the primitive substitution $\varphi_{b,m}$ defined by

$$\varphi_{b,m}(k) = k(k+1)(k+2)\cdots(k+(b-1)),$$

for $k \in \{0, 1, \ldots, m-1\}$, where the letters are interpreted modulo m. The word $\mathbf{t}_{b,m}$ is ultimately periodic if and only if $b \equiv 1 \pmod{m}$ [1]. We make the assumption that $\mathbf{t}_{b,m}$ is aperiodic.

To clarify the notation, from now on we assume that letters are elements of the group $\mathbb{Z}_m$, so that we can naturally add letters. Moreover, we keep b and m fixed and simply write φ for $\varphi_{b,m}$.

Let $\pi\colon \mathbb{Z}_m \to \mathbb{Z}_m$ denote the permutation defined by setting $\pi(k) = k + b - 1$. In other words, the permutation π maps k to the final letter of the word $\varphi(k)$. We set q to be the order of π, that is, the least positive integer such that $q(b-1) \equiv 0 \pmod{m}$.

To describe the winning shift $W(\varphi)$ of φ, it is crucial to know words of $\mathcal{L}(\varphi)$ of length 2 and 3. Our proof is almost verbatim from [14].

Lemma 5.1. *We have*

- $\mathcal{L}_\varphi(2) = \{\pi^i(k-1)k\colon k \in \mathbb{Z}_m, 0 \leq i < q\}$ *and*
- $\mathcal{L}_\varphi(3) = \{\pi^i(k-1)k(k+1)\colon k \in \mathbb{Z}_m, 0 \leq i < q\} \cup \{(k-1)k\pi^{-i}(k+1)\colon k \in \mathbb{Z}_m, 0 \leq i < q\}$.

Proof. Set $L_0 = \{(k-1)k\colon k \in \mathbb{Z}_m\}$. Clearly $L_0 \subseteq \mathcal{L}_\varphi(2)$. Let L_{j+1} to be the set of factors of length 2 of the words in $\varphi(L_j)$. By the definition of π, we have $L_j = \{\pi^i(k-1)k\colon k \in \mathbb{Z}_m, 0 \leq i < j+1\}$. Since $L_q = L_{q-1}$, we have $\mathcal{L}_\varphi(2) = L_{q-1}$.

By the form of φ, either the first two letters of a factor of length 3 are equal or its last two letters are. The claim thus follows from the form of the factors of length 2. □

The following lemma concerning the synchronization delay of φ is proven in [4]; we repeat the proof here.

Lemma 5.2. *The substitution φ has synchronization delay $2b$.*

Proof. Consider a word w of $\mathcal{L}(\varphi)$ of length $2b$. If w contains a factor $k\ell$ with $\ell \neq k+1$, then the factor $k\ell$ cannot occur inside a φ-image, so the position where ℓ occurs marks a synchronization point. If such a factor does not occur in w, then the word w is of the form $k(k+1)\cdots(k+2b-1)$, that is, $w = \varphi(k(k+b))$. Suppose for a contradiction that w has ancestor $x_1x_2x_3$. Due to the form of w, we have $x_2 = x_1 + b$ and $x_3 = x_1 + 2b$, that is, $x_1(x_1+b)(x_1+2b) \in \mathcal{L}(\varphi)$. This is impossible as $x_1 + b \neq x_1 + 1$ and $x_1 + 2b \neq x_1 + b + 1$ due to our assumption that $b \neq 1$. Thus the only ancestor of w is $k(k+b)$. We have thus shown that $L \leq 2b$.

Fix $k \in \mathbb{Z}_m$. Since $k - b = k - 1 + (q-1)(b-1)$, we see that $(k-b)k \in \mathcal{L}(\varphi)$ by Lemma 5.1. Consider the prefix u of $\varphi((k-b)k)$ of length $2b-1$. This prefix has $\varphi(k-1)$ as a suffix, and its prefix of length $b-1$ is a suffix of $\varphi(k-b-1)$. Because $(k-1-b)(k-1) \in \mathcal{L}(\varphi)$, the word u has two ancestors proving that $L \geq 2b$. □

Since φ is permutive, it now follows that every choice sequence in $W(\varphi)$ having length at least $2b+1$ is obtainable by substitution from a shorter choice sequence. Next we describe the choice sequences of length at most $2b$.

Proposition 5.3. *Let α in $W(\varphi)$ be an irreducible choice sequence of length n.*

(i) If $2 \leq n \leq b+1$, then $\alpha = \diamond 1^{n-2}a$ with $\diamond \in \{1,\ldots,m\}$ and $a \in \{2,\ldots,q\}$.

(ii) If $b+2 \leq n \leq 2b$, then $\alpha = \diamond 1^{n-2}a$ or $\alpha = \diamond 1^{\ell}21^{b-1}2$ with $\diamond \in \{1,\ldots,m\}$ and $a \in \{2,\ldots,q\}$.

Moreover, each word of such form is in $W(\varphi)$.

Proof. Consider first the case $2 \leq n \leq b+1$. Write $\alpha = \diamond ur$ with letters $\diamond$ and r, and let w be a winning play in the game with choice sequence α. First we argue that the prefix of w of length $n-1$ is of the form $k(k+1)\cdots(k+n-2)$ for some $k \in \mathbb{Z}_m$, that is, it equals $\varphi_{n-1,m}(k)$. If this were not the case, then this prefix equals $xijy$ for some words x and y and letters i and j such that $j \neq i+1$. Thus w has decomposition $|xi| \bmod b$. Since w is a winning play, Bob cannot choose inside a φ-image, and it must thus be that $|jy|$ is a positive multiple of b. This is impossible as now $|\alpha| > |xijy| \geq b+1$. Due to the restricted form of the prefix of w of length $n-1$, we see that Bob cannot make any choices between his first and last turns, so $\alpha = \diamond 1^{n-2}r$. Suppose for a

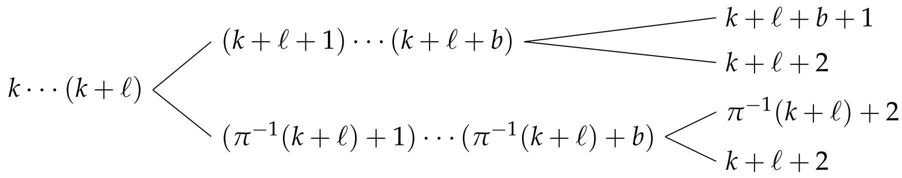

Figure 4: The subtree of Alice's winning strategy after Bob has chosen k in the game with choice sequence $\diamond 1^{\ell} 2 1^{b-1} 2$.

contradiction that $r > q$. Now Bob can pick a letter c such that $c \notin \{\pi^i(k+n-1) \colon 0 \leq i < q\}$. It follows that $\pi^{-1}(k+n-2)c$ is an ancestor of the played word $\varphi_{n-1,m}(k)c$. This is however a contradiction with Lemma 5.1. Therefore $r \leq q$. It is now clear that any word of the form $\diamond 1^{n-2} r$ with $\diamond \in \{1, \ldots, m\}$ and $r \in \{2, \ldots, q\}$ is in $W(\varphi)$: after Bob has chosen k, Alice forces him to play $\varphi_{n-1,m}(k)$ after which she lets him choose among the q letters c such that $\pi^{-1}(k+n-2)c$ is in $\mathcal{L}_\varphi(2)$.

Suppose then that $b+2 \leq n \leq 2b$. If α contains exactly two letters that are greater than 1, one at the beginning and one at the end, then α must again be of the form $\diamond 1^{n-2} a$ with $\diamond \in \{1, \ldots, m\}$ and $a \in \{2, \ldots, q\}$ (after Bob has chosen k, Alice forces him to play $\varphi_{n-b-1}(k)\varphi(\pi^{-1}(k+n-b-2)+1)$ after which she lets him choose among the q letters c such that $\pi^{-1}(k+n-b-2)(\pi^{-1}(k+n-b-2)+1)c \in \mathcal{L}(\varphi)$; see Lemma 5.1). Otherwise write $\alpha = \diamond urvs$ with letters $\diamond$, r, and s such that $r, s > 1$, and let w again be a winning play in the game with choice sequence α. Analogous to the arguments of the preceding paragraph, we see that $|\alpha| > |\diamond urv| \geq 2b+1$ unless the prefix of w of length $|u|+1$ is of the form $\varphi_{|u|+1,m}(k)$ for some $k \in \mathbb{Z}_m$. Again, we have $u = 1^{|u|}$ and, further, $v = 1^{b-1}$. Assume for a contradiction that $r \geq 3$. After $|u|+1$ rounds Bob can choose a letter c such that $c \notin \{k+|u|+1, \pi^{-1}(k+|u|)+1\}$. Clearly the word played so far has decomposition $|u|+1 \bmod b$, so during her next $b-1$ turns Alice must let Bob complete the φ-image beginning with c. During his final turn Bob can pick a letter d such that $d \neq c+1$. It follows that the played word has the word $\pi^{-1}(k+|u|)cd$ as an ancestor. By Lemma 5.1, this ancestor is not in $\mathcal{L}_\varphi(3)$, so Bob wins. This is a contradiction, so $r = 2$. The preceding arguments also show that w must have $\varphi_{|u|+1,m}(k)(k+|u|+1)$ or $\varphi_{|u|+1,m}(k)(\pi^{-1}(k+|u|)+1)$ as a prefix. Let us consider the former case. Since Bob wins if he can choose inside a φ-image, Alice must now force Bob to play $\varphi_{n-1,m}(k)$ to ensure that the word played so far has multiple ancestors. If $s \geq 3$, then as his ultimate move Bob can pick a letter c such that $c \notin \{k+n-1, \pi^{-1}(k+n-2)+1\}$. Then w has unique ancestor $\pi^{-1}(k+n-2-b)\pi^{-1}(k+n-2)c$. Our assumption that $b \neq 1$ implies by Lemma 5.1 that $\pi^{-1}(k+n-2)+1 = c$, which is impossible by the choice of c. Thus $s = 2$, that is, $\alpha = \diamond 1^{|u|} 2 1^{b-1} 2$. It is now straightforward to derive a winning strategy for Alice for any $\diamond \in \{1, \ldots, m\}$. The subtree of length $n-1$ of such a strategy is depicted in Figure 4; it is readily verified that the corresponding strategy is winning for Alice using Lemma 5.1. The claim follows. $\square$

Since φ is permutive, all long enough choice sequences α in $W(\varphi)$ are of the form $\sigma_i(\diamond)\sigma_b(w)a$, where $\diamond wa \in W(\varphi)$ for letters $\diamond$ and a. Combining this with Proposition 5.3, we see that the winning shift $W(\varphi)$ indeed has the same form as described in Section 3. Either α is of the form $\diamond 1^{|\alpha|-2} a$ with $\diamond \in \{1, \ldots, m\}$ and $a \in \{2, \ldots, q\}$ or $\alpha = \diamond 1^{\ell} 2 1^{b^k-1} 2$, where $\diamond \in \{1, \ldots, m\}$, k is the largest k such that $b^k < |\alpha|$ and $0 \leq \ell \leq b^k - b^{k-1} - 1$.

Proposition 5.3 together with Theorem 4.12 implies that for $n \geq 2$ the first difference function $\Delta(n)$ for $\mathbf{t}_{b,m}$ takes only two values: $(q-1)m$ and qm. Using induction, we can derive the values of $\Delta(n)$ and $C(n)$ (the factor complexity function of $\mathbf{t}_{b,m}$) for any $n \geq 1$; see Table 2. These functions have been derived by Š. Starosta with other methods [14].

n	$\Delta(n)$	$C(n)$
1	$m-1$	m
$2 \leq n \leq b+1$	$(q-1)m$	$qm(n-1) - m(n-2)$
$b^{k+1} + \ell + 1$ $k \geq 0, 1 \leq \ell \leq b^{k+1} - b^k$	qm	$qm(n-1) - m(b^{k+1} - b^k)$
$2b^{k+1} - b^k + \ell + 1$ $k \geq 0, 1 \leq \ell \leq b^{k+2} - 2b^{k+1} + b^k$	$(q-1)m$	$qm(n-1) - m(b^{k+1} - b^k + \ell)$

Table 2: The values of the first difference function $\Delta(n)$ and the factor complexity function $C(n)$ of the generalized Thue-Morse word $\mathbf{t}_{b,m}$.

Acknowledgments

The work of the first author was supported by the Finnish Cultural Foundation by a personal grant.

References

[1] J.-P. Allouche and J. Shallit. Sums of digits, overlaps, and palindromes. *Discrete Mathematics and Theoretical Computer Science* 4.1 (2000), 1–10.

[2] J.-P. Allouche and J. Shallit. *Automatic Sequences. Theory, Applications, Generalizations*. Cambridge University Press, 2003.

[3] R. P. Anstee, L. Rónyai, and A. Sali. Shattering News. *Graphs and Combinatorics* 18.1 (2002), 59–73.
DOI: 10.1007/s003730200003.

[4] L. Balková. Factor frequencies in generalized Thue-Morse words. *Kybernetika* 48.3 (2012), 371–385.

[5] S. Brlek. Enumeration of factors in the Thue-Morse word. *Discrete Applied Mathematics* 24 (1989), 83–96.
DOI: 10.1016/0166-218X(92)90274-E.

[6] A. de Luca and S. Varricchio. Some combinatorial properties of the Thue–Morse sequence and a problem in semigroups. *Theoretical Computer Science* 63.3 (1989), 333–348.
DOI: 10.1016/0304-3975(89)90013-3.

[7] A. Frid. On uniform D0L words. *15th Symposium on Theoretical Aspects of Computer Science. STACS'98*. Lecture Notes in Computer Science 1373. Springer, 1998, pp. 544–554.
URL: http://iml.univ-mrs.fr/~frid/Papers/Frid_3.ps.

[8] D. Gale and F. M. Stewart. Infinite games with perfect information. In: *Contributions to the Theory of Games*. Vol. 2. Annals of Mathematics Studies, no 28. Princeton University Press, Princeton, 1953, pp. 245–266.

[9] M. Lothaire. *Combinatorics on Words*. Encyclopedia of Mathematics and Its Applications 17. Addison-Wesley, 1983.

[10] M. Lothaire. *Algebraic Combinatorics on Words*. Encyclopedia of Mathematics and Its Applications 90. Cambridge University Press, 2002.

[11] B. Mossé. Puissances de mots et reconnaissabilitié des points fixes d'une substitution. *Theoretical Computer Science* 99 (1992), 327–334.
DOI: 10.1016/0304-3975(92)90357-L.

[12] B. Mossé. Reconnaissabilitié des substutions et complexité des suites automatiques. *Bulletin de la Société Mathématique de France* 124.2 (1996), 329–346.

[13] J. Peltomäki and V. Salo. On winning shifts of generalized Thue-Morse substitutions. *Proceedings of the Fourth Russian Finnish Symposium on Discrete Mathematics*. TUCS Lecture Notes 26. Turku Center for Computer Science, 2017, pp. 123–132.
URL: http://urn.fi/URN:ISBN:978-952-12-3547-4.

[14] Š. Starosta. Generalized Thue-Morse Words and palindromic richness. *Kybernetika* 48.3 (2012), 361–370.

[15] I. Törmä and V. Salo. Playing with subshifts. *Fundamenta Informaticae* 132 (2014), 131–152.
DOI: 10.3233/FI-2014-1037.

[16] J. Tromp and J. Shallit. Subword complexity of the generalized Thue-Morse word. *Information Processing Letters* 54.6 (1995), 313–316.
DOI: 10.1016/0020-0190(95)00074-M.