RAIRO-Theor. Inf. Appl. 48 (2014) 369–389
DOI: 10.1051/ita/2014015

Available online at:
www.rairo-ita.org

BINARY PATTERNS IN BINARY CUBE-FREE WORDS: AVOIDABILITY AND GROWTH *

Robert Mercaş[1], Pascal Ochem[2], Alexey V. Samsonov[3] and Arseny M. Shur[3]

Abstract. The avoidability of binary patterns by binary cube-free words is investigated and the exact bound between unavoidable and avoidable patterns is found. All avoidable patterns are shown to be D0L-avoidable. For avoidable patterns, the growth rates of the avoiding languages are studied. All such languages, except for the overlap-free language, are proved to have exponential growth. The exact growth rates of languages avoiding minimal avoidable patterns are approximated through computer-assisted upper bounds. Finally, a new example of a pattern-avoiding language of polynomial growth is given.

Mathematics Subject Classification. 68Q70, 68R15.

1. Introduction

Factorial languages, *i.e.*, languages closed under taking factors of their words, constitute a wide and important class. Each factorial language can be defined by a set of forbidden (avoided) structures: factors, patterns, powers, Abelian powers, *etc.* In this paper, we consider languages avoiding sets of patterns.

Pattern avoidance is one of the classical topics in combinatorics of words. Recall that patterns are words over the auxiliary alphabet of variables. These variables

Keywords and phrases. Formal languages, avoidability, avoidable pattern, cube-free word, overlap-free word, growth rate, morphism.

* *Robert Mercaş was supported during his work by the Alexander von Humboldt Foundation while affiliated at Otto-von-Guericke-Universität Magdeburg, Germany.*

[1] Christian-Albrechts-Universität zu Kiel, Institut für Informatik, Germany. rgm@informatik.uni-kiel.de

[2] CNRS, LIRMM, France. Pascal.Ochem@lirmm.fr

[3] Ural Federal University, Ekaterinburg, Russia. vonosmas@gmail.com; Arseny.Shur@usu.ru

Article published by EDP Sciences

© EDP Sciences 2014

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

admit arbitrary non-empty words over the main alphabet as values. A word over the main alphabet *meets* the pattern if some factor of this word can be obtained from the pattern by assigning values to the variables, and *avoids* the pattern otherwise.

The main question concerning the avoidance of any set of forbidden structures is whether the language of all avoiding words over the main alphabet is finite or infinite. The set is called *unavoidable* in the first case and *avoidable* in the second case. We use the terms *k-(un)avoidable* to specify the cardinality of the main alphabet.

If a set of structures is avoidable, then the second question is how big is the avoiding language in terms of growth. In general, a simple constraint usually defines either a finite language or a language of exponential growth. So, the examples of languages having subexponential (*e.g.*, polynomial) growth are quite valuable.

For languages avoiding patterns, the main question is far from being satisfactorily answered even for the case of a single pattern. A complete description of the pairs (alphabet, pattern) such that the pattern is avoidable over the alphabet is known only for patterns with at most three variables [8,9,11,13,18,24] and for the patterns that are not avoidable over any alphabet [4,25]. There are very few papers about avoidable sets of patterns; we only mention a result by Petrov [15]. The only exception is the set $\{\mathsf{xxx}, \mathsf{xyxyx}\}$, defining the binary *overlap-free* language which is quite well presented in literature starting from the seminal paper by Thue [24].

There are some scattered results concerning the second question (*cf.* [5, 14]). To the best of our knowledge, the only example of a pair (alphabet, pattern) such that the language avoiding the pattern over the alphabet grows subexponentially with the length, was found in [3]: a 7-ary pattern avoidable over the quaternary alphabet. All infinite languages avoiding a binary pattern grow exponentially (combined [7, 11]). However, the binary overlap-free language has polynomial growth [16].

In this paper we start a systematic study of both questions formulated above for the languages specified by a pair of forbidden patterns. It is quite natural to begin with the binary main alphabet and consider the patterns of two variables also. For the first step, it is also natural to fix one of the patterns to be xxx, which is the shortest pattern avoidable over two letters. This step is in line with other studies of binary cube-free words with additional constraints (see, *e.g.*, [2]). In this setting, the aim of this paper is to describe the avoidability of binary patterns by the binary cube-free words and the order of growth of avoiding languages. This description is given by the following theorem. Recall that an avoidable set of structures is called *D0L-avoidable* if it is avoided by an infinite word generated by the iteration of a morphism.

Theorem 1.1 (main theorem)**.** *Let* $P \in \{\mathsf{x}, \mathsf{y}\}^*$ *be a binary pattern.*
1) *The set* $\{\mathsf{xxx}, P\}$ *of patterns is 2-avoidable if and only if* P *contains as a factor at least one of the words*

$$\mathsf{xyxyx}, \mathsf{xxyxxy}, \mathsf{xxyxyy}, \mathsf{xxyyxx}, \mathsf{xxyyxyx}, \mathsf{xyxxyxy}, \mathsf{xyxxyyxy}, \tag{1.1}$$

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

considered up to negation and reversal.
2) *All 2-avoidable sets* $\{\mathsf{xxx}, P\}$ *are 2-D0L-avoidable.*
3) *For all 2-avoidable sets* $\{\mathsf{xxx}, P\}$*, except for the set* $\{\mathsf{xxx}, \mathsf{xyxyx}\}$*, the avoiding binary language has exponential growth.*

This is an "aggregate" theorem, the proof of which does not follow a single main line but uses quite different techniques. So, we present this proof as a sequence of lesser theorems. Some of these theorems contain refinements to the main theorem (*e.g.*, lower bounds for the growth rates of avoiding languages).

Statement 3 of Theorem 1.1 leaves little hope to find a subexponentially growing binary language avoiding a pair of patterns; so, we finish the paper by showing an example of such a language avoiding a triplet of binary patterns.

The text is organized as follows. After necessary preliminaries, in Section 3 we prove statement 1 of Theorem 1.1; our proof immediately implies statement 2. In Section 4 we finish the proof of Theorem 1.1, exhibiting exponential lower bounds for the cube-free languages avoiding the pattern xyxyxx and all patterns from (1.1), except for the pattern xyxyx. In Section 5 we estimate actual growth rates of avoiding languages through the upper bounds obtained by computer. Finally, in Section 6 we give a new example of a language of polynomial growth. This language consists of cube-free words avoiding a pair of binary patterns.

2. Preliminaries

We study finite, right infinite, and two-sided infinite sequences over the main alphabet $\{0, 1\}$ and call them *words*, *ω-words*, and *Z-words*, respectively. We also consider *patterns*, which are words over the alphabet of variables $\{\mathsf{x}, \mathsf{y}\}$. Standard notions of *factor*, *prefix*, and *suffix* of a word are used. For a word w, we write $|w|$ for its length, $w[i]$ for its ith letter, and $w[i \dots j]$ for its factor starting in the ith position and ending in the jth position. Thus, $w = w[1 \dots |w|]$. Letters in an ω-word are numbered starting with 1. For a binary word or pattern w, its *negation* is the word (resp., pattern) $\bar{w}$ such that $|w| = |\bar{w}|$ and $w[i] \neq \bar{w}[i]$ for any i. The *reversal* of w is the word $w[|w|] \dots w[1]$. A word w has *period* p if $w[1 \dots |w|-p] = w[p+1 \dots |w|]$. The *exponent* of a word is the ratio between its length and its minimal period. A word is *β-free* if the exponent of any of its factors is less than β. Two words are *conjugates* if they can be represented as uv and vu, for some words u and v. If a word uv has an integer exponent greater than 1, then vu has the same exponent.

A *language* is just a set of words. A language is *factorial* if it is closed under taking factors of its elements. Any factorial language L is determined by its set of *minimal forbidden words*, *i.e.*, the words that are not in L while all their proper factors are in L. The *growth rate* of a factorial language L is defined as $\mathsf{Gr}(L) = \lim_{n\to\infty}(C_L(n))^{1/n}$, where $C_L(n)$ is the number of words of length n in L. An infinite language L grows exponentially (subexponentially) if $\mathsf{Gr}(L) > 1$ (resp., $\mathsf{Gr}(L) = 1$). A word w is said to be *(two-sided) extendable* in the language L, if,

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

for any n, L contains a word of the form uwv such that $|u|, |v| \geq n$. The set of all extendable words in L is denoted by $\mathsf{e}(L)$.

A *morphism* is any map f from words to words satisfying the condition $f(w) = f(w[1]) \dots f(w[|w|])$ for each word w. A morphism is *non-erasing* if the image of any non-empty word is non-empty, and *n-uniform* if the images of all letters have length n. An n-uniform morphism f is called *k-synchronizing* if for any factor of length k of any word $f(w)$, the starting positions of all occurrences of this factor in $f(w)$ are equal modulo n.

A word w *meets* a pattern P if an image of P under some non-erasing morphism is a factor of w; otherwise, w *avoids* P. The images of the pattern xx (resp., xxx; xyxyx) are called *squares* (resp., *cubes*, *overlaps*). The words avoiding xx (resp., xxx; both xxx and xyxyx) are *square-free* (resp., *cube-free*, *overlap-free*).

If f is a non-erasing morphism and $f(a) = au$ for a letter a and a non-empty word u, then an infinite iteration of f generates an ω-word denoted by $\mathbf{f} = f^\infty(a)$. The ω-words obtained in this way are called *D0L-words* or *purely morphic* words. The images of letters under a morphism f are called *f-blocks*. Note that the D0L-word $\mathbf{f}$ is a product of f-blocks, and also of f^n-blocks for any $n > 1$, because the morphism f^n generates the same D0L-word $\mathbf{f}$.

The *Thue–Morse morphism* is defined by the rules $\theta(0) = 01$, $\theta(1) = 10$ and generates the *Thue–Morse word* $\mathbf{t} = \theta^\infty(0)$. The factors of $\mathbf{t}$ are *Thue–Morse factors*. We use the notation $\mathbf{t}_k = \theta^k(0)$ and $\bar{\mathbf{t}}_k = \theta^k(1)$ for *θ^k-blocks*. The properties listed in Lemma 2.1 below are well known and follow by induction from the facts that $\mathbf{t}$ is a product of θ-blocks and $\theta(\mathbf{t}) = \mathbf{t}$. The third property was first proved by Thue [24]. In the same paper, Thue proved that $\mathbf{t}$ is an overlap-free word.

Lemma 2.1.
1) *The number of Thue–Morse factors of length n is $\Theta(n)$.*
2) *For any fixed k, the number of pairs of equal adjacent θ^k-blocks in any Thue–Morse factor of length n is $n/(3\times 2^k) + O(1)$.*
3) *If vv is a Thue–Morse factor, then v is either a θ^k-block or a product of three alternating θ^k-blocks, for some $k \geq 0$. The position in which vv ends in $\mathbf{t}$ is divisible by 2^k but not by 2^{k+1}.*

A set of patterns (in particular, a single pattern) is *2-avoidable* if there exists a binary ω-word avoiding this set, and *2-D0L-avoidable* if such a D0L-word exists. The existence of an avoiding ω-word is clearly equivalent to the existence of an infinite set of avoiding finite words.

3. Avoidable and unavoidable patterns

In this section we classify the binary patterns avoidable by binary cube-free words. As was already mentioned, the pattern xyxyx is avoided by the Thue–Morse word. The following observation can be easily checked by hand or by computer.

Observation 3.1. All binary cube-free patterns of length at most 5, except for the pattern xyxyx, are unavoidable by binary cube-free words.

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

TABLE 1. Longest avoiding cube-free words for unavoidable patterns.

Pattern	Longest avoiding cube-free word u	$\lvert u\rvert$
xxyyxy	010100101101001011010010011001100	33
xyxxyx	001101011001010011010110010011011001010011010110011010011	56
xyxyyx	001100100110110010011011001001011	33
xyyxxy	001101101001010010110 1100	25
xyxxyyx	0011001100100101101001011010010100101101100	43

Next we focus our attention on the patterns of length 6. For both avoidability and growth, the patterns can be studied up to negation and reversal. Thus, we obtain the list of eight patterns:

$$\mathsf{xxyxxy}, \mathsf{xxyxyx}, \mathsf{xxyxyy}, \mathsf{xxyyxx}, \mathsf{xxyyxy}, \mathsf{xyxxyx}, \mathsf{xyxyyx}, \mathsf{xyyxxy}. \tag{3.1}$$

The pattern xxyxyx is obviously avoided by the Thue–Morse word as it has the factor xyxyx. The pattern xxyxxy is also avoided by the Thue–Morse word, as was first mentioned in [8]. (For the complete set of binary patterns avoided by the Thue–Morse word see [20].) The last four words from the list (3.1) are unavoidable, as can be easily checked by computer. The longest cube-free words avoiding these patterns are listed in Table 1. The remaining two patterns xxyyxx and xxyxyy are avoidable, see Theorems 3.2 and 3.4 below.

It follows immediately from the classification of patterns of length 6 that almost all binary patterns of length 7 are avoidable. Only three patterns of length 7, namely,

$$\mathsf{xxyyxyx}, \mathsf{xyxxyxy}, \mathsf{xyxxyyx},$$

have no proper avoidable factors. The last of these patterns is unavoidable (see Tab. 1), while the first two are avoidable (see Thm. 3.4). Finally, there is a unique pattern xyxxyyxy of length 8 for which all proper prefixes and suffixes are unavoidable. But this pattern is avoidable by the Thue–Morse word [20].

Thus, we have reduced statement 1 of Theorem 1.1 to the proof of Theorems 3.2 and 3.4. Since all avoidability proofs are obtained by constructing D0L-words, we also get statement 2 of Theorem 1.1.

Theorem 3.2. *There exists a binary cube-free D0L-word avoiding the pattern* xxyyxx.

Consider the morphism μ defined by the equalities $\mu(0) = 010$, $\mu(1) = 011$[4], and the D0L-word $\mathbf{m} = \mu^\infty(0)$. Some properties of the word $\mathbf{m}$ are gathered in the following lemma.

[4]The morphism μ' defined by $\mu'(0) = 001$, $\mu'(1) = 011$, also avoids $\{\mathsf{xxx}, \mathsf{xxyyxx}\}$ (see [19]; independently discovered by J. Shallit, private communication). We prefer the morphism μ because its study allows us to prove that the avoiding language grows exponentially (see Thm. 4.3).

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

Lemma 3.3. *Let k be an arbitrary nonnegative integer.*
1) *One has* $\mathbf{m}[3k{+}1] = 0$ *and* $\mathbf{m}[3k{+}2] = 1$.
2) *The last letter in the block* $\mu^k(a)$ *is* a. *All other letters in* $\mu^k(0)$ *and* $\mu^k(1)$ *coincide.*
3) *If* u *is a factor of* $\mathbf{m}$ *and* $|u| = 3^k$, *then the starting positions of all occurrences of* u *in* $\mathbf{m}$ *are equal modulo* 3^k.
4) *If* $\mathbf{m}$ *contains a square* uu *and* $3^k \le |u| < 3^{k+1}$, *then* $|u| \in \{3^k, 2 \times 3^k\}$.
5) *Suppose that* $\mathbf{m}[r_1 3^k{+}c \dots r_2 3^k{+}c{-}1]$ *is a square for some integers* r_1, r_2, c *such that* $0 < c \le 3^k$. *Then the word* $\mathbf{m}[r_1 3^k{+}1 \dots r_2 3^k]$ *is a square as well.*

Proof. Properties 1 and 2 follow immediately from the definition of μ. Let us prove property 3 by induction on k.

The base cases are $k = 0$ (holds trivially) and $k = 1$, which follows directly from property 1. Now we let $k \ge 2$ and prove the inductive step. Assume to the contrary that two occurrences of the factor u of length 3^k have starting positions j_1 and j_2 that are different modulo 3^k. These positions are also the starting positions of the occurrences of the factor $u[1 \dots 3^{k-1}]$. Hence, $j_1 \equiv j_2 \pmod{3^{k-1}}$ by the inductive assumption. By property 2, both considered occurrences of u are preceded by the same $(j_1 \bmod 3^{k-1}) - 1$ letters. Thus, $\mathbf{m}$ contains a factor u' such that $|u'| = 3^k$, u' is a product of μ^{k-1}-blocks, and the starting positions of two occurrences of u' are different modulo 3^k. Hence, $\mathbf{m}$ also contains the factor $\mu^{-1}(u')$ of length 3^{k-1} such that the starting positions of two occurrences of $\mu^{-1}(u')$ are different modulo 3^{k-1}, in contradiction with the inductive assumption. Therefore, property 3 is proved.

Property 4 is an immediate consequence of property 3. In order to prove property 5, we note that property 2 implies the equality $\mathbf{m}[r_1 3^k{+}1 \dots r_1 3^k{+}c{-}1] = \mathbf{m}[r_2 3^k{+}1 \dots r_2 3^k{+}c{-}1]$. Hence, the two considered words are conjugates. But all conjugates of a square are squares. □

Proof of Theorem 3.2. Let us prove that $\mathbf{m}$ is cube-free and avoids xxyyxx. Aiming at a contradiction, first assume that $\mathbf{m}$ contains a cube; consider the shortest one, say u^3. Then $|u| > 2$ in view of Lemma 3.3, 1. Hence $|u| \equiv 0 \pmod 3$ by Lemma 3.3, 4. Using Lemma 3.3, 5, we find a cube u'^3 which is a product of μ-blocks. Then $\mathbf{m}$ contains the cube $(\mu^{-1}(u'))^3$, in contradiction with the choice of u^3.

The argument for the pattern xxyyxx is essentially the same. If $\mathbf{m}$ has a factor $uuvvuu$, then Lemma 3.3, 1 implies that at least one of the numbers $|u|, |v|$ is greater than 2. Then this number is divisible by 3 by Lemma 3.3, 4, and hence the other number is divisible by 3 too (Lem. 3.3, 3). Therefore, we can apply Lemma 3.3, 5 to get a factor $u'u'v'v'u'u'$ such that both u' and v' are products of μ-blocks. Then $\mathbf{m}$ contains a shorter forbidden factor $\mu^{-1}(u'u'v'v'u'u')$, contradicting to the choice of $uuvvuu$. □

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

Theorem 3.4. *There exist binary cube-free D0L-words avoiding the patterns* xxyxyy, xxyyxyx, *and* xyxxyxy, *respectively*[5].

Our proof involves a rather short computer check based on the following two lemmas.

Lemma 3.5 ([17])**.** *A morphism $f : \{0,1\} \to \{0,1\}$ is cube-free if and only if the word $f(0011010110110010010100 11)$ is cube-free.*

Lemma 3.6. *Suppose that an ω-word $\mathbf{f}$ is generated by a k-synchronizing n-uniform binary morphism f, and $P \in \{\mathsf{xxyxyy}, \mathsf{xyyxyx}\}$. Then $\mathbf{f}$ meets P if and only if $\mathbf{f}$ contains the factor $g(P)$ for some morphism g satisfying $|g(\mathsf{x})|, |g(\mathsf{y})| < k$.*

Proof. We assume that the word $\mathbf{f}$ contains a factor of the form $g(P)$ such that $\max\{|g(\mathsf{x})|, |g(\mathsf{y})|\} \geq k$ and prove that $\mathbf{f}$ must contain a shorter image of P. Let $x' = g(\mathsf{x}), y' = g(\mathsf{y}), |x'| \geq k$. The starting positions of all occurrences of x' in $\mathbf{f}$ are equal modulo n by the definition of k-synchronizing morphism. Considering the occurrences inside $g(P)$, we see that $|x'| \equiv |x'y'| \equiv 0 \pmod{n}$ if $P = \mathsf{xxyxyy}$ and $|x'y'y'| \equiv |x'y'| \equiv 0 \pmod{n}$ if $P = \mathsf{xyyxyx}$. Thus, in both cases $|x'|$ and $|y'|$ are divisible by n. The assumption $|y'| \geq k$ leads to the same result.

Now we can write $x' = x_1x_2x_3, y' = y_1y_2y_3$, where x_1, y_1 (respectively, x_2, y_2; x_3, y_3) are suffixes (respectively, products; prefixes) of f-blocks, $|x_1| = |y_1| = r$, $|x_3| = |y_3| = l$, $l + r = n$. An f-block is determined either by its prefix of length l or by its suffix of length r. Thus, $\mathbf{f}$ contains another image of P of length $|g(P)|$: the starting position of this image is either r symbols to the right or l symbols to the left from the starting position of $g(P)$. This new image $h(P)$ is a product of f-blocks. As a result, $h(\mathsf{x})$ and $h(\mathsf{y})$ are products of f-blocks also. Hence, $\mathbf{f}$ contains an image of P under the composition of f^{-1} and h; this image is shorter than $g(P)$, as required. □

Proof of Theorem 3.4. Consider the morphisms h_1, h_2, and h_3 such that

$$\begin{array}{lll} h_1(0) = 0110010 & h_2(0) = 01001 & h_3(0) = 010011 \\ h_1(1) = 1001101 & h_2(1) = 10110 & h_3(1) = 011001. \end{array}$$

Checking the condition of Lemma 3.5 by computer, we obtain that all these morphisms are cube-free. Furthermore, it can be directly verified that h_1, h_2, and h_3 are 6-, 6-, and 5-synchronizing, respectively. Hence, if the D0L-word $\mathbf{h}_1$ generated by h_1 meets the pattern $P = \mathsf{xxyxyy}$, then by Lemma 3.6, $\mathbf{h}_1$ contains an image of P of length at most $5 \times 6 = 30$. Thus, it is enough to check all factors of $\mathbf{h}_1$ of length at most 30. Any such factor is contained in the image of a factor of $\mathbf{h}_1$ of length 6; this factor, in turn, belongs to the image of a factor of length 2, while all factors of length 2 can be found in $h_1(0)$. Therefore, we just need to examine all

[5]2-D0L-avoidability of the pattern xxyxyy was first observed by J. Cassaigne who found a 12-uniform avoiding cube-free morphism (private communication). This pattern is also avoided by a cube-free "quasi-morphism" defined in [19].

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

factors of length up to 30 in the word $h_1^3(0)$. A computer check shows that there are no images of P among such factors. So, we conclude that $\mathbf{h}_1$ avoids both cubes and the pattern xxyxyy.

A similar argument for the morphism h_2 and the pattern xxyyxyx, containing xyyxyx, shows that it is enough to examine the factors of length up to 35 in the word $h_2^4(0)$. A computer check implies the desired avoidability result. In the same way, we check the factors of length up to 28 in $h_3^3(0)$ to show that the corresponding D0L-word avoids the pattern xyxxyxy. The theorem is proved. □

4. Lower bounds for the growth rates

In this section we prove lower bounds for the growth rates of the languages avoiding the sets $\{\mathsf{xxx}, P\}$, where $P = \mathsf{xyxyxx}$ or P is any of the patterns listed in (1.1), except for the pattern xyxyx. In particular, the results of this section imply statement 3 of Theorem 1.1.

The bounds are obtained using two different methods. The first method uses block replacing in the factors of D0L-words, and is purely analytic. We apply this method to the patterns xyxyxx, xxyxxy, xxyyxx. The second method uses morphisms that act on the ternary alphabet and map ternary square-free words to binary cube-free words avoiding the given patterns. This method requires some computer search and check; we apply it to the remaining four patterns. (The second method can be applied for all patterns, but the analytic bounds are a bit better.)

4.1. Replacing blocks in D0L-words

Theorem 4.1. *The number of binary cube-free words avoiding the pattern* xyxyxx *grows exponentially with the rate of at least* $2^{1/24} \approx 1.0293$.

Proof. Let L be the language of all binary cube-free words avoiding xyxyxx. Recall that L contains all Thue–Morse factors. Consider the "distorted" θ^5-block

$$\mathbf{t}' = 0110\,1001\,1001\,\mathbf{1}0110\,1001\,0110\,0110\,1001, \tag{4.1}$$

obtained from the block $\mathbf{t}_5$ by inserting the letter 1 in the 13th position, and its negation $\bar{\mathbf{t}}'$ obtained by inserting a 0 in the same way into $\bar{\mathbf{t}}_5$. Let S be the set of all ω-words that can be obtained from the Thue–Morse word $\mathbf{t}$ by replacing some of its θ^5-blocks by the corresponding distorted blocks. Available places for inserting letters are shown below:

$$\mathbf{t} = \underbrace{\mathbf{t}_2\bar{\mathbf{t}}_2\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2\mathbf{t}_2\bar{\mathbf{t}}_2}_{\mathbf{t}_5}\underbrace{\bar{\mathbf{t}}_2\mathbf{t}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\bar{\mathbf{t}}_2\mathbf{t}_2}_{\bar{\mathbf{t}}_5}\underbrace{\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2}_{\bar{\mathbf{t}}_5}\underbrace{\mathbf{t}_2\bar{\mathbf{t}}_2\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2\bar{\mathbf{t}}_2\mathbf{t}_2}_{\mathbf{t}_5}\cdots \tag{4.2}$$

Let $\mathbf{z} \in S$. It is easy to check manually that $\mathbf{z}$ does not contain short cubes; as it will be shown below, $\mathbf{z}$ does not contain long overlaps, and hence has no cubes at

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

all. Now, our goal is to prove that $\mathbf{z}$ avoids the pattern xyxyxx.

Claim. Let $w = uvuvuu$ be a minimal forbidden word for L. Then $u \in \{0, 1, 01, 10\}$.

Assume that $|u| > 1$. Then $u[i] \neq u[i{+}1]$ for all i and, moreover, $u[|u|] \neq u[1]$. Indeed, otherwise $w[i...2|uv|{+}i{+}1]$ is an image of xyxyxx, a contradiction with the minimality of w. Hence, $u \in \{(01)^s, (10)^s\}$. Since uu is not forbidden, $s = 1$. The claim is proved.

Let us consider the overlaps in $\mathbf{z}$. The case analysis below is performed up to negation. Each overlap surely contains at least one inserted letter (we assume that the inserted letters are 1's, because there will be no need to consider different inserted letters simultaneously). Two short overlaps can be easily observed inside the word $\mathbf{t}'$, see (4.1). They are $\mathbf{t}'[5...14] = 1001100110$ and $\mathbf{t}'[11...18] = 01101101$. These overlaps obviously avoid the pattern xyxyxx. One can easily check that there is no other overlap of period ≤ 10. Note that the words 0011001**1** and 1**1**011 are not Thue–Morse factors and thus their occurrences in $\mathbf{z}$ indicate an inserted letter (the bold one).

Now assume to the contrary that some word $\mathbf{z} \in S$ meets the pattern xyxyxx. Let $w = uvuvuu$ be the shortest word among the images of xyxyxx in all words $\mathbf{z} \in S$. We already know that $|uv| > 10$. So, if v contains the inserted letter 1 then one of the corresponding "indicators" 0011001**1** and 1**1**011 occurs inside uvu. Hence, the same letter was inserted in the other occurrence of v. If we delete both these inserted letters from w, we will get a shorter image of xyxyxx, contradicting to the choice of w. Thus, w contains inserted letters only inside u. Since $uvuvu$ is an overlap and hence not a Thue–Morse factor, it contains an inserted letter. Recall that $|u| \leq 2$ by the claim.

Assume that the letter 1 was inserted inside the second (counting from the left) occurrence of u. Then if $|u| = 1$, the Thue–Morse word contains the square vv. If $|u| = 2$ then $u = 10$, because the inserted letter is preceded by the same letter. So, the 0 in the first occurrence of u is not an inserted letter. Thus, $0v0v$ is a square in $\mathbf{t}$. Then the word v or the word $0v$ should be either a θ^k-block or a product of three alternating θ^k-blocks (Lem. 2.1, 3). But v ends with 10011001, see (4.1), so we get a contradiction.

Now note that if an inserted 1 is in the third occurrence of u, then the corresponding indicator 0011001**1** occurs in the right occurrence of uvu in w. Hence 0011001**1** also occurs in the prefix uvu of w. Thus, 1 was also inserted inside the second occurrence of u, which is impossible as we have shown already. So, the only remaining position for the inserted letter is in the left occurrence of u. Then $|u| = 1$ (otherwise, $\mathbf{t}$ contains an overlap), and $vuvu$ is a factor of $\mathbf{t}$. But the inserted letter is preceded by the same letter, so $uvuvu$ must be a Thue–Morse factor. This contradiction finishes the proof of the fact that the word $\mathbf{z}$ avoids xyxyxx.

Assuming $|uv| > 10$, the above argument remains valid for the image $w = uvuvu$ of the pattern xyxyx. Then $\mathbf{z}$ does not contain overlaps with periods greater than 10. Hence we conclude that $\mathbf{z}$ is cube-free.

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

Thus, we have proved that all finite factors of the word $\mathbf{z}$ belong to L. To finish the proof, we take a large enough number n and consider all Thue–Morse factors of length n. For each factor, we perform the insertions of letters into θ^5-blocks according to both (4.1) and the negation of (4.1), in all possible combinations. Thus we obtain 2^k words from L, where k stands for the number of θ^5-blocks in the processed factor. Note that the words obtained from different factors are different (for instance, such words contain indicators in different positions). A Thue–Morse factor of length n contains $n/32 + O(1)$ "regular" θ^5-blocks plus those θ^5-blocks occurring on the border of two equal θ^5-blocks, see (4.2). Using Lemma 2.1, 2, we obtain the total of $n/24 + O(1)$ blocks. Taking Lemma 2.1, 1 into account, we see that we constructed $\Theta(n)2^{n/24+O(1)}$ words from L, and the lengths of these words cover the interval $[n, 25n/24 + O(1)]$ of length $\Theta(n)$. Therefore, the growth rate of L is at least $2^{1/24}$, as desired. □

Theorem 4.2. *The number of binary cube-free words avoiding the pattern* xxyxxy *grows exponentially with the rate of at least* $2^{1/24} \approx 1.0293$.

Proof. As in the proof of Theorem 4.1, we get an exponential lower bound using multiple insertions into the Thue–Morse word. But now we need to insert rather long words, not just letters. Let L be the language of all binary cube-free words avoiding xxyxxy. Recall that L contains all Thue–Morse factors. Consider the word

$$\mathbf{t}' = 0110\,1001\,1001\,0110\,1001\,0110\,\mathbf{01010\,01\,1001\,0110\,1001\,0110}\,0110\,1001, \quad (4.3)$$

obtained from the θ^5-block $\mathbf{t}_5$ by inserting the marked factor s of length 23 in the 25th position, and its negation $\bar{\mathbf{t}}'$ obtained by inserting $\bar{s}$ in the 25th position of $\bar{\mathbf{t}}_5$. One can check directly that both $\mathbf{t}'$ and $\bar{\mathbf{t}}'$ are cube-free and avoid xxyxxy. Note that $\mathbf{t}'[25...29] = 01010$, but $\mathbf{t}'[1...28]$ is an overlap-free word ending with the square $\theta(100100)$, and $\mathbf{t}'[26...55]$ is a Thue–Morse factor. Let S be the set of all ω-words that can be obtained from the Thue–Morse word $\mathbf{t}$ by replacing some of its θ^5-blocks by the corresponding blocks $\mathbf{t}'$, $\bar{\mathbf{t}}'$. Let us consider a successive pair of inserted factors in $\mathbf{z} \in S$ (here $a, b \in \{0, 1\}$):

$$\mathbf{z} = \cdots \underbrace{\overbrace{a\,\bar{a}\,a\,\bar{a}\,a}^{r} \cdots b\,\bar{b}\,b\,\bar{b}\,b}_{\text{overlap-free word}} \cdots \quad (4.4)$$

We see that r is a Thue–Morse factor, $rb\bar{b}$ is an overlap-free word with the suffix $(\theta(\bar{b}bb))^2$. Moreover, assume for a moment that the left of the two considered insertions is withdrawn; then *the factor* $rb\bar{b}$ *still would occur in the same place.*

Let us show that an ω-word $\mathbf{z} \in S$ contains no overlap except for 01010 and 10101. Assume to the contrary that such overlaps exist. Consider the overlap $w = uvuvu$ which has the shortest period (among the overlaps in all $\mathbf{z} \in S$) and is not extendable (*i.e.*, is not contained in a longer factor of $\mathbf{z}$ with the same period). In view of (4.4), w should contain the factor 10101 or 01010. We assume w.l.o.g. that

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

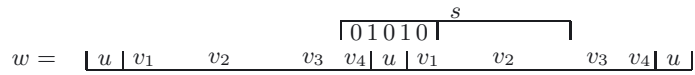
w contains 01010 and $w[i...i{+}4] = 01010$ is the rightmost occurrence of this factor in w. This occurrence in certainly not inside the prefix uvu of w. Suppose that this occurrence is inside the suffix uvu of w. Then we have $w[i{-}|uv|...i{-}|uv|{+}4] = 01010$. Both these occurrences of 01010 are prefixes of the occurrences of s in $\mathbf{z}$. Since the leftmost of these occurrences of s is obviously inside w, the rightmost one is also inside w due to non-extendability of w. Moreover, non-extendability of w implies that

- the rightmost occurrence of s is not a suffix of w, because s is always followed by $\mathbf{t}_3$;
- the leftmost occurrence of s is not a prefix of w, because s is always preceded by $\bar{\mathbf{t}}_3$.

According to this remark, we can place s inside uvu in five different ways shown in the left column of the table below. In each case we delete both mentioned occurrences of s in w and get an overlap with a smaller period in contradiction with the choice of w (see the right column of the table).

Position of s	Overlap after deletion of s
$u = u_1u_2u_3$, $s = u_2$, $u_3 \neq \lambda$	$u_3vu_1u_3vu_1u_3$
$v = v_1v_2v_3$, $s = v_2$	$uv_1v_3uv_1v_3u$
$u = u_1u_2$, $v = v_1v_2$, $s = u_2v_1$, $u_1, u_2, v_1 \neq \lambda$	$u_1v_2u_1v_2u_1$
$u = u_1u_2$, $v = v_1v_2$, $s = v_2u_1$, $v_2, u_1, u_2 \neq \lambda$	$u_2v_1u_2v_1u_2$
$u = u_1u_2u_3$, $s = u_3vu_1$, $u_2 \neq \lambda$	$u_2u_2u_2$

Thus, it remains to consider the case when the rightmost occurrence of 01010 in $w = uvuvu$ strictly contains the middle u. Since $\mathbf{t}'$ contains no overlaps except for 01010, we conclude that $|s| < |uv|$. Then the mutual location of the factors in w looks like in the following picture.



The word v_3 begins and ends with 0, and v_4 also begins with 0, see (4.3). Then the word $v_3v_3v_4$ begins with a shorter overlap, and this overlap contains at least five zeroes. Since the word $v_3v_3v_4$ occurs in an ω-word from S, we get a contradiction with the minimality of the period of w. Thus, we have proved that the "long" overlap w does not exist. Therefore, all ω-words from S contain no overlaps except for 01010 and 10101 and, in particular, are cube-free.

Now assume that $\mathbf{z} \in S$ contains an image of the pattern xxyxxy, *i.e.*, the factor $w = uuvuuv$ for some nonempty words u, v. W.l.o.g., this factor is preceded in $\mathbf{z}$ by 0. Since the word $0w$ is not an overlap, the word v ends with 1. Then $0w$ contains both factors $0uu$ and $1uu$. But one of the words $0uu$, $1uu$ is an overlap, *i.e.*, is equal to 01010 (resp., 10101). Let $v = v'1$ and consider both cases.

Case 1. $0w = 0\,0101v'1\,0101v'1$. By (4.3), v' ends with 1. Then the word w is followed by 0. Hence, $w0$ is an overlap, which is impossible.

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

Case 2. $0w = 0\,1010v'1\,1010v'1$. There is no factor 01010 or 10101 on the border between the left and the right uuv. Hence, the factors s and $\bar{s}$ in w, if any, are inside uuv (recall that w is not extendable to the right, because $\mathbf{z}$ has no long overlaps). Therefore, after deleting all occurrences of s and $\bar{s}$ in w, we will still have a square of the form $(1010\ldots)^2$. But the Thue–Morse word has no such squares, see Lemma 2.1, 3. This contradiction proves that the ω-word $\mathbf{z}$ avoids the pattern xxyxxy.

It remains to estimate the total number of factors in all words $\mathbf{z}$. Repeating the argument from the proof of Theorem 4.1, we arrive at the same bound $2^{1/24}$. □

Theorem 4.3. *The number of binary cube-free words avoiding the pattern* xxyyxx *grows exponentially with the rate of at least* $2^{1/18} \approx 1.0392$.

Proof. Proving this lower bound, we cannot rely on the Thue–Morse word, because it meets the pattern xxyyxx. Instead, we apply the insertion technique to the D0L-word $\mathbf{m}$ generated by the morphism μ, introduced in Section 3. The word $\mathbf{m}$ is a product of μ-blocks, as well as of μ^2-blocks, and has no other occurrences of such blocks by Lemma 3.3, 3. Consider the word

$$\mathbf{m}' = 010\,011\,010\,\mathbf{01010} \tag{4.5}$$

obtained by attaching the factor 01010 to the block $\mu^2(0)$. Let S be the set of all ω-words obtained from $\mathbf{m}$ by replacing some of the blocks $\mu^2(0)$ by the words $\mathbf{m}'$ (in other words, by inserting the factor 01010 after some blocks $\mu^2(0)$).

($\triangle$) If $u \in \mu^2(0, 1^*)$, and one inserts 01010 after u in a word $u\mathbf{v} \in S$, then u is followed by 010 (resp., $\mathbf{v}$ is preceded by 1010) both before and after insertion.

Note that 0101 is not a factor of $\mathbf{m}$ by Lemma 3.3, 1, and hence we use this word as a "marker". Let us show that S avoids $\{\mathsf{xxx}, \mathsf{xxyyxx}\}$. Assume to the contrary that some $\mathbf{z} \in S$ has a forbidden factor, and w is the shortest one among all forbidden factors of all words $\mathbf{z} \in S$. Since w is not a factor of $\mathbf{m}$, it contains at least one marker 0101.

The word w equals either to uuu or to $uuvvuu$, for some words u, v. If u or v contains the factor 01010, then one can cancel the corresponding insertions inside each occurrence of this word, thus getting a shorter forbidden factor in contradiction with the choice of w. Thus, all inserted factors inside w are on the borders of its parts.

Let $w = uuu$. Using the fact that $\mathbf{m}'$ is always followed by a μ^2-block, it is easy to check that $|u| \geq 5$. If uu contains 01010 somewhere in the middle, then $01010 = z_1 z_2$ and $u = z_2 u' z_1$. Hence, after cancelling the two insertions inside uuu, one obtains a shorter cube $u'u'u'$, a contradiction. Finally, if uuu ends with 0101, then this marker is a suffix of u, and we get the previous case. Thus, w has no markers, a contradiction.

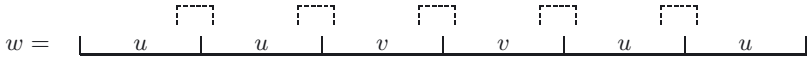
Now let $w = uuvvuu$. First consider the case where either u or v lies strictly inside some factor 01010 (and hence, is equal to 01 or 10). If $u = 01$, then $vv = 0z$,

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

where z is either a product of μ^2-blocks or such a product with 01010 inserted in the middle. In the first case $0z$ is a factor of **m** and hence is not a square by Lemma 3.3, 4. In the second case, the right half of $0z$ cannot begin with 00 as $0z$ itself does; once again we see that $0z$ is not a square. The case $u = 10$ and $vv = z0$ is symmetric to the above one.

If $v = 01$ [$v = 10$], then u begins with 00 (resp., 0) and ends with 0 (resp., 00), implying that uu contains the cube 000, which is impossible.

Thus, the factors 01010 can be found inside w only in the following places:



(In addition, w can have the suffix 0101; in case of any other partial intersection of 01010 and w, the deletion of this occurrence of 01010 from **z** leaves w unchanged by ($\triangle$).)

Consider any square in **z** containing 01010 in the middle. Such a square xx can be written in the form $z_2x'z_1\, z_2x'z_1$, where $z_1z_2 = 01010$. Then x' is a square in **m**, and thus $|x'|$ equals 3^k or 2×3^k for some $k \geq 0$ by Lemma 3.3, 4. One can easily see that for any value $|x'| = 1, 2, 3, 6$, it is impossible to obtain both squares $x'x'$ and xx. Hence, x' must be a product of μ^2-blocks ending with the block $\mu^2(0)$. Now we proceed with the case analysis.

Case 1: both uu and vv contain 01010 in the middle. Then

$$w = z_2u'z_1\, z_2u'z_1\, z_4v'z_3\, z_4v'z_3\, z_2u'z_1\, z_2u'z_1, \text{ where } z_1z_2 = z_3z_4 = 01010\,.$$

Since u' and v' are products of μ^2-blocks, we have $z_1z_4 = z_3z_2 = 01010$. Hence, $u'u'v'v'u'u'$ is a factor of **m**, a contradiction.

Case 2: uu contains 01010, while vv not. Then $w = z_2u'z_1z_2u'z_1\, vv\, z_2u'z_1z_2u'z_1$ and u' is a product of μ^2-blocks. Four subcases are possible depending on the existence of insertions on the borders of u and v.

Case 2.1: no insertions. Then z_1vvz_2 is a product of μ^2-blocks, implying $|vv| \equiv 4 \pmod 9$. By Lemma 3.3, 4, $v = 10$ and $z_1vvz_2 = 011\,010\,010$. But this is not a μ^2-block, a contradiction.

Case 2.2: an insertion only on the left. Then $v = z_2v'$. Let $\bar{v} = v'z_2$. Deleting all three insertions of $z_1z_2 = 01010$ from $w = z_2u'z_1z_2u'z_1\, z_2v'z_2v'\, z_2u'z_1z_2u'z_1$, one discovers the forbidden factor $u'u'\bar{v}\bar{v}u'u'$, contradicting the minimality of w.

Case 2.3: an insertion only on the right, is symmetric to Case 2.2.

Case 2.4: insertions on both sides. Then $v = z_2v' = v''z_1$, where $v'v''$ is a product of μ^2-blocks. Hence $|vv| \equiv 5 \pmod 9$, which is impossible by Lemma 3.3, 4.

Case 3: vv contains 01010, while uu not. Note that in this case u cannot have the suffix 0101. Then

- $w = uu\, z_2v'z_1z_2v'z_1\, uu$;

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

- v' is a product of μ^2-blocks, ending with $\mu^2(0)$ (in particular, $v' = 010\ldots1010$);
- uu is a factor of $\mathbf{m}$ (in particular, u has no factor 0101).

If $|u| = 1$, then either the first letter of z_2 or the last letter of z_1 equals u, implying that w contains a cube of a letter, which is impossible. The assumption $|u| = 2$ (*i.e.*, $u = 10$) also leads to a contradiction for all values of z_1. Namely, if z_1 ends with 0, then uuz_2 begins with $(10)^3$; if $z_1 = 01$, then $z_1uu = 011010$ is not a valid beginning of a μ^2-block; finally, $z_1 = 0101$ must be followed by 0, not by 1. Thus, $|u| \equiv 0 \pmod 3$. Let us analyze the possible values of z_1.

Case 3.1: $z_1 = 0101$, $v = 0v'0101$, $w = uu0v'01010v'0101uu$. Since u cannot end with 0101, w has exactly two occurrences of 01010 ($u[1] = 0$). Let us put $v' = v''0$, $\bar{v} = 0v''$. Deleting both occurrences of 01010, we obtain the forbidden word $uu\bar{v}\bar{v}uu$ which is shorter than w, a contradiction.

Case 3.2: $z_1 = 010$, $v = 10v'010$, $w = uu10v'01010v'010uu$. If there is no factor 01010 on the left border of vv, then uu ends in $\mathbf{m}$ in the position equal to 7 modulo 9. If this factor appears there, then uu ends in $\mathbf{m}$ in the position equal to 3 modulo 9. Similarly, if there is the factor (resp., no factor) 01010 on the right border of vv, then uu begins in the position equal to 8 modulo 9 (resp., to 4 modulo 9). Since $|u| = 0 \pmod 3$, exactly one factor 01010 should occur at the borders of vv. If this factor is on the left, we put $v' = 010v''$. Then deleting both factors 01010 we obtain a shorter forbidden factor $uuv''010v''010uu$ to get a contradiction (observe that the deleted suffix 010 of the second u is replaced by the prefix 010 of v'). Similarly, if the factor is on the right, we put $v' = v''10$ to obtain, after the deletion, a shorter forbidden factor $uu10v''10v''uu$.

For **Case 3.3**: $z_1 = 01$ and **Case 3.4**: $z_1 = 0$, the same analysis as in Case 3.2 works.

Case 4: neither uu nor vv contains 01010 in the middle. Then both uu and vv are factors of $\mathbf{m}$. We obtain contradictions between the length of uu and its starting and ending positions in $\mathbf{m}$.

Case 4.1: 01010 was inserted at the left border of vv. Since 01010 is followed by 010011, we see that either $v = 010$ or $|v| \equiv 0 \pmod 9$ by Lemma 3.3, 4. In the first case, the starting position of uu equals 4 modulo 9, and its ending position equals 2 modulo 9, contradicting Lemma 3.3, 4. If $|v| \equiv 0 \pmod 9$, let the starting position of vv be equal to k modulo 9. Then the ending position of uu equals $k+4$ modulo 9, while its starting position equals either $k-5$ or k modulo 9, depending on the existence of the factor 01010 at the right border of vv. In both cases, we have a contradiction with Lemma 3.3, 4.

Case 4.2: 01010 was inserted only at the right border of vv. Similar to Case 4.1, we analyze the possible lengths of v ($|v| = 1$, $|v| = 6$, and $|v| \equiv 0 \pmod 9$), obtaining that the length of u cannot satisfy Lemma 3.3, 4.

Thus, we finished the case study, obtaining contradictions in all cases. Hence, the forbidden word w does not exist, and the set S avoids $\{\mathsf{xxx}, \mathsf{xxyyxx}\}$. Finally,

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

we estimate the total number of factors in all words $\mathbf{z} \in S$, similar to the proof of Theorem 4.1. The word $\mathbf{m}$ has $\Theta(n)$ factors of length n; this follows, *e.g.*, from Pansiot's classification theorem (see [10]). It is clear that such a factor contains $n/2 + O(1)$ zeroes and then, $n/18 + O(1)$ factors $\mu^2(0)$. The latter quantity coincides with the number of places for insertions of the factor 01010. Thus, from the factors of $\mathbf{m}$ of length n we can construct $\Theta(n)2^{n/18+O(1)}$ factors of words from S. The lengths of these factors cover the interval of length $\Theta(n)$. Therefore, the growth rate of the binary language avoiding $\{\mathsf{xxx}, \mathsf{xxyyxx}\}$ is at least $2^{1/18}$, as required. □

4.2. Mapping ternary square-free words

In this section, we explore another approach for getting lower bounds. Namely, the fact that the language of ternary square-free words has exponential growth leads to the following simple observation.

Observation 4.4. If an n-uniform injective morphism $f : \{0,1,2\}^* \to \{0,1\}^*$ transforms any square-free ternary word to a binary word avoiding $\{\mathsf{xxx}, P\}$, then the number of such binary words grows exponentially at rate at least $\alpha^{1/n}$, where α is the growth rate of the language of ternary square-free words.

The morphisms with the desired properties can be obtained using the method described in [13]. The number α is known with a quite high precision: $1.3017597 < \alpha < 1.3017619$ (*cf.* [23]).

Theorem 4.5. *The number of binary cube-free words avoiding the pattern P, where $P \in \{\mathsf{xxyxyy}, \mathsf{xxyyxyx}, \mathsf{xyxxyxy}, \mathsf{xyxxyyxy}\}$, grows exponentially with the rate of at least*

- $\alpha^{1/14} \approx 1.0190$ *for* $P = \mathsf{xxyxyy}$*;*
- $\alpha^{1/13} \approx 1.0205$ *for* $P = \mathsf{xxyyxyx}, \mathsf{xyxxyxy}$*;*
- $\alpha^{1/10} \approx 1.0267$ *for* $P = \mathsf{xyxxyyxy}$.

Proof. In the proof of Theorem 3.4 we used morphic preimages to reduce the proof of pattern avoidance to the exhaustive search of forbidden factors in short words. Since we cannot iterate morphisms acting on alphabets of different sizes, here we need a different argument for such a reduction. For this purpose, we construct binary words avoiding simultaneously cubes, the pattern P, and large squares. We use the notation S_t for the t-ary pattern $(\mathsf{x}_1 \dots \mathsf{x}_t)^2$.

Consider the morphisms g_1, g_2, g_3, and g_4 such that

$$\begin{array}{ll} g_1(0) = 01011001100101 & g_2(0) = 0100110011011 \\ g_1(1) = 00110110010011 & g_2(1) = 0100101101001 \\ g_1(2) = 00101001101011 & g_2(2) = 0011011001001 \end{array}$$

$$\begin{array}{ll} g_3(0) = 0010110110011 & g_4(0) = 0101100110 \\ g_3(1) = 0010110011011 & g_4(1) = 0101001011 \\ g_3(2) = 0010011010011 & g_4(2) = 0100110010. \end{array}$$

For any square-free word $w \in \{0,1,2\}^*$ we claim that

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

- the word $g_1(w)$ avoids $\{\mathsf{xxx}, \mathsf{xxyxyy}, S_8\}$;
- the word $g_2(w)$ avoids $\{\mathsf{xxx}, \mathsf{xxyyxyx}, S_9\}$;
- the word $g_3(w)$ avoids $\{\mathsf{xxx}, \mathsf{xyxxyxy}, S_{10}\}$;
- the word $g_4(w)$ avoids $\{\mathsf{xxx}, \mathsf{xyxxyyxy}, S_8\}$.

To prove this claim, we notice that for every binary pattern P considered in this section, both variables x and y are involved in a square. This implies that in a word containing only squares of bounded length, potential occurrences of P and of cubes have bounded length as well. So we can check exhaustively that $g_i(w)$ avoids cubes and P for all short square-free words w. Let a *large square* be an occurrence of S_{t_i}, where $t_1 = t_4 = 8$, $t_2 = 9$, $t_3 = 10$. There remains to prove that if w is square-free, then $g_i(w)$ does not contain large squares. The proof is the same for all four morphisms.

Let $n_i = |g_i(a)|$, $a \in \{0, 1, 2\}$. First we check that the morphism g_i is $2n_i$-synchronizing. Indeed, any factor of $g_i(w)$ of length $2n_i$ contains a g_i-image of some letter a; but it is easy to see that for any letters $a, b, c \in \{0, 1, 2\}$, the factor $g_i(a)$ only appears in $g_i(bc)$ as a prefix or as a suffix. Then we check that no large square appears in the g_i-image of a ternary square-free word of length 5. So, a potential large square uu in $g_i(w)$ is such that $|u| > 2n_i$ and thus $|u| = qn_i$ for some integer $q \geq 3$ by the synchronization property. So uu is contained in the image of a word of the form $w = avbvc$ with $a, b, c \in \Sigma_3$ and the center of uu lies in $g_i(b)$. Moreover, $a \neq b$ and $b \neq c$ since w is square-free. This implies that abc is square-free and that $g_i(abc)$ contains a square $u'u'$ with $|u'| = n_i$. Now $u'u'$ is a large square because $n_i > t_i$ for all our morphisms g_i. This is a contradiction since no large square appears in the g_i-image of a ternary square-free word of length 5. The claim is proved, and the theorem follows. □

Proving Theorem 4.5, we actually showed that the considered binary patterns can be avoided by binary cube-free words simultaneously with large squares. So, a natural problem is to find the exact bound for the length of these large squares. The following theorem gives this bound for all patterns listed in (1.1).

Theorem 4.6. *Let* $P \in \{\mathsf{xxyxyx}, \mathsf{xxyxxy}, \mathsf{xxyxyy}, \mathsf{xxyyxx}, \mathsf{xxyyxyx}, \mathsf{xyxxyxy}, \mathsf{xyxxyyxy}\}$ *and let* $t(P)$ *be the number such that the set of patterns* $\{\mathsf{xxx}, P, S_{t(P)}\}$ *is 2-avoidable while the set* $\{\mathsf{xxx}, P, S_{t(P)-1}\}$ *is 2-unavoidable. Then*

$$t(P) = \begin{cases} 4 & \text{if } P = \mathsf{xxyyxx}, \\ 5 & \text{if } P \in \{\mathsf{xxyxyy}, \mathsf{xxyyxyx}, \mathsf{xyxxyxy}, \mathsf{xyxxyyxy}\}, \\ 7 & \text{if } P \in \{\mathsf{xxyxxy}, \mathsf{xxyxyx}\}. \end{cases}$$

and the binary language avoiding $\{\mathsf{xxx}, P, S_{t(P)}\}$ *has exponential growth.*

Proof. Below we list the morphisms mapping ternary square-free words to the binary words avoiding the required sets. The proof of avoidability and exponential

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

growth is the same as for Theorem 4.5.

$P = \mathsf{xxyyxx}$, length $= 62$
$0 \to 00100101101100101001101101001001101011001010011011001001101011$
$1 \to 00100101101100101001101100100110101100101001101101001001101011$
$2 \to 00100101101100101001101011001001101100101001101101001001101011$

$P = \mathsf{xxyxyy}$, length $= 88$
$0 \to 00100110101100101001100110101100110010100110101100100110110010100$
1100110101100101001101
$1 \to 00100110101100101001100110101100100110110010100110101100110010100$
1100110101100101001101
$2 \to 00100110101100101001100110101100100110110010100110011010110010100$
1101011001100101001101

$P = \mathsf{xyxxyxy}$, length $= 49$
$0 \to 00110010110110011010010011001101011001010011010 11$
$1 \to 00110010110110011010010011001011011001010011010 11$
$2 \to 00110010110110010011010110010100110110010011010 11$

$P = \mathsf{xxyyxyx}$, length $= 32$
$0 \to 00100110110100100110011011010011$
$1 \to 00100101101001001101101001011011$
$2 \to 00100101100110110100100110011011$

$P = \mathsf{xyxxyyxy}$, length $= 28$
$0 \to 00100101101001100110101100 11$
$1 \to 00100101101001100101101100 11$
$2 \to 00100101100110110100101100 11$

$P = \mathsf{xxyxyx}$, length $= 44$
$0 \to 00100110011010011001011001101001011011001101$
$1 \to 00100110010110110011001011001101001100101101$
$2 \to 00100110010110011010010110110011001011001101$

$P = \mathsf{xxyxxy}$, length $= 66$
$0 \to 00101001100101100110100110010110100110101100110100110010100110101 11$
$1 \to 00101001100101100110100110010100110101100110100110010110100110101 11$
$2 \to 00101001100101100110100101100101001101011001101001100101100110101 11$

Unavoidability of shorter squares is verified by computer search. □

5. Growth rates: numerical results

A general method to obtain upper bounds for the growth rates of factorial languages was proposed in [22]. An open-source implementation of this method can be found in [1]. We adjust this method for each pattern under consideration and calculate the upper bounds for the growth rates of avoiding binary cube-free language. Here is a high-level overview of the method.

Let L be a factorial language and M be its set of minimal forbidden words. If L is an infinite language avoiding a pattern, then M is also infinite. We construct

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

TABLE 2. Growth rates of binary cube-free languages avoiding binary patterns: upper bounds.

Pattern	Upper bound	Pattern	Upper bound
xxyxxy	1.098891	xyxyx	1 (previously known)
xxyxyy	1.226850	xxyyxyx	1.310975
xyxyxx	1.138449	xyxxyxy	1.281612
xxyyxx	1.322304	xyxxyyxy	1.348932

a family $\{M_i\}$ of finite subsets of M such that

$$M_1 \subseteq M_2 \subseteq \cdots \subseteq M_i \subseteq \cdots \subseteq M, \quad M_1 \cup M_2 \cup \cdots \cup M_i \cup \cdots = M.$$

Let L_i be the binary factorial language with the set of minimal forbidden words M_i. One has

$$L \subseteq \cdots \subseteq L_i \subseteq \cdots \subseteq L_1, \quad L_1 \cap L_2 \cap \cdots \cap L_i \cap \cdots = L.$$

It is not hard to show that the sequence of growth rates $\{\mathsf{Gr}(L_i)\}$ decreases and converges to $\mathsf{Gr}(L)$. The languages L_i are regular, and then the number $\mathsf{Gr}(L_i)$ can be found with any degree of precision. Increasing i, one can make the upper bound arbitrarily close to $\mathsf{Gr}(L)$.

Thus, to obtain an upper bound for $\mathsf{Gr}(L)$ one should make three steps. First, build a set of minimal forbidden words M_i for the chosen i. Second, convert this set into a deterministic finite automaton recognizing L_i (the automaton should be both accessible and coaccessible). And finally, calculate the number $\mathsf{Gr}(L_i)$. If we calculate M_i by some search procedure and store it in a trie, then the second step can be implemented as a modified Aho–Corasick algorithm for pattern matching that converts the trie into an automaton having the desired properties. At the third step we calculate the growth rate of L_i with any prescribed precision by an efficient (linear in the size of automaton) iterative algorithm. The second and third steps are common for all factorial languages.

For each pattern we use an ad-hoc procedure for constructing the set of minimal forbidden words for avoiding languages. In most cases we bound the length of the constructed forbidden words with some constant. We iterate over the candidate forbidden words in the order of increasing length and check that they do not contain proper forbidden factors, using already built shorter forbidden words for pruning. In practice, the described method allows us to construct and handle sets of thousands of forbidden words and automata of millions of vertices efficiently. Some numerical results are presented in Table 2. For each of the processed languages, the sequence of obtained upper bounds converges very fast. So, the actual value of the growth rate in each case is likely to be quite close to the given upper bound.

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

6. A language of polynomial growth

Statement 3 of Theorem 1.1, proved in Section 4, tells us that xyxyx is the only binary pattern that is avoided by a subexponentially-growing infinite set of binary cube-free words. In this section, we present two binary patterns P_1 and P_2 such that the binary language avoiding $\{\mathsf{xxx}, P_1, P_2\}$ has polynomial growth. This language contains the binary overlap-free language and is incomparable with the binary (7/3)-free language (the latter one is the biggest binary β-free language of polynomial growth [12]). Thus, this is an essentially new example of a language of polynomial growth.

Theorem 6.1. *The binary cube-free language avoiding both the patterns* xyxyxx *and* xxyxyx *has polynomial growth.*

Proof. Let L be the language of all binary cube-free words avoiding both xyxyxx and xxyxyx. Obviously, both L and its extendable part $\mathsf{e}(L)$ contain the set of all Thue–Morse factors. We aim to prove that this set coincides with $\mathsf{e}(L)$. The definition of extendable word implies that any word from $\mathsf{e}(L)$ is a factor of a Z-word all finite factors of which also belong to $\mathsf{e}(L)$.

For any word from L, the factors

$$000, 010101, 010100, 11001001, 10010011, 010010010,$$

and their negations are forbidden. Hence, a word $w \in \mathsf{e}(L)$ has no factor 01010, because any its extension to the right contains 010101 or 010100. Similarly, w has no factor 00100: extending this word, we inevitably meet one of the words 000, 11001001, 10010011, or $(010)^3$. The same argument applies for 10101 and 11011.

Claim. If a Z-word $\mathbf{z}$ has no factors 000, 01010, 00100, and their negations, then $\mathbf{z}$ is a product of θ-blocks.

If two squares of letters in a word begin in positions of different parity, then this word surely contains one of the listed factors. To see this, just consider the closest pair of such squares. So, all squares of letters in $\mathbf{z}$ occur in positions of the same parity. Hence, one can factorize $\mathbf{z}$ into the factors of length 2 in a way that splits any square of a letter, thus getting the desired product.

Consider a Z-word $\mathbf{z}$ all factors of which belong to $\mathsf{e}(L)$. By the claim, $\mathbf{z}$ is a product of θ-blocks. Consider its Thue–Morse preimage $\mathbf{z}' = \theta^{-1}(\mathbf{z})$. The Z-word $\mathbf{z}'$ avoids the patterns xxx, xxyxyx, and xyxyxx. Indeed, if $\mathbf{z}'$ contains an image of a pattern under f, then $\mathbf{z}$ contains an image of the same pattern under θf. Hence, $\mathbf{z}'$ has no factors listed in the claim, and we conclude that it is a product of θ-blocks. Then $\mathbf{z}$ is a product of θ^2-blocks. Repeating this argument inductively, we obtain that $\mathbf{z}$ is a product of θ^n-blocks for any n. Therefore, any finite factor of $\mathbf{z}$ is a factor of some θ^n-block, *i.e.*, a Thue–Morse factor, as desired.

The set of Thue–Morse factors contains $\Theta(n)$ words of length n, and then has the growth rate 1. But the languages L and $\mathsf{e}(L)$ always have the same growth

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

rate (see [21], Thm. 3.1), so our language L grows subexponentially. To prove that this growth is polynomial, some additional work is needed.

Let us take an overlap $w = 0v0v0 \in L$ with $|v| > 2$ and analyze how it can be extended within L. The words $0w, w0$ are images of xxyxyx and xyxyxx, respectively, so, $0w, w0 \notin L$. Note that v begins or ends with 1, because w has no factor 000. Assuming w.l.o.g. that $v = 1v'$ and extending w to the right by one symbol, we get a longer overlap: $w1 = 01v'01v'01$. We see that $w11$ and $w101$ contain images of xyxyxx. Assume that $w100 = 01v'01v'0100 \in L$.

If the last letter of v' is 1, then v ends with 11, because $010100 \notin L$. Then v' cannot begin with 1, because the factor 11011 in the middle of w means that w contains a forbidden factor (compare to the beginning of the proof). But if $v' = 0v''$, we see that the word $w100 = 01\,0v''010v''0100$ meets the pattern xyxyxx ($x \to 0$, $y \to v''01$). So, v ends with 0 and then with 10. Then the word $w1001$ ends with 1001001, guaranteeing that $w10011, w10010 \notin L$. Thus, we have proved the following property.

(▲) Suppose that $w = uvuvu \in L$, $|uv| \geq 4$, and $|u| > 1$. Then w can be extended within L by at most three letters to each side.

Finally, we estimate the number of words in L that are not (7/3)-free. These words contain overlaps with $|u| \geq |v|/2$. From (▲) it follows that the set of words in L containing overlaps such that $|u| > 1$ and $|u| \geq |v|/2$, is finite. So, it remains to consider the case $|u| = 1$ (and then $|v| \leq 2$). If $|uv| = 2$, the overlap is 01010 or 10101. It cannot be extended within L. Now let $|uv| = 3$. Such an overlap must contain the factor 00100 or 11011, which cannot be extended within L to both sides simultaneously by more than one letter. Then the words from L containing an overlap of period 3 have the form $0010010z$ or $10010010z$ up to reversal and negation. The number of such words grows polynomially because the word z is overlap-free. Since the number of (7/3)-free words is also polynomial, we get a polynomial upper bound on the number of words in L. □

Remark 6.2. Concerning the bounds for the degree of the polynomial growth of the language L considered in Theorem 6.1, we have shown, in fact, that one can take the upper bound derived for the (7/3)-free language in [6]. The obvious lower bound stems from the fact that L contains all overlap-free words.

Acknowledgements. We are grateful to J. Cassaigne and N. Rampersad for helpful discussions, and to anonymous referees for thorough reading of the text and valuable comments.

References

[1] Growth-rate-calculator. Library for calculating growth rates of factorial formal languages (2013). Available at http://code.google.com/p/growth-rate-calculator/.

[2] G. Badkobeh, S. Chairungsee and M. Crochemore, Hunting redundancies in strings. In *Proc. 15th Developments in Language Theory. DLT 2011*, Vol. 6795 of *Lect. Notes Sci.* Springer, Berlin (2011) 1–14.

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press

[3] K.A. Baker, G.F. McNulty and W. Taylor, Growth problems for avoidable words. *Theoret. Comput. Sci.* **69** (1989) 319–345.
[4] D.A. Bean, A. Ehrenfeucht and G. McNulty, Avoidable patterns in strings of symbols. *Pacific J. Math.* **85** (1979) 261–294.
[5] J.P. Bell and T.L. Goh, Exponential lower bounds for the number of words of uniform length avoiding a pattern. *Inform. Comput.* **205** (2007) 1295–1306.
[6] V.D. Blondel, J. Cassaigne and R. Jungers, On the number of α-power-free binary words for $2 < \alpha \leq 7/3$. *Theoret. Comput. Sci.* **410** (2009) 2823–2833.
[7] F.-J. Brandenburg, Uniformly growing k-th power-free homomorphisms. *Theoret. Comput. Sci.* **23** (1983) 69–82.
[8] J. Cassaigne, Unavoidable binary patterns. *Acta Informatica* **30** (1993) 385–395.
[9] J. Cassaigne, Motifs évitables et régularités dans les mots (Thèse de Doctorat). *Tech. Report.* LITP-TH 94-04 (1994).
[10] C. Choffrut and J. Karhumäki, Combinatorics of words. In vol. 1 of *Handbook of Formal Languages*, edited by G. Rozenberg and A. Salomaa. Springer-Verlag (1997) 329–438.
[11] P. Goralcik and T. Vanicek, Binary patterns in binary words. *Internat. J. Algebra Comput.* **1** (1991) 387–391.
[12] J. Karhumäki and J. Shallit, Polynomial versus exponential growth in repetition-free binary words. *J. Combin. Theory. Ser. A* **104** (2004) 335–347.
[13] P. Ochem, A generator of morphisms for infinite words. *RAIRO: ITA* **40** (2006) 427–441.
[14] P. Ochem, Binary words avoiding the pattern AABBCABBA. *RAIRO: ITA* **44** (2010) 151–158.
[15] A.N. Petrov, Sequence avoiding any complete word. *Mathematical Notes of the Academy of Sciences of the USSR* **44** (1988) 764–767.
[16] A. Restivo and S. Salemi, Overlap free words on two symbols. In *Automata on Infinite Words*, edited by M. Nivat and D. Perrin, *Ecole de Printemps d'Informatique Théorique, Le Mont Dore, 1984*, vol. 192 of *Lect. Notes Sci.* Springer-Verlag (1985) 198–206.
[17] G. Richomme and F. Wlazinski, About cube-free morphisms. In *STACS 2000, Proc. 17th Symp. Theoretical Aspects of Comp. Sci.*, vol. 1770 of *Lect. Notes Sci.* Edited by H. Reichel and S. Tison. Springer-Verlag (2000) 99–109.
[18] P. Roth, Every binary pattern of length six is avoidable on the two-letter alphabet. *Acta Informatica* **29** (1992) 95–107.
[19] A.V. Samsonov and A.M. Shur, Binary patterns in binary cube-free words: Avoidability and growth. In *Proc. 14th Mons Days of Theoretical Computer Science*. Univ. catholique de Louvain, Louvain-la-Neuve (2012) 1–7. electronic.
[20] A.M. Shur, Binary words avoided by the Thue–Morse sequence. *Semigroup Forum* **53** (1996) 212–219.
[21] A.M. Shur, Comparing complexity functions of a language and its extendable part. *RAIRO: ITA* **42** (2008) 647–655.
[22] A.M. Shur, Growth rates of complexity of power-free languages. *Theoret. Comput. Sci.* **411** (2010) 3209–3223.
[23] A.M. Shur, Growth properties of power-free languages. *Comput. Sci. Rev.* **6** (2012) 187–208.
[24] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912) 1–67.
[25] A.I. Zimin, Blocking sets of terms. *Mat. Sbornik* **119** (1982) 363–375; In Russian. English translation in *Math. USSR Sbornik* **47** (1984) 353–364.

Communicated by Ch. Choffrut.
Received March 14, 2014. Accepted March 17, 2014.

https://doi.org/10.1051/ita/2014015 Published online by Cambridge University Press