Subword complexity and power avoidance

Jeffrey Shallit

School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
Canada
`shallit@uwaterloo.ca`

Arseny Shur
Dept. of Algebra and Fundamental Informatics
Ural Federal University
620000 Ekaterinburg
Russia
`arseny.shur@urfu.ru`

May 30, 2020

Abstract

We begin a systematic study of the relations between subword complexity of infinite words and their power avoidance. Among other things, we show that
– the Thue-Morse word has the minimum possible subword complexity over all overlap-free binary words and all $(\frac{7}{3})$-power-free binary words, but not over all $(\frac{7}{3})^+$-power-free binary words;
– the twisted Thue-Morse word has the maximum possible subword complexity over all overlap-free binary words, but no word has the maximum subword complexity over all $(\frac{7}{3})$-power-free binary words;
– if some word attains the minimum possible subword complexity over all square-free ternary words, then one such word is the ternary Thue word;
– the recently constructed 1-2-bonacci word has the minimum possible subword complexity over all *symmetric* square-free ternary words.

Keywords: combinatorics on words, subword complexity, power-free word, critical exponent, Thue-Morse word

1   Introduction

Two major themes in combinatorics on words are *power avoidance* and *subword complexity* (also called *factor complexity* or just *complexity*). In power avoidance, the main goals are to construct infinite words avoiding various kinds of repetitions (see, e.g., [26]), and to count or estimate the number of length-n finite words avoiding these repetitions (see, e.g., [37]). In subword complexity, the main goal is to find explicit formulas for, or estimate, the number of distinct blocks of length n appearing in a given infinite word (see, e.g., [13]). In this paper we combine these two themes, beginning a systematic study of infinite binary and ternary words satisfying power avoidance restrictions. We follow two interlaced lines of research. First, given a power avoidance restriction, we study the set of infinite words satisfying this restriction, focusing on upper and lower bounds on their subword complexity, and examples of words of "large" and "small" complexity. Second, given a subword complexity restriction, we seek lower bounds on the powers avoided by infinite words of restricted complexity, and for words attaining these bounds. We also tried to cover the remaining white spots with open questions and conjectures. Most of the results are gathered in Table ; precise definitions are given below in Section .

The paper is organized as follows. After definitions, we study α-power-free infinite binary words for $\alpha \leq 7/3$ in Section 2. The number of distinct blocks of length n in this case is quite small and all infinite words are densely related to the Thue-Morse word. We provide answers for most (but not all) questions about complexity of these words. After that, in Section 3 we study low-complexity α-power-free infinite binary and ternary words for the case where α is big enough to provide sets of length-n blocks of size exponential in n. Here we leave more white spots, but still obtain quite significant results, especially about square-free ternary words. Finally, in Section 4 we briefly consider high-complexity infinite words and relate the existence of words of "very high" complexity to an old problem by Restivo and Salemi.

1.1   Definitions and notation

Throughout we let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k-1\}$. By Σ_k^* we mean the set of all finite words over Σ_k, including the empty word ε. By Σ_k^ω we mean the set of all one-sided right-infinite words over Σ_k; throughout the paper they are referred to as "infinite words". The length of a finite word w is denoted by $|w|$.

If $x = uvw$, for possibly empty words u, v, w, x, then we say that u is a *prefix* of x, w is a *suffix* of x, and v is a *subword* (or *factor*) of x. A prefix (resp., suffix, factor) v of x is *proper* if $v \neq x$. A factor v of x can correspond to different factorizations of x: $x = u_1 v w_1 = u_2 v w_2 = \cdots$. Each factorization corresponds to an *occurrence* of v in x; the *position* of an occurrence is the length of the prefix of x preceding v. Thus occurrences of v (and, in particular, occurrences of letters) are linearly ordered by their positions, and we may speak about the "first" or "next" occurrence. An infinite word is *recurrent* if every factor has infinitely many occurrences.

Any map $f : \Sigma_k \to \Sigma_m^*$ $(k, m \geq 1)$ can be uniquely extended to all finite and infinite words over Σ_k by putting $f(a_0 a_1 \cdots) = f(a_0) f(a_1) \cdots$, where $a_i \in \Sigma_k$ for all i. Such extended maps

Table 1: Infinite power-avoiding binary and ternary words of small and large complexity. Question marks indicate conjectured results.

Avoidance restriction	Small complexity	Large complexity
Binary words		
all overlap-free	minimum: Thue-Morse word	maximum: twisted Thue-Morse word
symmetric overlap-free	minimum: Thue-Morse word	maximum: Thue-Morse word
$(\frac{7}{3})$-power-free	minimum: Thue-Morse word	maximum: no; upper bound: $< 4n$
symmetric $(\frac{7}{3})$-power-free	minimum: Thue-Morse word	maximum: Thue-Morse word
$(\frac{7}{3})^+$-power-free	Thue-Morse word is not of minimum complexity	exponential
$(\frac{5}{2})^+$-power-free	minimum(?): $2n$ (new word)	
$(\frac{5+\sqrt{5}}{2})$-power-free	minimum: $n+1$ (Fibonacci word)	
Ternary words		
symmetric $(\frac{7}{4})^+$-power-free	minimal(?) growth: $12n+O(1)$ (Arshon word)	
symmetric $(\frac{5+\sqrt{5}}{4})$-power-free	minimal growth: $6n+O(1)$ (1-3-bonacci word)	
square-free	minimum(?): ternary Thue word	exponential
symmetric square-free	minimum: $6n-6$ (1-2-bonacci word)	
$(\frac{5}{2})$-power-free	minimum(?): $2n+1$ (new word)	

are called *morphisms*. A morphism is a *coding* if it maps letters to letters.

The Thue-Morse word

$$\mathbf{t} = t_0 t_1 t_2 \cdots = 0110100110010110 \cdots$$

is a well-studied infinite word with many equivalent definitions (see, e.g., [1]). The first is that t_i counts the number of 1's, modulo 2, in the binary expansion of i. The second is that $\mathbf{t}$ is the fixed point, starting with 0, of the morphism μ sending 0 to 01 and 1 to 10.

A *language* is a set of finite words over Σ_k. A language L is said to be *factorial* if $x \in L$ implies that every factor of x is also in L. If $\mathbf{u}$ is a one-sided or two-sided infinite word, then by $\mathrm{Fac}(\mathbf{u})$ we mean the factorial language of all finite factors of $\mathbf{u}$. We call a language $L \subseteq \Sigma_k^*$ *symmetric* if $f(L) = L$ for any bijective coding $f : \Sigma_k^* \to \Sigma_k^*$. An infinite word $\mathbf{u}$ is *symmetric* if $\mathrm{Fac}(\mathbf{u})$ is symmetric. For example, the Thue-Morse word is symmetric.

The so-called *subword complexity* or *factor complexity* of an infinite word $\mathbf{x}$ is the function $p_{\mathbf{x}}(n)$ that maps n to the number of distinct subwords (factors) of length n in $\mathbf{x}$. If the context is clear, we just write $p(n)$. A more general notion is the *growth function* (or *combinatorial complexity*, or *census function*) of a language L; it is the function $p_L(n)$ counting the number of words in L of length n. Thus, $p_{\mathbf{x}}(n) = p_{\mathrm{Fac}(\mathbf{x})}(n)$. These complexity functions can be roughly classified by their *growth rate* $\limsup_{n \to \infty}(p(n))^{1/n}$; for factorial languages, the lim sup can be replaced by lim. Exponential (resp., subexponential) words and languages have growth rate > 1 (resp., 1); growth rate 0 implies a finite language. Infinite words constructed by some regular procedure (e.g., those generated by morphisms) usually have small, often linear, complexity (see, e.g., [2]).

We say that an infinite word has *minimum* (resp., *maximum*) subword complexity in a set of words S if $p_{\mathbf{u}}(n) \leq p_{\mathbf{v}}(n)$ (resp., $p_{\mathbf{u}}(n) \geq p_{\mathbf{v}}(n)$) for every word $\mathbf{v} \in S$ and every $n \geq 0$.

An integer power of a word x is a word of the form $x^n = \overbrace{xx \cdots x}^{n}$. If x is nonempty then by x^{ω} we mean the infinite word $xxx \cdots$. Integer powers can be generalized to fractional powers as follows: by x^{α}, for a real number $\alpha \geq 1$, we mean the prefix of length $\lceil \alpha|x| \rceil$ of the infinite word x^{ω}. If u is a finite word and x is the shortest word such that u is a prefix of x^{ω}, then the ratio $|u|/|x|$ is called the *exponent* of u. The *critical* (or *local*) *exponent* of a finite or infinite word $\mathbf{u}$ is the supremum of the exponents of its factors. Thus, for example, the French word `contentent` (as in *ils se contentent*) has a suffix that is the $(3 + \sqrt{5})/2 = 2.61803 \cdots$'th power of the word `nte`, as well as the $(8/3)$'th power of this word; the exponent of the word `contentent` is 1, and its critical exponent is $8/3$.

We say a finite or infinite word *is α-power-free* or *avoids α-powers* if it has no factors that are β-powers for $\beta \geq \alpha$. Similarly, a finite or infinite word *is α^+-power-free* or *avoids α^+-powers* if it has no factors that are β-powers for $\beta > \alpha$. In what follows, we use only the term "α-power", assuming that α is either a number or a "number with a $+$". We write $L_{k,\alpha}$ for the language of all finite k-ary α-power-free words. The growth functions of these power-free languages were studied in a number of papers; see the survey [37] for details. For our study, we need the following rough classification of growth functions of infinite power-free languages [16, 18, 28, 33, 40]: the languages $L_{2,\alpha}$ for $2^+ \leq \alpha \leq 7/3$ have polynomial growth functions, while all other power-free languages are conjectured to have exponential growth functions. This conjecture has been proved for all power-free languages over the alphabets of even size up to 10 and of odd size up to 101. The "polynomial plateau" of binary power-free languages possesses several distinctive properties due to their intimate connection to the Thue-Morse word (see, e.g., [37, Section 2.2]).

An infinite word is called *periodic* if it has a suffix x^{ω} for some nonempty word x; otherwise, it is called *aperiodic*. Obviously, all power-free words are aperiodic. The minimum

subword complexity of an aperiodic word is $n+1$, reached by the class of *Sturmian words* [21].

A finite word x from a language L is *right extendable* in L if for every integer n there is a word v such that $|v| > n$ and $xv \in L$. Left extendability is defined in a dual way. Further, x is *two-sided extendable* in L if for every integer n there are words u, v such that $|u|, |v| > n$ and $uxv \in L$. We write

$$\begin{aligned} \mathsf{rext}(L) &= \{x \in L \mid x \text{ is right extendable in } L\} \\ \mathsf{ext}(L) &= \{x \in L \mid x \text{ is two-sided extendable in } L\} \end{aligned}$$

Note that all factors of an infinite word $\mathbf{u}$ are right extendable in $\mathsf{Fac}(\mathbf{u})$. It is known [36] that for every language L the languages $\mathsf{ext}(L)$ and $\mathsf{rext}(L)$ have the same growth rate as L.

For a word u over $\Sigma_2 = \{0, 1\}$, we say that we *flip* a letter a we replace it with $1 - a$. The word $\overline{u}$ obtained from u by flipping all letters is the *complement* of u.

2 Minimum and maximum subword complexity in small languages

The 2^+-power-free words are commonly called *overlap-free* due to the following equivalent characterization: *a word w contains an α-power with $\alpha > 2$ if and only if two different occurrences of some factor in w overlap.* It is known since Thue [39] that the Thue-Morse word $\mathbf{t}$ is overlap-free. The morphism μ satisfies the following very strong property.

Lemma 1 ([34]). *For every real $\alpha > 2$, an arbitrary word u avoids α-powers iff the word $\mu(u)$ does.*

Moreover, all binary $(7/3)$-power-free (in particular, overlap-free) words can be expressed in terms of the morphism μ. Below is the "infinite" version of a well-known result proved by Restivo and Salemi [28] for overlap-free words and extended by Karhumäki and Shallit [16] to all $(\frac{7}{3})$-power-free words. The second statement of this lemma was proved in [29].

Lemma 2. *Let $\mathbf{u}$ be an infinite $(7/3)$-power-free binary word, $k \geq 0$ be an integer. Then $\mathbf{u}$ is uniquely representable in the form*

$$\mathbf{u} = x_o \mu(x_1 \mu(\cdots x_k \mu(\mathbf{v}) \cdots)) = x_0 \mu(x_1) \cdots \mu^k(x_k) \mu^{k+1}(\mathbf{v}), \tag{1}$$

where $\mathbf{v}$ is also an infinite $(7/3)$-power-free binary word, $x_0, \ldots, x_k \in \{\varepsilon, 0, 1, 00, 11\}$. Moreover, for every $i \geq 1$ the condition $|x_i| = 2$ implies either $|x_{i-1}| = 0$, or $|x_{i-1}| = \cdots = |x_0| = 1$, or $|x_{i-1}| = \cdots = |x_j| = 1$, $|x_{j-1}| = 0$ for some $j \in \{1, \ldots, i-1\}$.

The factorization (1) implies that an infinite $(7/3)$-power-free binary word contains the words $\mu^k(0)$ and $\mu^k(1)$ as factors, for all $k \geq 0$. So we immediately get two corollaries of Lemma 2.

Corollary 3. *Every $(7/3)$-power-free infinite binary word contains, as factors, all elements of $\mathsf{Fac}(\mathbf{t})$. In particular, this is true of every overlap-free word.*

Corollary 4. *The Thue-Morse word* **t** *has the minimum subword complexity among all binary (7/3)-power-free (in particular, overlap-free) infinite words.*

The subword complexity $p_\mathbf{t}$ of the Thue-Morse sequence **t** has been known since the independent work of Brlek [5], de Luca and Varricchio [11], and Avgustinovich [4]. For $n \geq 2$ it is as follows:

$$p_\mathbf{t}(n+1) = \begin{cases} 4n - 2^i, & \text{if } 2^i \leq n \leq 3 \cdot 2^{i-1}; \\ 2n + 2^{i+1}, & \text{if } 3 \cdot 2^{i-1} \leq n \leq 2^{i+1}. \end{cases} \tag{2}$$

We now consider the analogue of the Thue-Morse sequence, where for $n \geq 0$ we count the number of 0's, mod 2, (instead of the number of 1's, mod 2) in the binary representation of n. By convention, we assume that the binary expansion of 0 is ε. We call this word

$$\mathbf{t}' = 00100110100101100110100110010110100110 \cdots = $$
$$00\mu(1)\mu^2(0) \cdots \mu^{2n}(0)\mu^{2n+1}(1) \cdots \tag{3}$$

the *twisted Thue-Morse word*. The word $\mathbf{t}'$ was mentioned in [32] and has appeared previously in the study of overlap-free and (7/3)-power-free words [12]. It is the image, under the coding $\{0, 2\} \to 0$, $1 \to 1$, of the fixed point of the morphism $0 \to 02$, $1 \to 21$, $2 \to 12$, and is known to be overlap-free.

We now state one of our main results. The proof follows after a series of preliminary statements.

Theorem 5. *The twisted Thue-Morse word* $\mathbf{t}'$ *has maximum subword complexity among all overlap-free infinite binary words, and is the unique word with this property, up to complement.*

Remark 6. The word $\mathbf{t}'$ has linear subword complexity, as is proved below, and so contains, as factors, only a small fraction of all right-extendable overlap-free words: the number of such words has superlinear growth (see [17]). This fact is explained in a broader context in Theorem 16.

Remark 7. The uniqueness of $\mathbf{t}'$, stated in Theorem 5, differs strikingly from the situation with minimum complexity, described by Corollary 4. Namely, the language of factors Fac(**t**), and thus the subword complexity $p_\mathbf{t}(n)$, is shared by a continuum of infinite words. The explicit construction of all such words can be found in [34]. (Precisely, Section 2 of [34] describes two-sided infinite words with the language of factors Fac(**t**), but all their suffixes have exactly the same factors.)

A word w is *minimal forbidden* for a factorial language L if $w \notin L$, while all proper factors of w belong to L. Every overlap-free infinite word **u** having a factor not in Fac(**t**) contains a minimal forbidden word x for Fac(**t**); moreover, every such word x appearing in **u** is right extendable in the language $L_{2,2^+}$. These forbidden words are classified in the following lemma.

6

Lemma 8 ([35]). *Let a_k be the last letter of $\mu^k(0)$. The minimal forbidden words for $\mathrm{Fac}(\mathbf{t})$ are exactly the following words and their complements:*

(a) 000;

(b) $r_k = a_k\mu^k(010)0$, $k \geq 0$;

(c) $s_k = a_k\mu^k(101)0$, $k \geq 0$.

Among these, only the words r_k (and their complements) are right extendable for $L_{2,2+}$.

Remark 9. This claim is also easy to prove using the `Walnut` prover [20].

Lemma 10. *Let $\mathbf{u}$ be an infinite overlap-free binary word and let $k \geq 0$. If $\mathbf{u} = xr_k\cdots$ or $\mathbf{u} = x\bar{r}_k\cdots$ for some word x, then $|x| \leq 2^k - 1$. In particular, the words $r_k, \bar{r}_k$ have, in total, at most one occurrence in $\mathbf{u}$.*

Proof. By induction on k. First, consider the case $k = 0$. We have $r_0 = 00100$. Since $\mathbf{u}$ is overlap free, the letter following xr_0 in $\mathbf{u}$ is 1. On the other hand, if x is nonempty, it must end with either 0 or 1, and in both cases xr_01 has an overlap. So x must be empty.

Now the induction step. Assume the claimed result is true for $k' < k$; we prove it for k. Let $\mathbf{u} = xr_k\mathbf{v} = xa_k\mu^k(010)0\mathbf{v}$ and assume $|x| \geq 2^k$. The factorization (1) of $\mathbf{u}$ implies $\mathbf{v} = 1\mu(\mathbf{v}')$ for some infinite word $\mathbf{v}'$, which is overlap-free by Lemma 1. Next, note that $x = y\bar{a}_k$ for some nonempty word y. Indeed, $\mu^k(010)$ begins with 0110. If x ends with a_k, then $\mathbf{u}$ contains the factor a_ka_k0110, preceded by some letter(s); this certainly creates an overlap. Thus
$$\mathbf{u} = y\bar{a}_ka_k\mu^k(010)01\mu(\mathbf{v}') = y\mu(\bar{a}_k\mu^{k-1}(010)0\mathbf{v}').$$

On the other hand, $\mathbf{u} = x_0\mu(\mathbf{u}')$ for some $x_0 \in \{\varepsilon, 0, 00, 1, 11\}$ and some infinite word $\mathbf{u}'$ by Lemma 2. The uniqueness of factorization in Lemma 2 implies $y = x_0\mu(x')$ for some word x' (possibly empty). Observing that $\bar{a}_k = a_{k-1}$, we finally write

$$\mathbf{u} = x_0\mu(\mathbf{u}'), \quad \text{where } \mathbf{u}' = x'a_{k-1}\mu^{k-1}(010)0\mathbf{v}'.$$

Applying the inductive hypothesis to $\mathbf{u}'$, we obtain $|x'| \leq 2^{k-1} - 1$. Having $|x| = |x_0| + 2|x'| + 1$ and $|x| \geq 2^k$, we obtain $|x'| = 2^{k-1} - 1$. Since $|x'|$ has the maximum possible length, the inductive hypothesis guarantees that both words $0\mathbf{u}'$ and $1\mathbf{u}'$ contain overlaps. Then the word $\mathbf{u}'$ has some prefix uu that ends with 0, and also some prefix vv that ends with 1 (e.g., $\mathbf{u}'$ can be a word of the form $001001\cdots$). Hence $\mu(\mathbf{u}')$ has the prefix $\mu(u)\mu(u)$ that ends with 1, and the prefix $\mu(v)\mu(v)$ that ends with 0. This means that both words $0\mu(\mathbf{u}')$ and $1\mu(\mathbf{u}')$ contain overlaps, and thus $x_0 = \varepsilon$. So we have $|x| = 2|x'| + 1 = 2^k - 1$. This contradicts our assumption $|x| \geq 2^k$ and thus proves the inductive step. $\square$

Recall that a factor v of $\mathbf{u}$ is (right) $\mathbf{u}$-*special* if both $v0$ and $v1$ are factors of $\mathbf{u}$. The set of all special factors of $\mathbf{u}$ is denoted by $\mathsf{Spec}(\mathbf{u})$. We will omit $\mathbf{u}$ when it is clear from the context. We use the following familiar fact:

Lemma 11. *The number $D_{\mathbf{u}}(n) = \mathsf{Spec}(\mathbf{u}) \cap \Sigma_2^n$ is the first difference of the subword complexity of $\mathbf{u}$: $D_{\mathbf{u}}(n) = p_{\mathbf{u}}(n+1) - p_{\mathbf{u}}(n)$.*

Proof. Consider the function mapping every word from $\mathrm{Fac}(\mathbf{u})$ of length $n+1$ to its prefix of length n. Each special factor of $\mathbf{u}$ of length n has two preimages, while each non-special factor of length n has a single preimage. $\square$

Corollary 12. *For all $n \geq 1$ and every infinite binary word $\mathbf{u}$ we have*

$$p_{\mathbf{u}}(n) = 2 + \sum_{1 \leq i < n} D_{\mathbf{u}}(i). \tag{4}$$

Since $D_{\mathbf{u}}(0) = 1$ for every binary word, below we we restrict our attention to $\{D_{\mathbf{u}}(n)\}$ for $n \geq 1$. For example,

$$D_{\mathbf{t}}(n) = \begin{cases} 4, & \text{if } n = 2^k + i \text{ for some } k \geq 1, \, i > 0, \, i \leq 2^{k-1}; \\ 2, & \text{otherwise} \end{cases}$$

as was first computed in [5]. As an infinite word over $\{2, 4\}$, this sequence looks like

$$2242442244442222444444422222224 \cdots \tag{5}$$

where each subsequent block of equal letters is twice the size of the previous block of the same letter.

Let $\mathbf{u}$ be overlap-free. By Corollary 4, all $\mathbf{t}$-special factors are $\mathbf{u}$-special, so $D_{\mathbf{u}}(n) \geq D_{\mathbf{t}}(n)$ for all n. We call $\mathbf{u}$-special factor *irregular* if it is not $\mathbf{t}$-special.

Lemma 13. *For an overlap-free infinite binary word $\mathbf{u}$, all $\mathbf{u}$-special factors are Thue-Morse factors.*

Proof. By Lemmas 8 and 10, every word from $\mathrm{Fac}(\mathbf{u}) \setminus \mathrm{Fac}(\mathbf{t})$ occurs in $\mathbf{u}$ only once and thus is not $\mathbf{u}$-special. Hence $\mathsf{Spec}(\mathbf{u}) \subseteq \mathrm{Fac}(\mathbf{t})$. $\square$

Proof of Theorem 5. Let $\mathbf{u} = 0 \cdots$ be an overlap-free infinite binary word (the case $\mathbf{u} = 1 \cdots$ is parallel, so we omit it). We show the following four facts:

(i) at every position in $\mathbf{u}$, the first occurrence of at most one irregular $\mathbf{u}$-special factor begins;

(ii) for every $n \geq 0$, there is an irregular $\mathbf{t}'$-special factor v_n with the first occurrence beginning at position n of $\mathbf{t}'$;

(iii) if w_n is an irregular $\mathbf{u}$-special factor with the first occurrence beginning at position n of an overlap-free word $\mathbf{u}$, then $|w_n| \geq |v_n|$;

8

(iv) the inequality $|w_n| \geq |v_n|$ is strict for at least one value of n.

(i) Let v be an irregular $\mathbf{u}$-special factor. By Lemma 13, $v \in \text{Fac}(\mathbf{t})$; but either $v0$ or $v1$ is not a Thue-Morse factor by definition of irregularity. W.l.o.g., $v0 \notin \text{Fac}(\mathbf{t})$. Then some suffix of $v0$ is a minimal forbidden word for $\text{Fac}(\mathbf{t})$. By Lemma 8, this suffix equals r_k for some $k \geq 0$. So $v0 = v'r_k$, where $|v'| < 2^k$ by Lemma 10. Thus, the first occurrence of v in $\mathbf{u}$ is followed by 0 (and $v0 \notin \text{Fac}(\mathbf{t})$), while all other occurrences of v are followed by 1 (and $v1 \in \text{Fac}(\mathbf{t})$). Now assume that some proper prefix w of v is also an irregular $\mathbf{u}$-special factor. Since $v \in \text{Fac}(\mathbf{t})$, the occurrence of w as a prefix of v is not the first occurrence of w in $\mathbf{u}$. Hence the first occurrences of each two irregular $\mathbf{u}$-special factors begin in different positions.

(ii) From (3) it is easy to see that for every even k the prefix of $\mathbf{t}'$ of length $5 \cdot 2^k$ equals $v\mu^k(0100)$, where $v = 0 = \mu^k(0)$ for $k = 0$ and v has the common suffix $\mu^{k-1}(1)$ with $\mu^k(0)$ for $k > 0$. Similarly, for every odd k such a prefix equals $v\mu^k(1011)$, where v has the common suffix $\mu^{k-1}(0)$ with $\mu^k(1)$. According to the above description of the irregular special factors, $\mathbf{t}'$ has first occurrences of irregular special factors beginning at each position:

Factor	Position	Length	
0010	0	4	
0100110	1	7	
10011010010110	2	14	(6)
0011010010110	3	13	
$\vdots$	$\vdots$	$\vdots$	
$r\mu^k(010), r$ is a suffix of $\mu^{k-1}(0)$	$2^{k-1} + i,\ 0 \leq i < 2^{k-1}$	$4 \cdot 2^k - 2^{k-1} - i$	

(iii) If v is an irregular $\mathbf{u}$-special factor and its first occurrence is followed by a, then va is not Thue-Morse and thus has, by Lemma 8, the suffix r_k or s_k for some $k \geq 0$ (because v is Thue-Morse). Hence by Lemma 10, v is a suffix of the prefix $r'\mu^k(010)$ of $\mathbf{u}$, where $0 < |r'| \leq 2^k$. Therefore the factor v of length $3 \cdot 2^k + i$ occurs in $\mathbf{u}$ for the first time at position at most $2^k - i$ $(1 \leq i \leq 2^k)$. Inverting this, the factor w_n such that $2^k \leq n < 2^{k+1}$ $(k \geq 0)$ is of length at least $3 \cdot 2^{k+1} + 2^{k+1} - n = 2^{k+3} - n$. But this number is exactly the length of v_n for every $n \geq 1$; see (6). Since v_0 is the shortest irregular special factor, we proved the required statement.

(iv) Note that the lengths of irregular special factors, given in (6), and the first letter of u allow one to restore the whole word in a unique way, and this word is $\mathbf{t}'$. Hence, for any other word $\mathbf{u}$ starting with 0 we have $|w_n| > |v_n|$ for some n.

To finish the proof, consider the function $f : \text{Spec}(\mathbf{u}) \rightarrow \text{Spec}(\mathbf{t}')$ that maps every $\mathbf{t}$-special factor to itself and every irregular factor w_n to the irregular factor v_n. The function f is well-defined by (ii) and injective by (i). By (iii), ϕ never increases the length of the word. Finally, (iv) implies that f decreases the length of some w_{n_0}. Together, these facts

imply that for every n,

$$\sum_{1 \leq i < n} D_{\mathbf{u}}(i) \leq \sum_{1 \leq i < n} D_{\mathbf{t}'}(i),$$

and the inequality is strict for $n = n_0$. The claim of the theorem is now immediate from (4). □

As a consequence of the proof, we can determine a closed form for the subword complexity of $\mathbf{t}'$.

Proposition 14. *The number of special $\mathbf{t}'$-factors of length $n > 0$ is given by the formula*

$$D_{\mathbf{t}'}(n) = \begin{cases} 4, & 2^{k+1} < n \leq 3 \cdot 2^k \text{ for some } k \geq 0; \\ 3, & n = 4 \text{ or } 3 \cdot 2^k < n \leq 7 \cdot 2^{k-1} \text{ for some } k \geq 1; \\ 2, & \text{otherwise.} \end{cases}$$

Proof. The proposition states that the sequence $\{D_{\mathbf{t}'}(n)\}$ ($n \geq 1$) can be written as the following word over $\{2, 3, 4\}$:

$$2243443244443324444444333322224\cdots$$

Comparing this to (5), we see that some 2's have been changed to 3's. This means an additional $\mathbf{t}'$-special factor for each corresponding length, and this factor must be irregular. According to (6), the set of all positions of 3's indeed coincides with the set of lengths of irregular $\mathbf{t}'$-special factors, thus proving the proposition. □

Corollary 15. *The maximum factor complexity of a binary overlap-free infinite word is the factor complexity of the twisted Thue-Morse word $\mathbf{t}'$ and is given, for $n \geq 4$, by the formula*

$$p_{\mathbf{t}'}(n+1) = \begin{cases} 4n - 3 \cdot 2^{i-2}, & \text{if } 2^i \leq n \leq 3 \cdot 2^{i-1}; \\ 3n + 3 \cdot 2^{i-2}, & \text{if } 3 \cdot 2^{i-1} \leq n \leq 7 \cdot 2^{i-2}; \\ 2n + 5 \cdot 2^{i-1}, & \text{if } 7 \cdot 2^{i-2} \leq n \leq 2^{i+1}. \end{cases} \tag{7}$$

Proof. Immediate from Theorem 5, Proposition 14, and formula (4). □

A table of the first few values of the subword complexity of $\mathbf{t}'$ follows:

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$p_{\mathbf{t}'}(n)$	1	2	4	6	10	13	17	21	24	26	30	34	38	42	45	48	50	52	56

2.1 Beyond overlap-free words

Now we turn to the case of α-power-free infinite binary words for arbitrary α from the half-open interval $(2, \frac{7}{3}]$. By Corollary 4, the Thue-Morse word and its complement are words of minimum complexity. Theorem 16 below shows that the asymptotic growth of subword complexity belongs to a very small range. Theorem 17 demonstrates that there is no $(7/3)$-power-free infinite binary word of maximum complexity.

Theorem 16. *Let $\alpha \leq 7/3$. Every infinite binary word $\mathbf{u}$ avoiding α-powers has linear subword complexity. Moreover, for every $n > 0$ one has $p_{\mathbf{u}}(n) < \frac{6}{5} \cdot p_{\mathbf{t}}(n)$.*

Proof. First assume that $\mathbf{u} = \mu^m(\mathbf{v})$ for some word $\mathbf{v}$ and some $m \geq 0$. Then $\mathbf{v}$ also avoids α-powers by Lemma 1. According to Lemma 8, the shortest word that can be a factor of $\mathbf{v}$, but does not occur in $\mathbf{t}$, is either 00100 or 11011. Hence every factor of $\mathbf{u}$ that is contained in four consecutive blocks of the form $\mu^m(a)$, $a \in \Sigma_2$, is a factor of $\mathbf{t}$. Thus the shortest factor of $\mathbf{u}$ that is not in $\mathbf{t}$ has the length at least $3 \cdot 2^m + 2$. (If $\mathbf{v}$ contains 00100, this factor is r_k from Lemma 8.) So we have

$$\mathbf{u} = \mu^m(\mathbf{v}) \implies p_{\mathbf{u}}(n) = p_{\mathbf{t}}(n) \text{ for every } n = 0, 1, \ldots, 3 \cdot 2^m + 1. \tag{8}$$

Now let $\mathbf{u}$ be arbitrary. Still, $\mathbf{u}$ satisfies (8) with $m = 0$, so $p_{\mathbf{u}}(n) = p_{\mathbf{t}}(n)$ for $n \leq 4$. So we take an arbitrary $n \geq 5$ and choose a unique integer $m \geq 1$ satisfying the condition $3 \cdot 2^{m-1} + 1 < n \leq 3 \cdot 2^m + 1$. Consider the factorization of $\mathbf{u}$ of type (1):

$$\mathbf{u} = x_0\mu(x_1\mu(\cdots x_{m-1}\mu(\mathbf{v})\cdots)) = x_0\mu(x_1)\cdots\mu^{m-1}(x_{m-1})\mu^m(\mathbf{v}).$$

By (8), all factors of $\mu^m(\mathbf{v})$ of length n are Thue-Morse factors, so we have

$$p_{\mathbf{u}}(n) - p_{\mathbf{t}}(n) \leq |x_0\mu(x_1)\cdots\mu^{m-1}(x_{m-1})| = \sum_{i=0}^{m-1} 2^i|x_i|.$$

This upper bound is a bit loose; to tighten it, consider x_{m-1}. By Lemma 2, $|x_{m-1}| \leq 2$. Let $|x_{m-1}| = 2$ (w.l.o.g., $x_{m-1} = 00$). Then

$$\mu^{m-1}(x_{m-1})\mu^m(\mathbf{v}) = \mu^{m-1}(0)\mu^{m-1}(0100110)\cdots$$

Since 0100110 is a Thue-Morse factor, so is $\mu^{m-1}(0100110)$. Hence the number of length n words in $\mathrm{Fac}(\mathbf{u})\backslash\mathrm{Fac}(\mathbf{t})$ is at most $2^{m-1} + \sum_{i=0}^{m-2} 2^i|x_i|$. Applying the second statement of Lemma 2 to x_{m-1}, we easily obtain $\sum_{i=0}^{m-2} 2^i|x_i| \leq \sum_{i=0}^{m-2} 2^i < 2^{m-1}$. Therefore,

$$p_{\mathbf{u}}(n) - p_{\mathbf{t}}(n) < 2^m. \tag{9}$$

Next let $x_{m-1} = 0$ and let $\mathbf{v} = abc\cdots$, $a, b, c \in \{0, 1\}$. If $a = b = c$ or $a = b = 1$, then $\mathbf{u}$ is not $(7/3)$-power free. Otherwise, $1abc$ is a Thue-Morse factor, as well as the suffix

11

$\mu^{m-1}(0)\mu^m(abc)$ of $\mu^m(1abc)$. Then all length-n words of $\mathbf{u}$, that are not in $\mathrm{Fac}(\mathbf{t})$, begin in $\mathbf{u}$ on the left of $\mu^{m-1}(x_{m-1})$. But $\sum_{i=0}^{m-2} 2^i|x_i| < 2^m$, so we again have (9). By the same reason we obtain (9) in the case $x_{m-1} = \varepsilon$.

Finally, we apply (2) to get

$$p_\mathbf{t}(n) \geq p_\mathbf{t}(3 \cdot 2^{m-1} + 2) = 2 \cdot (3 \cdot 2^{m-1} + 1) + 2^{m+1} = 5 \cdot 2^m + 2 \qquad (10)$$

and compare (9) to (10) to obtain the required inequality. $\qquad\square$

From the proof of Theorem 16 we have the following: for arbitrary $(7/3)$-power-free infinite words $\mathbf{u}$,

(a) if $n \leq 4$, then $p_\mathbf{u}(n) = p_\mathbf{t}(n)$;

(b) if $5 \leq n \leq 7$, we write $\mathbf{u} = x_0\mu(\mathbf{v})$ and see that $p_\mathbf{u}(n) \leq p_\mathbf{t}(n) + 1$; the only length n factor of $\mathbf{u}$ that is possibly not in $\mathrm{Fac}(\mathbf{t})$ is the prefix of $\mathbf{u}$;

(c) if $8 \leq n \leq 13$, we write $\mathbf{u} = x_0\mu(x_1)\mu^2(\mathbf{v})$ and analyze cases to get $p_\mathbf{u}(n) \leq p_\mathbf{t}(n) + 2$ for $n = 8,9$ and $p_\mathbf{u}(n) \leq p_\mathbf{t}(n) + 3$ for $n = 10, 11, 12, 13$. Further, we note that every word $\mathbf{u}_1$ having the property $p_{\mathbf{u}_1}(8) = p_\mathbf{t}(8) + 2$ and beginning with 0 is of the following form:

$$\mathbf{u}_1 = 00\mu(1)\mu^2(01\cdots) = 00\,10\,0110\,1001\cdots ,$$

and $p_{\mathbf{u}_1}(n) = p_\mathbf{t}(n) + 2$ for all n in the considered range (the two "additional" factors are those beginning at positions 0 and 1). On the other hand, every word $\mathbf{u}_2$ with the property $p_{\mathbf{u}_1}(10) = p_\mathbf{t}(10) + 3$, starting with 0, has the form

$$\mathbf{u}_2 = 0\mu(00)\mu^2(101\cdots) = 0\,0101\,1001\,0110\,1001\cdots ,$$

and one has $p_{\mathbf{u}_2}(8) = p_\mathbf{t}(8) + 1$ (the only additional factor 10110010 begins at position 2), $p_{\mathbf{u}_2}(9) = p_\mathbf{t}(9) + 2$ (the additional factors begin at positions 1 and 2), and $p_{\mathbf{u}_2}(n) = p_\mathbf{t}(n) + 3$ for $n = 10, 11, 12, 13$ (the additional factors begin at positions 0, 1, and 2). We see that $\mathbf{u}_1$ reaches the maximum possible complexity for $n = 8$ but not for $n = 10$, while $\mathbf{u}_2$ reaches this maximum for $n = 10$ but not for $n = 8$.

Thus we have proved the following result:

Theorem 17. *There is no $(7/3)$-power-free infinite binary word of maximum complexity.*

If there is no maximum subword complexity, it makes sense to look at some sort of "asymptotically maximal" complexity. For a function $y(n)$ of linear growth, let its *linear growth constant* be $\limsup_{n\to\infty} \frac{y(n)}{n}$. For overlap-free infinite binary words, the maximum subword complexity has linear growth constant $7/2$, as can be easily derived from (7). From (2) we see that such a constant for the Thue-Morse word is $10/3$. By Theorem 16, this means that the linear growth constants of all $(7/3)$-power-free infinite binary words are upper bounded by 4.

Open Question 18. *What is the maximum linear growth constant for the subword complexity of a $(7/3)$-power-free infinite binary word? Which words have such complexity?*

2.2 The symmetric case

Theorem 19. *The only possible subword complexity function of a $(7/3)$-power-free symmetric infinite binary word is the function $p_{\mathbf{t}}(n)$.*

Proof. Let $\mathbf{u}$ be a $(7/3)$-power-free infinite binary word such that $p_{\mathbf{u}}(n) \neq p_{\mathbf{t}}(n)$. By Corollary 4, $\mathbf{u}$ contains a factor that is not Thue-Morse. By Lemma 8, $\mathbf{u}$ contains one of the factors r_k, s_k, or their complements. Assume that $\mathbf{u}$ contains $r_k = a_k \mu^k(010)0$. Taking the representation (1) of $\mathbf{u}$, we see that $\mathbf{u}$ has the suffix $a_k \mu^k(0100 \cdots)$. If the word $\mu^k(0100 \cdots)$ contains the factor $\bar{r}_k$, it contains either $\mu^k(111)$ or $\mu^k(0110110)$, which is impossible because $\mathbf{u}$ is $(7/3)$-power free. Thus, $\bar{r}_k$ cannot occur in $\mathbf{u}$ to the right of an occurrence of r_k. Similarly, if $\mathbf{u}$ contains $s_k = a_k \mu^k(101)0$, then the suffix $\mu^k(1010 \cdots)$ of $\mathbf{u}$ cannot contain $\bar{s}_k$ without containing the $(5/2)$-power $\mu^k(10101)$.

Repeating the same argument for the factors $\bar{r}_k, \bar{s}_k$ in $\mathbf{u}$ we conclude that $\mathbf{u}$ contains neither $r_k, \bar{r}_k$ simultaneously, nor $s_k, \bar{s}_k$ simultaneously. So $\mathbf{u}$ is not symmetric. Hence every $(7/3)$-power-free symmetric infinite binary word has subword complexity $p_{\mathbf{t}}(n)$. $\square$

3 Small subword complexity in big languages

In this section we study binary and ternary words. Note the interconnection of the results over Σ_2 and Σ_3 through the encodings of words in both directions. To ease the reading, we denote words over Σ_3 by capital letters and other words by small letters.

Our study follows two related questions about small subword complexity:

- Given a pair (k, α), how small can the subword complexity of an α-power-free infinite k-ary word be?

- Given an integer k and a function $f(n)$, what is the smallest power that can be avoided by an infinite k-ary word with a subword complexity bounded above by $f(n)$?

3.1 Ternary square-free words

Among α-power-free infinite ternary words, the most interesting are square-free ($=$ 2-power-free) words, the existence of which was established by Thue [38] and in particular $(\frac{7}{4})^+$-power-free words, because $\alpha = (\frac{7}{4})^+$ is the minimal power that can be avoided by an infinite ternary word, as was shown by Dejean [10].

13

Consider the ternary Thue word $\mathbf{T}$ [39], which is the fixed point of the morphism θ defined by $0 \to 012, 1 \to 02, 2 \to 1$:

$$\mathbf{T} = T_1 T_2 T_3 \cdots = 01202101210201202102012101202101210201210120210 \cdots$$

This word has critical exponent 2, which is not reached, so $\mathbf{T}$ is square-free. Also, $\mathbf{T}$ has two alternative definitions through the Thue-Morse word. The first definition says that for any $i \geq 1$, T_i is the number of zeroes between the i'th and $(i+1)$'th occurrences of 1 in $\mathbf{t}$. The second definition is

$$T_i = \begin{cases} 0, & \text{if } t_{i-1} t_i = 01; \\ 1, & \text{if } t_{i-1} = t_i; \\ 2, & \text{if } t_{i-1} t_i = 10. \end{cases} \tag{11}$$

The definition (11) easily implies a bijection between the length-n factors of $\mathbf{t}$ and length-$(n-1)$ factors of $\mathbf{T}$ for any $n \geq 3$. Hence, $p_{\mathbf{T}}(n) = p_{\mathbf{t}}(n+1)$ for all $n \geq 2$, and one can use formula (2). In [14], the complexity of $\mathbf{T}$ was computed directly from the morphism θ.

Conjecture 20. The ternary Thue word $\mathbf{T}$ has the minimum subword complexity over all square-free ternary infinite words.

The above conjecture is supported by the following result, showing that $\mathbf{T}$ is the only candidate for a square-free word of minimum complexity.

Theorem 21. *If a word $\mathbf{U}$ has minimum subword complexity over all square-free ternary infinite words, then $\mathrm{Fac}(\mathbf{U}) = \zeta(\mathrm{Fac}(\mathbf{T}))$, where ζ is a bijective coding.*

Before proving this theorem and presenting further results, we need to recall an encoding technique introduced in [31] by the second author as a development of a particular case of Pansiot's encoding [23]. In what follows, a, b, c are unspecified pairwise distinct letters from Σ_3. Ternary square-free words contain three-letter factors of the form aba, called *jumps* (of one letter over another). Jumps occur quite often: if a square-free word $\mathbf{u}$ has a jump aba at position i, then the next jump in u occurs at one of the positions $i+2$ ($\mathbf{u} = \cdots abaca \cdots$), $i+3$ ($\mathbf{u} = \cdots abacbc \cdots$), or $i+4$ ($\mathbf{u} = \cdots abacbab \cdots$). Note that a jump at position $i+1$ would mean that $\mathbf{u}$ has the square $abab$ at position i, while no jump up to position $i+5$ would lead to the square $bacbac$ at position $i+1$. Also note that a jump in a square-free word can be uniquely reconstructed from the previous (or the next) jump and the distance between them. Thus,

($\star$) a square-free ternary word $\mathbf{u}$ can be uniquely reconstructed from the following information: the leftmost jump, its position, the sequence of distances between successive jumps, and, for finite words only, the number of positions after the last jump.

The property ($\star$) allows one to encode square-free words by walks in the weighted K_{33} graph shown in Fig. 1. The weight of an edge is the number of positions between the positions of

14

two successive jumps. A square-free word $\mathbf{u}$ is represented by the walk visiting the vertices in the order in which jumps occur when reading $\mathbf{u}$ left to right. If the leftmost jump occurs in $\mathbf{u}$ at position $i > 1$, then we add the edge of length $i-1$ to the beginning of the walk; in this case the walk begins at an edge, not a vertex. A symmetric procedure applies to the end of $\mathbf{u}$ if $\mathbf{u}$ is finite. By $(\star)$, we can omit the vertices (except for the first one), keeping just the weights of edges and marking the "hanging" edges in the beginning and/or the end. Due to symmetry, we can omit even the first vertex, retaining all information about $\mathbf{u}$ up to renaming the letters. The result is a word over $\{1, 2, 3\}$ with two additional bits of information (whether the first/last letters are marked; for infinite words, only one bit is needed). This word is called a *codewalk* of $\mathbf{u}$ and denoted by $\mathsf{cwk}(\mathbf{u})$. For example, here is some prefix of $\mathbf{T}$ (with first letters of jumps written in boldface) and the corresponding prefix of its codewalk (the marked letter is underlined):

$$\mathbf{T} = 01\mathbf{2}02\mathbf{1}01\mathbf{2}1\mathbf{0}201\mathbf{2}021\mathbf{0}201\mathbf{2}101\mathbf{2}021\mathbf{0}121\mathbf{0}201\mathbf{2}101\mathbf{2}021\mathbf{0}20 \cdots$$

$$\mathsf{cwk}(\mathbf{T}) = \underline{2} \quad 2 \quad 1 \quad 2 \quad 3 \quad 3 \quad 2 \quad 1 \quad 2 \quad 2 \quad 1 \quad 2 \quad 2 \quad 1 \quad 2 \quad 3 \qquad \cdots$$
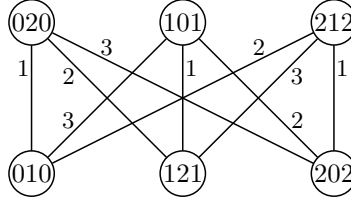

Figure 1: The graph of jumps in ternary square-free words. Vertices are jumps; two jumps that can follow each other in a square-free word are connected by an edge of weight i, where i is the number of positions between the positions of these jumps. Due to symmetry, the graph is undirected.

Note that two words have the same codewalk if and only if they are images of each other under bijective codings and thus have the same structure and the same properties related to power-freeness. A codewalk is *closed* if it corresponds to a closed walk without hanging edges in $K_{3,3}$; e.g., 212212 is closed and 212 is not.

Clearly, not all walks in the weighted K_{33} graph correspond to square-free words. However, there is a strong connection between square-freeness of a word and forbidden factors in its codewalk. Combining several results of [31], we get the following lemma. (More restrictions can be added to statement 2 of this lemma, but we do not need them in our proofs.)

Lemma 22 ([31]).

1. *If a codewalk has (a) no factors $11, 222, 223, 322, 333$, and (b) no factors of the form $vabv$, where $v \in \{1, 2, 3\}^*$, $a, b \in \{1, 2, 3\}$ and the codewalk vab is closed, then the word with this codewalk is square free.*

2. *The codewalk of a square-free word has no proper factors* 11 *and* vav *for all* $v \in \{1,2,3\}^*$, $a, b \in \{1,2,3\}$ *such that* va *is a closed codewalk. In particular, such a codewalk contains no squares of closed codewalks.*

Proof of Theorem 21. Thue [39] showed that a square-free infinite ternary word contains all six factors of the form ab and all six factors of the form abc. As for the jumps, Thue proved that any two factors *from different parts of the* $K_{3,3}$ *graph in Fig.* 1 can absent. (This is an optimal result since it is easy to see that at least two jumps from each part must be present.) All three possible cases (up to symmetry) with two absent jumps are depicted in Fig. 2. We say that a square-free ternary word is of type i, $i \in \{1,2,3\}$, if it lacks two jumps connected by an edge of weight i. Note that **T** has no factors 010 and 212 and thus is of type 3.

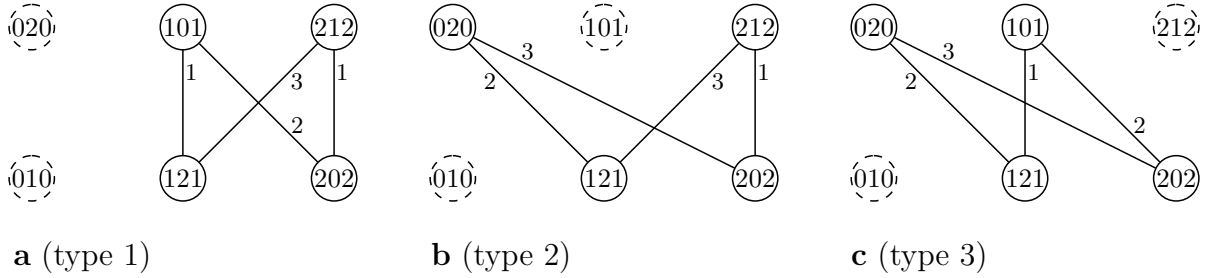**a** (type 1)   **b** (type 2)   **c** (type 3)

Figure 2: Avoidance of two jumps in ternary square-free words. "Type" is the weight of the edge between the avoided jumps.

Assume that a square-free infinite ternary word **U** has the minimum subword complexity among all such words. Then **U** avoids two jumps, otherwise $p_{\mathbf{U}}(3) > p_{\mathbf{T}}(3)$. W.l.o.g., the two missing jumps are those indicated in Fig. 2, depending on the type of **U** (if this is not the case, we replace **U** with its image under an appropriate bijective coding). Note that all four remaining jumps occur in **U** infinitely often, since any suffix of **U** is square-free and thus must contain four jumps. Hence **U** has eight factors of length 4, containing a jump. Let us compute $p_{\mathbf{U}}(4)$. For this, we need to consider the factors without jumps. First let **U** have type 1. 1021 and 2012 are factors of **U** because they are the only right extensions of 102 and 201, respectively. If the factor 0120 is absent, then it is impossible to move by the edge of weight 2 from the vertex 101 to the vertex 202; since the codewalk of **U** cannot contain 11 or 333 by Lemma 22 this means that the codewalk of **U** always passes the cycle in Fig. 2a in the same direction. But this means that the codewalk of **U** contains squares of closed walks, which is also impossible by Lemma 22. So the word 0120 must be a factor of **U**. The same argument works for the factors 0210, 1201, 2102, which are responsible, respectively, for the moves from 202 to 101; from 212 to 121; and from 121 to 212. Thus $p_{\mathbf{U}}(4) = 14$.

If **U** has type 2, the same argument works. Namely, 2012 and 2102 occur in **U** as unique extensions of 201 and 210, respectively. The remaining four factors 0120, 0210, 1021, and 2012 are responsible for the moves by the edges of weight 3; if one of them is absent, this leads to squares of closed walks, thus violating of Lemma 22. Hence we again have $p_{\mathbf{U}}(4) = 14$. The situation changes if **U** has type 3: then each of the words 1021, 1201 can occur in **U**

16

only as the prefix (1021 is followed by 0, and thus can be preceded neither by 0 nor by 2; similar for 1201). The remaining factors must present by the same argument as in the previous cases. Note that $\mathbf{T}$ avoids 1021 and 1201, and thus $p_{\mathbf{T}}(4) = 12$. So $\mathbf{U}$ must have type 3 and $p_{\mathbf{U}}(4) = 12$. (Moreover, $\mathbf{T}$ and $\mathbf{U}$ have the same factors up to length 4.) To prove the theorem, it is enough to show that $\mathrm{Fac}(\mathbf{T}) \subseteq \mathrm{Fac}(\mathbf{U})$; the equality then follows by minimality of complexity of $\mathbf{U}$.

Consider the language C of all finite codewalks that can be read in the graph in Fig. 2c and correspond to square-free words. We define two sequences of codewalks by induction:

$$A_0 = 212, B_0 = 3,$$
$$A_{i+1} = B_i B_i A_i A_i A_i, B_{i+1} = B_i B_i A_i.$$

Below we prove the following three statements, which immediately imply the desired inclusion $\mathrm{Fac}(\mathbf{T}) \subseteq \mathrm{Fac}(\mathbf{U})$. We include in the preimage $\mathsf{cwk}^{-1}(X)$ only the words of type 3 avoiding the factors 010 and 212, as in Fig. 2c.

(i) $\mathrm{Fac}(\mathbf{U}) \supseteq \mathsf{cwk}^{-1}(A_i)$ for every $i \geq 0$;

(ii) $\mathsf{ext}(C) = \bigcup_{i \geq 0} \mathrm{Fac}(A_i)$;

(iii) $\mathrm{Fac}(\mathbf{T}) \subseteq \mathsf{cwk}^{-1}(\mathsf{ext}(C))$.

(i) Note that the codewalks $A_i A_i$, $A_i B_i$, and $B_i B_i$ are closed for all i. This fact immediately follows by induction from the definition (for the base case, one may consult Fig. 2c).

We prove by induction that for each i some suffix of $\mathsf{cwk}(U)$ is an infinite product of blocks A_i and B_i. For the base case note that $\mathsf{cwk}(U)$ has no factors 11 by Lemma 22. Hence, as the codewalk reaches one of the vertices 020, 202 (see Fig. 2c) it infinitely proceeds between these two vertices with the paths labeled by A_0 and B_0; each path occurs infinitely many times because $\mathsf{cwk}(U)$ is aperiodic. The walks 212212, 2123, and 33 are obviously closed.

Now we proceed with the inductive step. First let $i = 1$. The factors $A_0 B_0 A_0$ and $B_0 B_0 B_0$ does not appear in the codewalks of square-free words by Lemma 22. Hence, B_0's always occur in pairs. Further, the factor $B_0 A_0 A_0 B_0$ cannot appear far from the beginning of $\mathsf{cwk}(U)$, because otherwise it will be uniquely extended to $A_0 B_0 B_0 A_0 A_0 B_0 B_0 A_0$, which is the square of a closed codewalk, impossible by Lemma 22. Observing that the factor $A_0 A_0 A_0 A_0$ is also forbidden as the square of a closed codewalk, we see that there may be either one or three consecutive blocks A_0. Hence some suffix of $\mathsf{cwk}(U)$ can be partitioned into the blocks $B_0 B_0 A_0 A_0 A_0 = A_1$ and $B_0 B_0 A_0 = B_1$. As in the base case, both blocks must appear infinitely often to prevent periodicity. For the general case $i > 1$ the argument is essentially the same. The only difference is in proving the fact that $A_i B_i A_i$ and $B_i B_i B_i$ are forbidden: now B_i is a prefix of A_i by construction, so these two codewalks are extended to the right by B_i, which gives us the square of the closed codewalk $A_i B_i$ (resp., $B_i B_i$). The inductive step is finished.

Thus we know that $\mathsf{cwk}(U)$ contains A_i (and even $A_i A_i$) for any i. Clearly, A_i is not closed, so it corresponds to a walk from the vertex 020 to 202 or vice versa. Hence $\mathsf{cwk}^{-1}(A_i)$

consists of two words, each one corresponds to a walk in one direction. But A_i's in the factor A_iA_i correspond to walks in opposite directions (the codewalk A_iA_i is closed), so both words from $\mathsf{cwk}^{-1}(A_i)$ are factors of $\mathbf{U}$. Statement (i) is proved.

(ii) As shown in the proof of (i), each A_i occurs infinitely often in $\mathsf{cwk}(\mathbf{U})$. Hence $\mathsf{ext}(C) \supseteq \bigcup_{i \geq 0} \mathrm{Fac}(A_i)$. For the reverse inclusion, first note the following property implied by the proof of (i). There is a function $f(n)$ such that for every $\mathbf{U}$ of type 3 its suffix, equal to the product of the blocks A_n and B_n, starts before the position $f(n)$. Let $V \in \mathsf{ext}(C)$ and let n be such that $|V| \leq |B_n|$. By definition of two-sided extendable word, there exists a square-free infinite word $\mathbf{U}$ of type 3 such that $\mathsf{cwk}(\mathbf{U})$ contains V at position greater than $f(n)$. Then V is a factor of one of the codewalks A_nA_n, A_nB_n, B_nA_n, B_nB_n. In each case, V is a factor of A_{n+2}, and we have the desired inclusion $\mathsf{ext}(C) \subseteq \bigcup_{i \geq 0} \mathrm{Fac}(A_i)$.

(iii) Since T is recurrent, its codewalk is recurrent as well, implying $\mathrm{Fac}(\mathsf{cwk}(\mathbf{T})) \subseteq \mathsf{ext}(C)$. The result now follows. □

For symmetric words, the square-free ternary infinite word of minimum subword complexity does exist. Recall that the Fibonacci word $\mathbf{f}$ is the fixed point of the binary morphism defined by $\phi(0) = 01, \phi(1) = 0$. We define the coding $\xi : (0 \to 2, 1 \to 1)$ and write $\mathbf{f}_{12} = \xi(\mathbf{f})$. Now consider the *1-2-bonacci word* $\mathbf{F}_{12} \in \Sigma_3^\omega$, which is the word beginning with 01 and having the codewalk $\mathbf{f}_{12}$. This word was introduced by Petrova [25], who proved that $\mathbf{F}_{12}$ has critical exponent 11/6 (reachable) and no length-5 factors of the form $abcab$. Also, $\mathbf{F}_{12}$ appeared to have a nice extremal property [15, Proposition 13].

Theorem 23. *The 1-2-bonacci word $\mathbf{F}_{12}$ has the minimum subword complexity over all symmetric square-free ternary infinite words. This complexity equals $6n - 6$ for all $n \geq 2$.*

Proof. Let $\mathbf{u}$ be a symmetric square-free ternary infinite word. Since $\mathbf{u}$ is aperiodic, it has a special factor of length n for each $n \geq 0$. (In the ternary case, a word v is called a $\mathbf{u}$-special factor if at least two of the words $v0, v1, v2$ are factors of $\mathbf{u}$.) If v is $\mathbf{u}$-special, then the word $\zeta(v)$, where ζ is any bijective coding, is $\mathbf{u}$-special as well, because of the symmetry of $\mathbf{u}$. Thus, there are at least six $\mathbf{u}$-special factors of length n for each $n \geq 2$. Together with the fact $p_\mathbf{u}(2) = 6$ mentioned above, this gives the lower bound for the complexity of $\mathbf{u}$: $p_\mathbf{u}(n) \geq 6n - 6$ for all $n \geq 2$. So we are going to prove that the 1-2-bonacci word is symmetric and its complexity matches this lower bound.

The codewalk $\mathbf{f}_{12}$ of the 1-2-bonacci word $\mathbf{F}_{12}$ has no 3's and thus corresponds to a walk in the subgraph of the K_{33} graph (see Fig. 3). By definition of $\mathbf{F}_{12}$, this walk begins at the vertex 010. Let v be a factor of $\mathbf{f}_{12}$. Let us write $f_i = \xi(\phi^i(0))$ for all $i \geq 0$. Then there is i such that $f_i = uvw$ for some words u and w. Note that f_i occurs in $\mathbf{f}_{12}$ infinitely often, in particular, as a prefix and after each prefix f_{i+k}, where $k > 0$.

We call two codewalks *equivalent* (and write $u \sim v$) if the corresponding walks, beginning in the same vertex, end in the same vertex. For example, $21221 \sim 2$, because the walk 1221
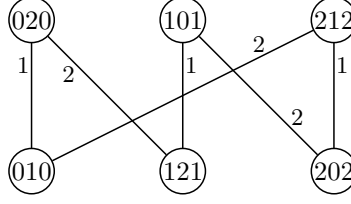
Figure 3: The graph of jumps in the 1-2-bonacci word.

is closed. Similar to [25], we observe that

$$
f_i \sim \begin{cases}
21, & \text{if } i \bmod 6 = 1; \\
212, & \text{if } i \bmod 6 = 2; \\
2, & \text{if } i \bmod 6 = 0 \text{ or } 3; \\
12, & \text{if } i \bmod 6 = 4; \\
1, & \text{if } i \bmod 6 = 5.
\end{cases}
$$

We note that the paths from the vertex 010, labeled by 21, 212, 2, 12, and 1, end in all five remaining vertices. Thus, when reading the codewalk of F_{12}, we read f_i starting from each vertex. Reading the word u from all vertices generates a bijection on the set of vertices; thus, we read v from each vertex. Therefore, $\mathbf{F}_{12}$ contains all six factors with the codewalk v. Since v is arbitrary, this means that $\mathbf{F}_{12}$ is symmetric.

Finally we compute the subword complexity of $\mathbf{F}_{12}$. Consider a $\mathbf{F}_{12}$-special factor V such that $|V| \geq 5$ and $Va, Vb \in \mathrm{Fac}(\mathbf{F}_{12})$. The codewalk of V has the form $xv\underline{1}$ for some $v \in \mathrm{Fac}(\mathbf{f}_{12})\backslash\{\varepsilon\}$, $x \in \{\varepsilon, \underline{1}, \underline{2}\}$. The factors Va and Vb of $\mathbf{F}_{12}$ have codewalks $xv1$ and $xv2$. Hence both $v1$ and $v2$ are factors of $\mathbf{f}_{12}$, so v is $\mathbf{f}_{12}$-special. If U is another $\mathbf{F}_{12}$-special factor of length $|V|$, then, similarly, its codewalk is of the form $yu\underline{1}$ for some $\mathbf{f}_{12}$-special factor u and $y \in \{\varepsilon, \underline{1}, \underline{2}\}$. Since $\mathbf{f}_{12}$ is a Sturmian word, it has only one special factor of each length. But every suffix of a special factor is special, so w.l.o.g. u is a suffix of v.

Let $v = v'u$ and assume v' nonempty. The words U and V have suffixes of equal length encoded by $u\underline{1}$. Note that v' encodes at least two letters of V (just two if $v' = 1$), while y encodes at most two letters of U (just two if $y = \underline{2}$). But $\mathbf{f}_{12} \in \{1,2\}^\omega$, so if $y = \underline{2}$ then $2u$ is $\mathbf{f}_{12}$-special and thus a suffix of v. Hence in this case the last letter in v' is 2, implying that v' encodes at least three letters of V. Therefore, $|V| > |U|$ in all cases, contradicting the choice of U. Then $v' = \varepsilon$ and $u = v$. Thus we proved that two $\mathbf{F}_{12}$-special factors of the same length have the same codewalk. Due to symmetry, this means that $\mathbf{F}_{12}$ has exactly six special factors of every length $n \geq 5$. Computing $p_{\mathbf{F}_{12}}(2) = 6$, $p_{\mathbf{F}_{12}}(3) = 12$, $p_{\mathbf{F}_{12}}(4) = 18$, $p_{\mathbf{F}_{12}}(5) = 24$, we then have $p_{\mathbf{F}_{12}}(n) = 6n - 6$ for all $n \geq 2$, as desired. $\qquad\square$

Remark 24. As demonstrated in the proof of Theorem 23, 6 is the minimal linear growth constant of an aperiodic symmetric infinite ternary word. However, there exist such words with linear growth constant 6 and critical exponent smaller than 11/6. An example of

19

such word is the 1-3-bonacci word $\mathbf{F}_{13}$ obtained similar to the 1-2-bonacci word: take the Fibonacci word $\mathbf{f}$, replace all 0's with 3's to get the codewalk $\mathbf{f}_{13}$ and take the word with this codewalk as $\mathbf{F}_{13}$. The critical exponent of $\mathbf{F}_{13}$ is $\frac{5+\sqrt{5}}{4} \doteq 1.8090\cdots$ [25]; the fact that $\mathbf{F}_{13}$ is symmetric and the equality $p_{F_{13}}(n) = 6n$ for all $n \geq 5$ can be proved as in Theorem 23.

Remark 24 suggests the following question.

Open Question 25. *What is the minimal critical exponent of a symmetric infinite ternary word with linear growth constant 6?*

As to the case of $(\frac{7}{4})^+$-power-free words, our knowledge is quite limited. The two well-known such words are the Arshon word [3] and the Dejean word [10]. Both are symmetric and recurrent; the codewalk of the Arshon word is the concatenation of some finite prefix and a morphic image of the fixed point of the morphism η defined by $\eta(0) = 010, \eta(1) = 011$ (see [24, Lemma 2]). This fixed point has subword complexity $2n$ for every $n \geq 1$. Thus, the Arshon word can be easily proved, similar to the proof of Theorem 23, to have linear growth constant 12. We have not computed such a constant for the Dejean word, but there is a numerical evidence that it is not smaller than 12. Thus if the answer to Open Question 25 is greater than $7/4$, which looks plausible, then the linear growth constant of the Arshon word is the minimum possible. Two further questions are natural.

Open Question 26. *Is there a symmetric $(\frac{7}{4})^+$-power-free infinite ternary word of minimum subword complexity? If yes, is the Arshon word an example of such word?*

Open Question 27. *What is the minimal linear growth constant of a $(\frac{7}{4})^+$-power-free infinite ternary word?*

3.2   Other ternary words

When the squares in ternary words are allowed, we can build words of smaller complexity. Here we consider the special case of words with the complexity upper bounded by the function $2n + 1$. Let

$$\mathbf{G} = 012020102012010201202\cdots \tag{12}$$

be the fixed point of the morphism γ defined by $0 \to 01, 1 \to 2, 2 \to 02$. The word $\mathbf{G}$ is a recoding of the sequence A287104 from Sloane's *Encyclopedia*.

Lemma 28. $p_{\mathbf{G}}(n) = 2n + 1$ *for all $n \geq 0$.*

Proof. We use standard techniques (see, e.g., [7]), so we try to keep the proof short. It is sufficient to prove that there are exactly two $\mathbf{G}$-special words of length n for all $n \geq 1$. For small n, one can check by exhaustive search that there is a unique special word ending in 0 and a unique such word ending in 1 (a $\mathbf{G}$-special word cannot end in 2 because $21, 22 \notin$

Fac(**G**)). Let $V0$ be special; then $V01, V02 \in$ Fac(**G**), implying that $\gamma(V)012, \gamma(V)0102 \in$ Fac(**G**) and thus $\gamma(V)01$ is special. Similarly, if $U1$ is special, then $U10, U12 \in$ Fac(**G**); $\gamma(U)201, \gamma(V)202 \in$ Fac(**G**) and thus $\gamma(U)20$ is special. Since each suffix of a special word is special, there are special words of every length ending in 0 and in 1. Now assume that the lemma is false; then for some n one has $D_{\mathbf{G}}(n) > 2$, $D_{\mathbf{G}}(1) = \cdots = D_{\mathbf{G}}(n-1) = 2$.

Some case analysis is needed; all cases are similar, so we consider one of them. Assume that two special words of length n end with 0. Since their suffixes are special, and only one special word of length $n-1$ ends with 0, these two words are $aV0$ and $bV0$, where $a, b \in \Sigma_3$. Let $a = 0, b = 1$ (the other case is $a = 1, b = 2$). Then we can write $V = 2V'$. We have

$$02V'01, \ 02V'02, \ 12V'01, \ 12V'02 \in \text{Fac}(\mathbf{G}) \quad \text{and hence}$$
$$2\gamma^{-1}(V')0, \ 2\gamma^{-1}(V')2, \ 1\gamma^{-1}(V')0, \ 1\gamma^{-1}(V')2 \in \text{Fac}(\mathbf{G})$$

Then $2\gamma^{-1}(V'), 1\gamma^{-1}(V')$ are two special words of length $< n$, ending with the same letter 1; this is impossible by the choice of n. Studying all cases in the same way, we reach the same contradiction. Thus the lemma holds. $\qquad\square$

Theorem 29. *The critical exponent of the word* **G** *is* $2 + \frac{1}{\lambda^2 - 1} = 2.4808627\cdots$, *where* $\lambda = 1.7548777\cdots$ *is the real zero of the polynomial* $x^3 - 2x^2 + x - 1$.

Proof. The critical exponent of **G** can be computed by Krieger's method [19]. We recall the necessary tools suitable for analyzing **G** specifically, rather than in full generality.

For a word $w \in \Sigma_k^*$, we let $|w|_a$ denote the number of occurrences of the letter a in w. The *Parikh vector* of w is the vector $\vec{P}(w) = (|w|_0, \ldots, |w|_{k-1})$. By *norm* of a vector we mean the sum of its coordinates; so $\|\vec{P}(w)\| = |w|$. If w is a prefix of x^ω for some word x, we say that w has *period* $|x|$. In this case, all factors of w of length $|x|$ share the same Parikh vector $\vec{P}(x)$, so we can speak about "Parikh vector of the period". If $|x|$ is the minimal period of w, we call x the *root* of w. The exponent of w then can be written as

$$\exp(w) = \frac{|w|}{|x|} = \frac{\|\vec{P}(w)\|}{\|\vec{P}(x)\|}.$$

The matrix A_f of a morphism $f : \Sigma_k^* \to \Sigma_m^*$ is a nonnegative integer $k \times m$ matrix, the i'th row of which is the Parikh vector of $f(i-1)$, where $i = 1, \ldots, k$. For example, the morphism γ has the matrix

$$A = A_\gamma = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

One has $\vec{P}(f(w)) = \vec{P}(w) \cdot A_f$. Note that the characteristic polynomial of A is $x^3 - 2x^2 + x - 1$, so the maximal (and unique) real eigenvalue of A is λ.

A *run* in a finite or infinite word w is an occurrence of a factor v of w such that (a) $\exp(v) \geq 2$ and (b) this occurrence cannot be extended in w to a longer factor with the same

minimal period. For example, the word **G** has run 2020 at position 2 with period 2, run 02010201 at position 3 with period 4, and run 2010201201020120 at position 4 with period 7; see (12). For **G**, as for any word containing squares, the critical exponent equals the supremum of exponents of its runs. Since **G** is a fixed point of a morphism, its runs can be grouped into infinite series in the following way:

- take a run V at position i with root X (let $\mathbf{G} = UV \cdots$, $|U| = i$);

- take the occurrence of $\gamma(V)$ at position $|\gamma(U)|$ and extend it to a run with period $|\gamma(X)|$;

- take the obtained run as V and repeat.

For example, the runs mentioned above form the beginning of a series:

$$
\begin{array}{llll}
2020 & \rightarrow & 02010201 & \rightarrow & \mathbf{2}010201201020120 & \rightarrow & \cdots \\
\text{at } 2 & & \text{at } 3 & & \text{at } 4 & & \\
U = 01 & & \gamma(U) = 012 & & \gamma^2(U) = 01202 & & \\
X = 20 & & \gamma(X) = 0201 & & \gamma^2(X) = 0102012 & & \\
& & \text{no extensions} & & \text{extended left by } 2, \text{right by } 0 & &
\end{array}
$$

In the case of **G**, the left (resp., right) extension is a longest common suffix (resp., prefix) of γ-images of corresponding letters. Thus, the left extension is either 2 (the common suffix of $\gamma(1)$ and $\gamma(2)$) or ε, and the right extension is either 0 (the common prefix of $\gamma(0)$ and $\gamma(2)$) or ε. It is possible to compute the exponents of any run in a series in a uniform way. Namely, if one has a series $\{V_m\}_0^\infty$ such that V_0 has root X, then the exponent of each V_m can be computed by the following formula:

$$
\exp(V_m) = \frac{\|\vec{P}(V_0) \cdot A^m + \sum_{i=1}^m \vec{P_i} \cdot A^{m-i}\|}{\|\vec{P}(X) \cdot A^m\|} \,, \tag{13}
$$

where $\vec{P_i}$ is the sum of the Parikh vectors of the words which were added to $\gamma(V_{i-1})$ to get V_i. Consider the series introduced above, with $V_0 = 2020$. From definition of γ it is easy to find that $\vec{P_i} = (0,0,0)$ for odd i and $\vec{P_i} = (1,0,1)$ for even i. Observing that $\vec{P}(X) = (1,0,1)$, $\vec{P}(V_0) = 2\vec{P}(x)$, we simplify (13) to get

$$
\exp(V_m) = 2 + \frac{\|(1,0,1) \cdot \left(A^{m-2} + A^{m-4} + \cdots + A^{m-2\lfloor m/2 \rfloor}\right)\|}{\|(1,0,1) \cdot A^m\|} \,. \tag{14}
$$

One can check that the sequence $\{\exp(V_m)\}_0^\infty$ is strictly increasing, so that its supremum equals its limit. The limit can be computed by standard machinery of Perron-Frobenius theory. We describe the idea, omitting the plain calculus. For large m, the vector $(1,0,1) \cdot A^m$ is very close to the eigenvector of the matrix A corresponding to its maximal eigenvalue λ; so the multiplication of this vector by A corresponds, up to a small error, to its multiplication

22

by λ. Next note that if we multiply the matrix in the numerator of (14) by $(A^2 - I)$, where I is the identity matrix, we obtain either $A^m - I$ or $A^m - A$, depending on the parity of m. Hence, as $m \to \infty$, the numerator multiplied by $(\lambda^2 - 1)$ approaches the denominator. Therefore,

$$\lim_{m \to \infty} \exp(V_m) = 2 + \frac{1}{\lambda^2 - 1},$$

as required. In the same way, we consider the series starting with the run 201201 at position 8. For this run, $\vec{P}_i = (1, 0, 0)$ for odd i and $\vec{P}_i = (0, 0, 1)$ for even i. Thus we get the following equality instead of (14):

$$\exp(V_m) = 2 + \frac{\|(1,0,0) \cdot \left(A^{m-1} + A^{m-3} + \cdots\right) + (0,0,1) \cdot \left(A^{m-2} + A^{m-4} + \cdots\right)\|}{\|(1,1,1) \cdot A^m\|}. \quad (15)$$

Observing that $(1, 0, 0)A + (0, 0, 1) = (1, 1, 1)$, we can reduce (15) to (14). (The particular nonnegative vector does not matter; what matters is that the vector in the numerator and the denominator is the same.) So we get the same limit as above.

For all other series of runs, the initial run is long enough to ensure that the vector $\sum_{i=1}^{m} \vec{P}_i \cdot A^{m-i}$ has much smaller norm than $\vec{P}(X) \cdot A^m$. The theorem is proved. $\qquad \square$

Conjecture 30. Among all ternary words with subword complexity bounded above by $2n + 1$, the word **G** has the lowest possible critical exponent.

3.3 Binary words

When we switch from $(\frac{7}{3})$-power-free to $(\frac{7}{3})^+$-power-free infinite binary words, the Thue–Morse word loses its status as the word of minimum complexity. Let us give an example of a $(\frac{7}{3})^+$-power-free infinite word with subword complexity incomparable to $p_{\mathbf{t}}(n)$. Consider the morphism $g : \Sigma_3 \to \Sigma_2$ defined by the rules

$$0 \to 01100100110\,1001\,0110\,1001\,1001,$$
$$1 \to 01100100110\,1001\,0110\,0110\,1001,$$
$$2 \to 01100100110\,1001\,1001\,0110\,1001.$$

It maps square-free ternary words to $(\frac{7}{3})^+$-power-free binary words (see [34, Section 3]).
Theorem 31. *Let* $\mathbf{g} = g(\mathbf{T})$, *where the morphism* g *and the ternary Thue word* $\mathbf{T}$ *are defined above. Then* $p_{\mathbf{g}}(n) < p_{\mathbf{t}}(n)$ *for infinitely many values of* n.

Proof. Let $m = 2^k$ for some $k \geq 2$ and compare the number of factors of length $n = 27m - 7$ in $\mathbf{g}$ and $\mathbf{t}$. Note that images of letters under g differ only by the factor of length 8 at position 15. Hence a length-n factor of $\mathbf{g}$ contains exactly m such "identifying" factors (one of them, possibly, only partially). Thus every length-n factor of $\mathbf{g}$ is uniquely indentified

by its g-preimage of length m and its ininial position inside the g-image of a letter. So $p_{\mathbf{g}}(n) \leq 27 \cdot p_{\mathbf{T}}(m) = 27 \cdot 3m = 81 \cdot 2^k$. By (2) we have $p_{\mathbf{t}}(n) = 2(n-1) + 2^{k+5} = 86 \cdot 2^k - 16$. Since $5 \cdot 2^k > 16$, we obtain $p_{\mathbf{g}}(n) < p_{\mathbf{t}}(n)$. $\qquad\square$

Open Question 32. *What is the minimun value of α such that some α-power-free infinite binary word $\mathbf{u}$ has smaller complexity than the Thue-Morse word? (Recall that "smaller" means $p_{\mathbf{u}}(n) \leq p_{\mathbf{t}}(n)$ for all n and $p_{\mathbf{u}}(n) < p_{\mathbf{t}}(n)$ for some n.)*

It is known [6] that the critical exponent of a Sturmian word is at least $(5 + \sqrt{5})/2 \doteq 3.61803 \cdots$, and the minimum is reached by the Fibonacci word $\mathbf{f}$ defined above.

Remark 33. The words of subword complexity $n + c$ for any integer constant $c > 1$ were characterized by Cassaigne [8, Proposition 8] as having the form $uf(\mathbf{v})$, where u is a finite word, $\mathbf{v}$ is a Sturmian word and f is a morphism. So it is not hard to prove that these words have the same minimum critical exponent $(5 + \sqrt{5})/2$.

So, for the critical exponents smaller than $(5 + \sqrt{5})/2$ we look at the infinite words with linear growth constant 2. As Theorem 35 below shows, such words can have the critical exponent as small as $5/2$. This gives an upper bound for the value of α in Open Question 32.

Remark 34. By backtracking, one can prove that the longest binary words avoiding $5/2$-powers and with subword complexity $\leq 2n$ are of length 38. They are

$$00110011010011001001101001100100110010,$$
$$00110011010011001001101001100100110011,$$

and their reversals and complements.

Theorem 35. *Let $\tau : \Sigma_3 \to \Sigma_2$ be the morphism defined by $0 \to 0$, $1 \to 01$, $2 \to 011$ and $\mathbf{G}$ be the word defined in Section . Then*

$$\tau(\mathbf{G}) = 0010110011001001100101100100110010110011001 \cdots$$

has the lowest critical exponent among all binary words with subword complexity $\leq 2n$. It (i) avoids $(\frac{5}{2})^+$-powers and (ii) has subword complexity exactly $2n$ for all $n > 0$.

Proof. In the proof we refer to the properties of the word $\mathbf{G}$ (Lemma 28 and Theorem 29) and their proofs. For (ii), we check by hand that for small n the word $\tau(\mathbf{G})$ has two special words, one ending with 0 and the other ending with 1. As in the proof of Lemma 28, we then assume that statement (ii) is false, choose n such that $D_{\tau(\mathbf{G})}(n) > 2$, $D_{\tau(\mathbf{G})}(n-1) = \cdots = D_{\tau(\mathbf{G})}(1) = 2$, and do case analysis. There are two $\tau(\mathbf{G})$-special length-n words ending with the same letter; these words have common suffix of length $n - 1$ by the choice of n. Consider the case where this suffix equals $0v0$ for some $v \in \mathrm{Fac}(\tau(\mathbf{G}))$. From the condition

$$00v00, 00v01, 10v00, 10v01 \in \mathrm{Fac}(\tau(\mathbf{G}))$$

24

we conclude that $0v = \tau(V)$ for some $V \in \text{Fac}(\mathbf{G})$, and all words $0V0, 0V2, 1V0, 1V2$ are factors of $\mathbf{G}$ (the second conclusion uses the fact that $\mathbf{G}$ has no factor 21). Then both $0V$ and $1V$ are $\mathbf{G}$-special, contradicting the fact, shown in the proof of Lemma 28, that $\mathbf{G}$ has at most one special word of a given length ending with a given letter. The cases where the common suffix has the form $0v1, 1v0$, or $1v1$ are similar and imply the same contradiction. Hence $D_{\tau(\mathbf{G})}(n) = 2$ for all $n \geq 1$, and then $p_{\tau(\mathbf{G})}(n) = 2n$.

For (i), we check that $\tau(\mathbf{G})$ contains the $(\frac{5}{2})$-power 0110011001, and all other runs with small periods have smaller exponents. For large periods, we should check the exponents of runs which are obtained as images of runs from two series described in the proof of Theorem 29. The images of runs in $\mathbf{G}$ are extended by runs in $\tau(\mathbf{G})$ by 1 or 2 letters on the right (τ-images of all letters begin with 0, τ-images of 1 and 2 begin with 01) and by 0 or 1 letters two the left (τ-images of 1 and 2 end with 1). For the series of runs in $\mathbf{G}$ starting with 2020 exactly two letters are added to each $\tau(V_m)$ to make a run in $\tau(G)$. So if we write v_m for the run in $\tau(\mathbf{G})$ obtained from the τ-image of V_m and B for the matrix of τ, we get from (14)

$$\exp(v_m) = 2 + \frac{\|(1,0,1) \cdot \left(A^{m-2} + A^{m-4} + \cdots + A^{m-2\lfloor m/2 \rfloor}\right) \cdot B\| + 2}{\|(1,0,1) \cdot A^m \cdot B\|}.$$

It is not hard to check that $\exp(v_0) = 5/2$ and $\exp(v_m) < 5/2$ for all $m > 0$. Similar computation can be performed for the series of runs in $\mathbf{G}$ starting with 201201; here $\exp(v_m) < 5/2$ for all m. As a result, we conclude that $\tau(\mathbf{G})$ has critical exponent $5/2$, as required. $\qquad \square$

4 Large subword complexity in big languages

If a language $L_{k,\alpha}$ has an exponential growth function, then it seems quite natural that there would be infinite α-power-free words over Σ_k having exponential subword complexity. For example, Currie and Rampersad [9, Prop. 9] gave an example of a squarefree word over Σ_3 having exponential subword complexity.

Additional examples of such words can be provided using some standard techniques. Below we give the examples for the minimal binary and minimal ternary power-free languages of exponential growth.

Theorem 36.

(a) There is an infinite binary $(\frac{7}{3})^+$-power-free word having exponential subword complexity.

(b) There is an infinite ternary $(\frac{7}{4})^+$-power-free word having exponential subword complexity.

Proof. (a) First, create an infinite square-free word over Σ_4 with exponential subword complexity. For this, take an infinite square-free word $\mathbf{u}$ over Σ_3 and an infinite word $\mathbf{v}$ of exponential complexity over $\{2, 3\}$. For each $i \geq 1$, replace the i'th occurrence of symbol 2

in $\mathbf{u}$ with the i'th symbol of $\mathbf{v}$. The resulting word obviously satisfies the desired properties. Now apply the 21-uniform morphism $h : \Sigma_4^* \to \Sigma_2^*$ from [16, Lemma 8]. The lemma guarantees that the image of a square-free word is $(7/3)^+$-power-free, and every uniform injective morphism preserves the property of having exponential subword complexity.

(b) Start with an infinite $(7/5)^+$-power-free word over Σ_4, which exists by Pansiot's result [23]. As in (a), replace the occurrences of 3 in this word by an infinite word over $\{3, 4\}$ with exponential subword complexity, getting an infinite $(7/5)^+$-power-free word over Σ_5 with exponential subword complexity. Now apply the morphism of Ochem [22, Theorem 4.2]. The result is guaranteed to be $(7/4)^+$-free and to have exponential subword complexity. $\square$

As usual, in such examples the growth rate of subword complexity is barely above 1. Are there words having the subword complexity comparable to the growth function of the whole language? It turns out that this problem is closely related to an old problem by Restivo and Salemi. In [27] they posed the following problem: *given two square-free words $u, v \in \Sigma_3^*$, provide an algorithm deciding whether there is a word $w \in \Sigma_3^*$ such that uwv is square-free.* The problem can be generalized to any language $L_{k,\alpha}$, and is still open except for the case of small binary languages (due to the existence of factorizations of type (1), it is easy to connect any right extendable u to any left extendable v by an appropriate Thue-Morse factor).

Clearly, the interesting part of the problem is formed by the case where u is right extendable and v is left extendable. To the best of our knowledge, there are no known tuples (k, α, u, v) such that u is right extendable in $L_{k,\alpha}$, v is left extendable in $L_{k,\alpha}$, and no word of the form uwv belongs to $L_{k,\alpha}$. For our purposes, we restrict ourselves to the consideration of two-sided extendable words.

We say that a language $L_{k,\alpha}$ has the *Restivo-Salemi property* if for every $u, v \in \mathsf{ext}(L_{k,\alpha})$ there is a word w such that $uwv \in \mathsf{ext}(L_{k,\alpha})$.

Theorem 37. *A power-free language $L_{k,\alpha}$ has the Restivo-Salemi property if and only if all words from $\mathsf{ext}(L_{k,\alpha})$ are factors of some α-power-free infinite recurrent k-ary word $\mathbf{u}$.*

An already mentioned result of [36] says that L and $\mathsf{ext}(L)$ have the same growth rate, so Theorem 37 implies the following.

Corollary 38. *If a power-free language $L_{k,\alpha}$ possesses the Restivo-Salemi property, then there is an α-power-free infinite k-ary word $\mathbf{u}$ with subword complexity having the same growth rate as $L_{k,\alpha}$.*

Proof of Theorem 37. For the forward implication we endow Σ_k^* with the radix order (the words are ordered by length, and the words of equal length are ordered lexicographically) and build the word $\mathbf{u}$ by induction. As the base case, we build the prefix $u_0 = 0$. For the inductive step, assume that the prefix u_n was constructed so far. Let v_n be the smallest in radix order word from $\mathsf{ext}(L_{k,\alpha})$ that is not a factor of u_n. Then we take w_n such that $u_n w_n v_n \in \mathsf{ext}(L_{k,\alpha})$ and put $u_{n+1} = u_n w_n v_n$. The resulting word $\mathbf{u}$ is α-power-free by construction. Further, every word $v \in \mathsf{ext}(L_{k,\alpha})$ is a factor of some u_n and thus of $\mathbf{u}$. Finally, for an arbitrary

$v \in \mathsf{ext}(L_{k,\alpha})$ and every n, there is a word x such that $|x| > |u_n|$ and $xv \in \mathsf{ext}(L_{k,\alpha})$; since $\mathbf{u}$ contains the factor xv, there is an occurrence of v in $\mathbf{u}$ outside the prefix u_n. Hence v occurs in $\mathbf{u}$ infinitely many times. Thus $\mathbf{u}$ is recurrent and we proved this implication.

Now turn to the backward implication. For arbitrary words $u, v \in \mathsf{ext}(L_{k,\alpha})$ each of them occurs in $\mathbf{u}$ infinitely often, so we can find a factor of the form uwv. This factor also occurs in $\mathbf{u}$ infinitely often, allowing us to find arbitrarily long words x, y such that $xuwvy$ is a factor of $\mathbf{u}$. Then $uwv \in \mathsf{ext}(L_{k,\alpha})$. Hence we proved the Restivo-Salemi property for $L_{k,\alpha}$. $\qquad\square$

Remark 39. It is worth mentioning that for small binary languages Theorem 37 works in an extremal form. Since $2^+ \leq \alpha \leq 7/3$ implies $\mathsf{ext}(L_{2,\alpha}) = \mathrm{Fac}(\mathbf{t})$, the language $L_{2,\alpha}$ trivially has the Restivo-Salemi property; as we know from Corollary 3, *all* α-power-free infinite binary words contain all words of $\mathsf{ext}(L_{2,\alpha})$ as factors.

The following conjecture is based on extensive numerical studies.

Conjecture 40. [30, Conjecture 1] All power-free languages satisfy the Restivo-Salemi property.

As an approach to Conjecture 40, we suggest the following.

Open Question 41. *Prove the converse of Corollary 38.*

References

[1] J.-P. Allouche and J. O. Shallit. The ubiquitous Prouhet-Thue-Morse sequence. In C. Ding, T. Helleseth, and H. Niederreiter, editors, *Sequences and Their Applications, Proceedings of SETA '98*, pages 1–16. Springer-Verlag, 1999.

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

[3] S. E. Arshon. Proof of the existence of asymmetric infinite sequences. *Mat. Sbornik* **2** (1937), 769–779. In Russian, with French abstract.

[4] S. V. Avgustinovich. The number of different subwords of given length in the Morse-Hedlund sequence. *Sibirsk. Zh. Issled. Oper.* **1** (1994), 3–7,103. In Russian. English translation in A. D. Korshunov, ed., *Discrete Analysis and Operations Research*, Kluwer, 1996, pp. 1-5.

[5] S. Brlek. Enumeration of factors in the Thue-Morse word. *Disc. Appl. Math.* **24** (1989), 83–96.

[6] A. Carpi and A. de Luca. Special factors, periodicity, and an application to Sturmian words. *Acta Inf.* **36**(12) (2000), 983–1006.

[7] J. Cassaigne. Complexité et facteurs spéciaux. *Bull. Belg. Math. Soc.* **4** (1997), 67–88.

[8] J. Cassaigne. Sequences with grouped factors. In *Proceedings DLT 1997*, pages 211–222. Aristotle University of Thessaloniki, 1997.

[9] J. Currie and N. Rampersad. Cubefree words with many squares. *Discrete Math. & Theoret. Comput. Sci.* **12**(3) (2010), 29–34.

[10] F. Dejean. Sur un théorème de Thue. *J. Combin. Theory. Ser. A* **13** (1972), 90–99.

[11] A. de Luca and S. Varricchio. Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups. *Theoret. Comput. Sci.* **63** (1989), 333–348.

[12] C. F. Du, J. Shallit, and A. M. Shur. Optimal bounds for the similarity density of the Thue-Morse word with overlap-free and (7/3)-power-free infinite binary words. *Int. J. Found. Comput. Sci.* **26**(8) (2015), 1147–1166.

[13] S. Ferenczi. Complexity of sequences and dynamical systems. *Discrete Math.* **206** (1999), 145–154.

[14] A. E. Frid and S. V. Avgustinovich. On bispecial words and subword complexity of D0L sequences. In C. Ding, T. Helleseth, and H. Niederreiter, editors, *Sequences and Their Applications, Proceedings of SETA '98*, pages 191–204. Springer-Verlag, 1999.

[15] D. Gasnikov and A. M. Shur. Ternary square-free partial words with many wildcards. In *Proceedings DLT 2016*, Vol. 9840 of *Lecture Notes in Computer Science*, pages 177–189. Springer, 2016.

[16] J. Karhumäki and J. Shallit. Polynomial versus exponential growth in repetition-free words. *J. Combin. Theory Ser. A* **104** (2004), 335–347.

[17] Y. Kobayashi. Enumeration of irreducible binary words. *Disc. Appl. Math.* **20** (1988), 221–232.

[18] R. Kolpakov and M. Rao. On the number of Dejean words over alphabets of 5, 6, 7, 8, 9 and 10 letters. *Theoret. Comput. Sci.* **412** (2011), 6507–6516.

[19] D. Krieger. On critical exponents in fixed points of non-erasing morphisms. *Theor. Comput. Sci.* **376**(1-2) (2007), 70–88.

[20] H. Mousavi. Automatic theorem proving in Walnut. Preprint, available at `https://arxiv.org/abs/1603.06017`, 2016.

[21] M. Morse and G. A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *Amer. J. Math.* **62** (1940), 1–42.

[22] P. Ochem. A generator of morphisms for infinite words. *RAIRO Theor. Inf. Appl.* **40** (2006), 427–441.

[23] J.-J. Pansiot. A propos d'une conjecture de F. Dejean sur les répétitions dans les mots. *Disc. Appl. Math.* **7** (1984), 297–311.

[24] E. A. Petrova and A. M. Shur. Constructing premaximal ternary square-free words of any level. In *Proceedings MFCS 2012*, Vol. 7464 of *Lecture Notes in Computer Science*, pages 752–763, 2012.

[25] E. A. Petrova. Avoiding letter patterns in ternary square-free words. *Electronic J. Combinatorics* **23**(1) (2016), P1.18.

[26] N. Rampersad and J. Shallit. Repetitions in words. In V. Berthé and M. Rigo, editors, *Combinatorics, Words and Symbolic Dynamics*, Vol. 159 of *Encyc. of Math. and Its Appl.*, pages 101–150. Cambridge University Press, 2016.

[27] A. Restivo and S. Salemi. Some decision results on non-repetitive words. In A. Apostolico and Z. Galil, editors, *Combinatorial algorithms on words*, Vol. F12 of *NATO ASI series*, pages 289–295. Springer-Verlag, 1985.

[28] A. Restivo and S. Salemi. Overlap free words on two symbols. In M. Nivat and D. Perrin, editors, *Automata on Infinite Words*, Vol. 192 of *Lecture Notes in Computer Science*, pages 198–206. Springer-Verlag, 1985.

[29] N. Rampersad, J. Shallit, and A. M. Shur. Fife's theorem for (7/3)-powers. In *Proc. 8th Internat. Conf. Words (WORDS 2011)*, Vol. 63 of *EPTCS*, pages 189–198, 2011.

[30] A. M. Shur. Two-sided bounds for the growth rates of power-free languages. In *Proc. 13th Int. Conf. on Developments in Language Theory. DLT 2009*, Vol. 5583, pages 466–477. Springer, 2009.

[31] A. M. Shur. On ternary square-free circular words. *Electronic J. Combinatorics* **17** (2010), R140.

[32] J. Shallit. Fife's theorem revisited. In G. Mauri and A. Leporati, editors, *DLT '11: Proceedings of the 15th International Conf. on Developments in Language Theory*, Vol. 6795 of *Lecture Notes in Computer Science*, pages 397–405. Springer-Verlag, 2011.

[33] A. M. Shur and I. A. Gorbunova. On the growth rates of complexity of threshold languages. *RAIRO Inform. Théor. App.* **44** (2010), 175–192.

[34] A. M. Shur. The structure of the set of cube-free Z-words in a two-letter alphabet. *Izvestiya Mathematics* **64** (2000), 847–871.

[35] A. M. Shur. Combinatorial complexity of rational languages. *Diskretn. Anal. Issled. Oper. Ser. 1* **12**(2) (2005), 78–99. In Russian.

[36] A. M. Shur. Comparing complexity functions of a language and its extendable part. *RAIRO Inform. Théor. App.* **42** (2008), 647–655.

[37] A. M. Shur. Growth properties of power-free languages. *Computer Sci. Review* **6** (2012), 187–208.

[38] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22.

[39] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67.

[40] I. N. Tunev and A. M. Shur. On two stronger versions of Dejean's conjecture. In *Proc. 37th Internat. Conf. on Mathematical Foundations of Computer Science. MFCS 2012*, Vol. 7464, pages 801–813, 2012.