Fewest repetitions in infinite binary words

Golnaz Badkobeh, Maxime Crochemore

Fewest repetitions in infinite binary words

Golnaz Badkobeh
King's College London, London, UK

Maxime Crochemore
King's College London, London, UK
and Université Paris-Est, France

July 25, 2012

Abstract

A square is the concatenation of a nonempty word with itself. A word has period p if its letters at distance p match. The exponent of a nonempty word is the quotient of its length over its smallest period.

In this article we give a proof of the fact that there exists an infinite binary word which contains finitely many squares and simultaneously avoids words of exponent larger than 7/3.

Our infinite word contains 12 squares, which is the smallest possible number of squares to get the property, and 2 factors of exponent 7/3. These are the only factors of exponent larger than 2.

The value 7/3 introduces what we call the finite-repetition threshold of the binary alphabet. We conjecture it is 7/4 for the ternary alphabet, like its repetitive threshold.

Keywords: combinatorics on words, repetitions, word morphisms.
MSC: 68R15 Combinatorics on words.

1 Introduction

Repetitions in words is a basic question in Theoretical Informatics, certainly because it is related to many applications although it has first been studied by Thue at the beginning of the twentieth century [11] with a pure theoretical objective. Related results apply to the design of efficient string pattern matching algorithm, to text compression methods and entropy analysis, as well as to the study of repetitions in biological molecular sequences among others.

The knowledge of the strongest constraints an infinite word can tolerate help for the design and analysis of efficient algorithms. The optimal bound on the maximal exponent of factors of the word has been studied by Thue and many other authors after him. One of the first discoveries was that an infinite binary word can avoid factors with an exponent larger than 2, called 2^+-powers. This has been extended by Dejean [3] to the ternary alphabet and her famous

conjecture on the repetitive threshold for larger alphabets has eventually been proved recently after a series of partial results by different authors (see [9, 2] and references therein).

Another constraint is considered by Fraenkel and Simpson [4]: their parameter to the complexity of binary infinite words is the number of squares occurring in them without any restriction on the number of occurrences. It is fairly straightforward to check that no infinite binary word can contain less than three squares and they proved that some of them contain exactly three. Two of these squares appear in the cubes `000` and `111` so that the maximum exponent is 3 in their word. In this article we produce an infinite word with few distinct squares and a smaller maximal exponent.

Fraenkel and Simpson's proof uses a pair of morphisms, one to get an infinite word by iteration, the other to produce the final translation on the binary alphabet. Their result has been proved with different pairs of morphisms by Rampersad et al. [8] (the first morphism is uniform), by Harju and Nowotka [5] (the second morphism accepts any infinite square-free word), and by Badkobeh and Crochemore [1] (the simplest morphisms).

In this article we show that we can combine the two types of constraints for the binary alphabet: producing an infinite word whose maximal exponent of its factor is the smallest possible while containing the smallest number of squares. The maximal exponent is 7/3 and the number of squares is 12 to which can be added two words of exponent 7/3.

It is known from Karhumäki and Shallit [6] that if an infinite binary word avoids 7/3-powers it contains an infinite number of squares. Proving that it contains more than 12 squares is indeed a matter of simple computation.

Shallit [10] has built an infinite binary word avoiding $7/3^+$-powers and all squares of period at least 7. His word contains 18 squares.

Our infinite binary word avoids the same powers but contains only 12 squares, the largest having period 8. As before the proof relies on a pair of morphisms satisfying suitable properties. Both morphisms are almost uniform (up to one unit). The first morphism is weakly square-free on a 6-letter alphabet, and the second does not even correspond to a uniquely-decipherable code but admits a unique decoding on the words produced by the first. To get the morphisms, we first examined carefully the structure of long words satisfying the conditions and obtained by backtracking computation. Then, we inferred the morphisms from the regularities found in the words.

After introducing the definitions and main results in the next section, we provide a weakly square-free morphism and the infinite square-free word on 6 letters it generates in Section 3. Section 4 shows how this word is translated into an infinite binary word satisfying the constraints. In the conclusion we define the new notion of finite-repetition threshold and state a conjecture on its value for the 3-letter alphabet.

2 Repetitions in binary words

A word is a sequence of letters drawn from a finite alphabet. We consider the binary alphabet $B = \{0, 1\}$, the ternary alphabet $A_3 = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, and the 6-letter alphabet $A_6 = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}\}$.

A square is a word of the form uu where u is a nonempty (finite) word. A word has period p if its letters at distance p are equal. The exponent of a nonempty word is the quotient of its length over its smallest period. Thus, a square is any word with an even integer exponent.

In this article we consider infinite binary words in which a small number of squares occur.

The maximal length of a binary word containing less than three square is finite. It can be checked that it is 18, e.g. 010011000111001101 contains only 00 and 11. But, as recalled above, this length is infinite if 3 squares are allowed to appear in the word. A simple proof of it relies on two morphisms f and h_0 defined as follows. The morphism f is defined from A_3^* to itself by

$$\begin{cases} f(\mathtt{a}) = \mathtt{abc}, \\ f(\mathtt{b}) = \mathtt{ac}, \\ f(\mathtt{c}) = \mathtt{b}. \end{cases}$$

It is known that the infinite word $\mathbf{f} = f(\mathtt{a})^\infty$ it generates is square-free (see [7, Chapter 2]). The morphism h_0 is from A_3^* to B^* and defined by

$$\begin{cases} h(\mathtt{a}) = \mathtt{01001110001101}, \\ h(\mathtt{b}) = \mathtt{0011}, \\ h(\mathtt{c}) = \mathtt{000111}. \end{cases}$$

Then the result is a consequence of the next statement.

Theorem 1 ([1]) *The infinite word* $\mathbf{h_0} = h_0(f(\mathtt{a})^\infty)$ *contains the* 3 *squares* 00, 11 *and* 1010 *only. The cubes* 000 *and* 111 *are the only factors occurring in* $\mathbf{h}$ *and of exponent larger than* 2.

It is impossible to avoid 2^+-powers and keep a bounded number of squares. As proved by Karhumäki and Shallit [6], the exponent has to go up to 7/3 to allow the property.

In the two following sections we define two morphisms and derive the properties that we need to prove the next statement.

Theorem 2 *There exists an infinite binary word whose factors have an exponent at most* 7/3 *and that contains* 12 *squares, the fewest possible.*

Our infinite binary word contain the 12 squares $\mathtt{0}^2$, $\mathtt{1}^2$, $(\mathtt{01})^2$, $(\mathtt{10})^2$, $(\mathtt{001})^2$, $(\mathtt{010})^2$, $(\mathtt{011})^2$, $(\mathtt{100})^2$, $(\mathtt{101})^2$, $(\mathtt{110})^2$, $(\mathtt{01101001})^2$, $(\mathtt{10010110})^2$, and the two words 0110110 and 1001001 of exponent 7/3.

Proving that it is impossible to have less than 12 squares in the previous statement results from the next table. It has been obtained by pruned backtracking sequential computation that avoids exhaustive search. It shows the

maximal length of binary words whose factors have an exponent at most 7/3, for each number s of squares, $0 \leq s \leq 11$.

s	$=0$	1	2	3	4	5	6	7	8	9	10	11
$\ell(s)$	$=3$	5	8	12	14	18	24	30	37	43	83	116

3 A weakly square-free morphism on six letters

In this section we consider a specific morphism used for the proof of Theorem 2. It is called g and defined from A_6^* to itself by:

$$\begin{cases} g(\mathtt{a}) = \mathtt{abac}, \\ g(\mathtt{b}) = \mathtt{babd}, \\ g(\mathtt{c}) = \mathtt{eabdf}, \\ g(\mathtt{d}) = \mathtt{fbace}, \\ g(\mathtt{e}) = \mathtt{bace}, \\ g(\mathtt{f}) = \mathtt{abdf}. \end{cases}$$

We prove below that the morphism is weakly square-free in the sense that $\mathbf{g} = g^\infty(\mathtt{a})$ is an infinite square-free word, that is, all its finite factors have an exponent smaller than 2. Note that however it is not square-free since for example $g(\mathtt{cf}) = \mathtt{eabdfabdf}$ contains the square $(\mathtt{abdf})^2$. This prevents from using characterisation of square-freeness of the morphism, or equivalently of the fixed points of the morphism. As far as we know only an ad hoc proof is possible.

The set of codewords $g(a)$'s ($a \in A_6$) is a prefix code and therefore a uniquely-decipherable code. Note also that any occurrence of $\mathtt{abac}$ in $g(w)$, for $w \in A_6^*$, uniquely corresponds to an occurrence of $\mathtt{a}$ in w. The proof below relies on the fact that not all doublets and triplets (words of length 2 and 3 respectively) occur in $\mathbf{g}$, as the next statements show.

Lemma 1 *The set of doublets occurring in* $\mathbf{g}$ *is*

$$D = \{\mathtt{ab}, \mathtt{ac}, \mathtt{ba}, \mathtt{bd}, \mathtt{cb}, \mathtt{ce}, \mathtt{da}, \mathtt{df}, \mathtt{ea}, \mathtt{fb}\}.$$

Proof. Note that all letters of A_6 appear in $\mathbf{g}$. Then doublets $\mathtt{ab}$, $\mathtt{ac}$, $\mathtt{ba}$, $\mathtt{bd}$, $\mathtt{ce}$, $\mathtt{df}$, $\mathtt{ea}$, $\mathtt{fb}$ appear in $\mathbf{g}$ because they appear in the images of one letter. The images of these doublets generate two more doublets, $\mathtt{cb}$ and $\mathtt{da}$, whose images do not create new doublets. ■

Lemma 2

The set of triplets in $\mathbf{g}$ *is*

$$T = \{\mathtt{aba}, \mathtt{abd}, \mathtt{acb}, \mathtt{ace}, \mathtt{bab}, \mathtt{bac}, \mathtt{bda}, \mathtt{bdf}, \mathtt{cba}, \mathtt{cea}, \mathtt{dab}, \mathtt{dfb}, \mathtt{eab}, \mathtt{fba}\}.$$

Proof. Triplets appear in the images of a letter or of a doublet. Triplets found in images of one letter are: $\mathtt{aba}$, $\mathtt{abd}$, $\mathtt{ace}$, $\mathtt{bab}$, $\mathtt{bac}$, $\mathtt{bdf}$, $\mathtt{eab}$, $\mathtt{fba}$. The images of doublets occurring in $\mathbf{g}$, in set D of Lemma 1, contain the extra triplets: $\mathtt{acb}$, $\mathtt{bda}$, $\mathtt{cba}$, $\mathtt{cea}$, $\mathtt{dab}$, $\mathtt{dfb}$. ■

Table 1: Gaps of abac: words between consecutive occurrences of abac in **g**. They are images of gaps between consecutive occurrences of a.

$g(\mathtt{b})$	=	babd	4
$g(\mathtt{cb})$	=	eabdfbabd	9
$g(\mathtt{bd})$	=	babdfbace	9
$g(\mathtt{ce})$	=	eabdfbace	9
$g(\mathtt{bdfb})$	=	babdfbaceabdfbabd	17

To prove that the infinite word **g** is square-free we first show that it contains no square with less than four occurrences of the word $g(\mathtt{a}) = \mathtt{abac}$. Then, we show it contains no square with at least four occurrences of it. The word abac is chosen because its occurrences in **g** correspond to $g(\mathtt{a})$ only, so they are used to synchronise the parsing of the word according to the codewords $g(\mathtt{a})$'s.

Lemma 3 *No square in* **g** *can contain less than four occurrences of* abac.

Proof. Assume by contradiction that a square ww in **g** contains less than four occurrences of abac. Let x be the shortest word whose image by g contains ww.

Then x is a factor of **g** that belongs to the set $\mathtt{a}((A_6 \setminus \{\mathtt{a}\})^*\mathtt{a})^5$. Since two consecutive occurrences of a in **g** are separated by a string of length at most 4 (the largest such string is indeed bdfb as a consequence of Lemma 2), the set is finite.

The square-freeness of all these factors has been checked via an elementary implementation of the test, which proves the result. ■

Proposition 1 *No square in* **g** *can contain at least four occurrences of* abac.

Proof. The proof is by contradiction: let k be the maximal integer for which $g^k(\mathtt{a})$ is square-free and let ww be a square occurring in $g^{k+1}(\mathtt{a})$ and containing at least 4 occurrences of abac. Distinguishing several cases according to the words between consecutive occurrences of abac (see Table 1), we deduce that $g^k(\mathtt{a})$ is not square-free, the contradiction.

The square ww can be written

$$\underbrace{v_0(\mathtt{abac}\cdots\mathtt{abac})u_1}\underbrace{v_1(\mathtt{abac}\cdots\mathtt{abac})u_2}$$

where v_0, u_1, v_1, and u_2 contain no occurrence of abac. It occurs in the image of a factor of **g**. The central part of w starting and ending with abac is the image of a unique word U factor of $g^k(\mathtt{a})$ due to the code property:

$$g(U) = v_0^{-1}wu_1^{-1} = v_1^{-1}wu_2^{-1}.$$

We split the proof in two parts according to whether abac occurs in u_1v_1 or not.

No abac in u_1v_1. We consider five cases according to the value of u_1v_1, the gap of abac (see Table 1).

1. $u_1v_1 =$ babd corresponds to $g(\mathtt{b})$ only. If either u_1 or v_1 is empty, then v_0 or u_2 is $g(\mathtt{b})$, in either case we get $\mathtt{b}U\mathtt{b}U$ or $U\mathtt{b}U\mathtt{b}$ that are squares. Else v_0 has a suffix d so it belongs to $g(\mathtt{b})$, and again $\mathtt{b}U\mathtt{b}U$ is a square in g.

2. $u_1v_1 =$ eabdfbabd corresponds to $g(\mathtt{cb})$ only. An occurrence of cb always belongs to $g(\mathtt{ab})$ therefore U has a prefix abd and a suffix aba, and the letter after aba is c. If v_1 is empty, u_2 has a prefix eabdfbabd so it is $g(\mathtt{cb})$ and again $U\mathtt{cb}U\mathtt{cb}$ is a square. If v_1 is not empty then v_0 has a suffix d, suffix of $g(\mathtt{b})$, therefore $\mathtt{b}U\mathtt{cb}U\mathtt{c}$ is a square.

3. $u_1v_1 =$ babdfbace corresponds to $g(\mathtt{bd})$. The word abda is a factor of $g(\mathtt{ba})$ only so U has a prefix aba and a suffix ba. If $|u_1| = 0$, $v_0 =$ babdfbace can only be $g(\mathtt{bd})$ so $\mathtt{bd}U\mathtt{bd}U$ is a square. Otherwise u_2 must have a prefix b; since U has a suffix ba the next letter after it is either b or c; as only $g(\mathtt{b})$ is prefixed by b the letter is b so u_2 has a prefix or is a prefix of $g(\mathtt{b})$, and we know that bab is always followed by d thus $U\mathtt{bd}U\mathtt{bd}$ is a square.

4. $u_1v_1 =$ eabdfbace corresponds to $g(\mathtt{ce})$ only. If u_1 is empty, v_0 is $g(\mathtt{ce})$ so $\mathtt{ce}U\mathtt{ce}U$ is a square. Otherwise, u_2 has a prefix or is a prefix of $g(\mathtt{c})$; the next letter after $g(\mathtt{c})$ is either b or e; (see Lemma 1); if it is b the right-most U has a suffix aba but the left-most U has a suffix fba, which cannot be. Therefore the letter after c is e and $U\mathtt{ce}U\mathtt{ce}$ is a square.

5. $u_1v_1 =$ babdfbaceabdfbabd. If $|v_1| > 12$, v_0 has a suffix $g(\mathtt{dfb})$ and the letter before it is b, so $\mathtt{bdfb}U\mathtt{bdfb}U$ is a square. If $0 < |v_1| \leq 12$, then $|u_1| \geq 5$, so u_2 has a prefix or is a prefix of $g(\mathtt{bd})$ so the next letter is either a or f. If it is a the right-most U has a suffix ba but v_0 is a suffix of or has a suffix $g(\mathtt{b})$; the letter before it is either $g(\mathtt{c})$ or $g(\mathtt{f})$; if it is c then U has a prefix abd and bdfbabd is from the concatenation of $g(\mathtt{c})$ and $g(\mathtt{b})$ or $g(\mathtt{dfb})$; in either case the left occurrence of U will have ea as a suffix, a contradiction since $\mathtt{fb}U\mathtt{bdfb}U\mathtt{bd}$ and $U\mathtt{bdfb}U\mathtt{bdfb}$ are both squares.

An occurrence of abac in u_1v_1. Then the suffix of u_1 is either aba, ab or a while the respective prefix of v_1 is c, ac or bac.

Note that c is followed either by b or e (Lemma 1) and that cb occurs only in the image of ab. Then if the occurrence of abac is followed by b, the occurrence of cb in v_0 is preceded by aba, and then there is a square starting 1, 2 or 3 positions before the occurrence of ww, which brings us back to the first case. Therefore, abac is followed by e.

The occurrence of abace comes from $g(\mathtt{ac})$, and by Lemma 2 u_1v_1 contains an occurrence of $g(\mathtt{bac})$. So, the occurrence of abace is preceded by d, and since da occurs only in the image of ba, the occurrence of da in u_2 is followed by bac, which yields a square starting 1, 2 or 3 positions after the occurrence of ww. Again this takes us back to the first case.

In all cases we deduce the existence of a square in $g^k(\mathtt{a})$, which is a contradiction with the definition of k. Therefore there is no square in $\mathbf{g}$ containing at least four occurrences of $\mathtt{abac}$. ■

The next corollary is a direct consequence of Lemma 3 and Proposition 1.

Corollary 1 *The infinite word* $\mathbf{g}$ *is square-free, or equivalently, the morphism* g *is weakly square-free.*

4 Binary translation

The second part of the proof of Theorem 2 consists in showing that the special infinite square-free word on 6 letters introduced in the previous section can be transformed into the desired binary word. This is done with a second morphism h from A_6^* to B^* defined by

$$\begin{cases} h(\mathtt{a}) = \mathtt{10011}, \\ h(\mathtt{b}) = \mathtt{01100}, \\ h(\mathtt{c}) = \mathtt{01001}, \\ h(\mathtt{d}) = \mathtt{10110}, \\ h(\mathtt{e}) = \mathtt{0110}, \\ h(\mathtt{f}) = \mathtt{1001}. \end{cases}$$

Note that the codewords of h do not form a prefix code, nor a suffix code, nor even a uniquely-decipherable code! We have for example $g(\mathtt{ae}) = \mathtt{10011 \cdot 0110} = \mathtt{1001 \cdot 10110} = g(\mathtt{fd})$. However, parsing the word $h(y)$ when y is a factor of $\mathbf{g}$ is unique due to the absence of some doublets and triplets in it (see Lemmas 1 and 2). For example $\mathtt{fd}$ does not occur, which induces the unique parsing of $\mathtt{100110110}$ as $\mathtt{10011 \cdot 0110}$.

Proposition 2 *The infinite word* $\mathbf{h} = h(g^\infty(\mathtt{a}))$ *contains no factor of exponent larger than* $7/3$. *It contains the* 12 *squares* $\mathtt{0}^2$, $\mathtt{1}^2$, $(\mathtt{01})^2$, $(\mathtt{10})^2$, $(\mathtt{001})^2$, $(\mathtt{010})^2$, $(\mathtt{011})^2$, $(\mathtt{100})^2$, $(\mathtt{101})^2$, $(\mathtt{110})^2$, $(\mathtt{01101001})^2$, $(\mathtt{10010110})^2$ *only. Words* $\mathtt{0110110}$ *and* $\mathtt{1001001}$ *are the only factors with an exponent larger than* 2.

The proof is based on the fact that occurrences of $\mathtt{10011}$ in $\mathbf{h}$ identify occurrences of $\mathtt{a}$ in $\mathbf{g}$ and on the unique parsing mentioned above. It proceeds by considering several cases according to the gaps between consecutive occurrences of $\mathtt{10011}$ (see Table 2), associated with gaps between consecutive occurrences of $\mathtt{a}$ in $\mathbf{g}$, which leads to analyse paths in the graph of Figure 1.

Proof. We show that if $\mathbf{h}$ would contain a square not in the list it would come from a square in $\mathbf{g}$, which cannot be since $\mathbf{g}$ is square-free (Corollary 1).

Suppose $\mathbf{h}$ contains the square w^2. It is a factor of $h(g^k(\mathtt{a}))$, for some integer k and can be written $\underbrace{v_0(h(\mathtt{a})\cdots h(\mathtt{a}))u_1}\,\underbrace{v_1(h(\mathtt{a})\cdots h(\mathtt{a}))u_2}$. The central part of

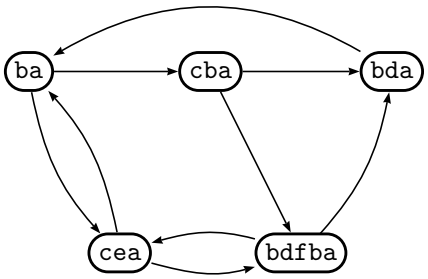


Figure 1: Graph showing immediate successors of gaps in the word $\mathbf{g}$: a suffix of it following an occurrence of $\mathtt{a}$ is the label of an infinite path.

Table 2: Gaps between consecutive occurrences of $\mathtt{10011}$ in $\mathbf{h}$.

$h(\mathtt{b})$	=	$\mathtt{01100}$	5
$h(\mathtt{cb})$	=	$\mathtt{0100101100}$	10
$h(\mathtt{bd})$	=	$\mathtt{0110010110}$	10
$h(\mathtt{ce})$	=	$\mathtt{010010110}$	9
$h(\mathtt{bdfb})$	=	$\mathtt{0110010110100101100}$	19

w is the image of a unique square-free factor U of $g^k(\mathtt{a})$ due to the unique parsing mentioned above:

$$h(U) = (h(\mathtt{a}) \cdots h(\mathtt{a})) = v_0^{-1} w u_1^{-1} = v_1^{-1} w u_2^{-1}.$$

We proceed through different cases as in the proof of Proposition 1.

No $h(\mathtt{a})$ in $u_1 v_1$.

1. $u_1 v_1 = \mathtt{01100}$ corresponds to $h(\mathtt{b})$ only.

   If $|v_1| > 1$, then v_0 belongs to $h(\mathtt{b})$, $\mathtt{b}U\mathtt{b}U$ is a square. Else $|u_1| \geq 4$ so u_2 belongs to $h(\mathtt{b})$, it cannot belong to $h(\mathtt{e})$ since $\mathtt{ae}$ is not a factor of $\mathbf{g}$, therefore $U\mathtt{b}U\mathtt{b}$ is a square of $\mathbf{g}$.

2. $u_1 v_1 = \mathtt{0110010110}$ corresponds to $h(\mathtt{bd})$.

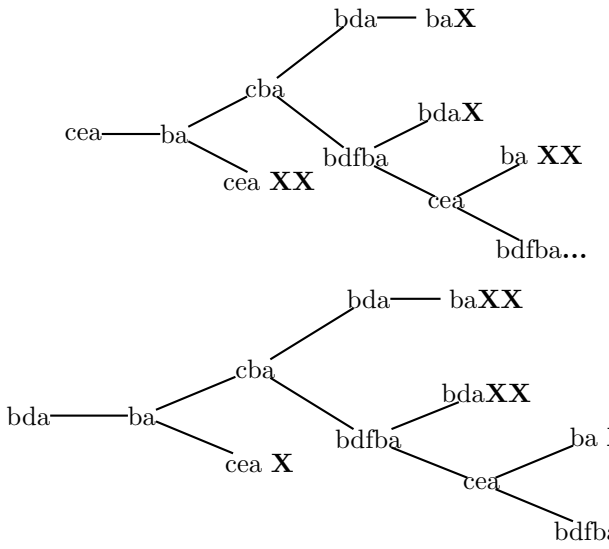
   $$v_0 \underbrace{(h(\mathtt{a}) \cdots h(\mathtt{a}))}\, h(\mathtt{bd}) \underbrace{(h(\mathtt{a}) \cdots h(\mathtt{a}))}\, u_2$$

   the word $\mathtt{abda}$ is a factor of $g(\mathtt{ba})$ only, so U has a prefix $\mathtt{abac}$ and a suffix $\mathtt{ba}$ (Note that U cannot be $\mathtt{aba}$ since $\mathtt{ababdaba}$ is not a factor of $\mathbf{g}$).

   $$v_0 \underbrace{(h(\mathtt{abac}) \cdots h(\mathtt{ba}))}\, h(\mathtt{bd}) \underbrace{(h(\mathtt{abac}) \cdots h(\mathtt{ba}))}\, u_2$$

If u_2 comes from or has a prefix $h(\mathtt{b})$ then the letter after $\mathtt{bab}$ is always $\mathtt{d}$ so we have the square $U\mathtt{bd}U\mathtt{bd}$. Then u_2 is a prefix of or has a prefix $h(\mathtt{c})$, the longest common prefix (LCP) of $h(\mathtt{c})$ and $h(\mathtt{b})$ is $\mathtt{01}$, so v_0 has a suffix $\mathtt{10010110}$, which is a suffix of $h(\mathtt{bd})$ or $h(\mathtt{ce})$. If v_0 comes from $h(\mathtt{bd})$ then we have the square $\mathtt{bd}U\mathtt{bd}U$. So v_0 is a suffix of $h(\mathtt{ce})$

$$h(\mathtt{ce})\,\underbrace{(h(\mathtt{abac})\cdots h(\mathtt{ba}))}\,h(\mathtt{bd})\,\underbrace{(h(\mathtt{abac})\cdots h(\mathtt{ba}))}\,h(\mathtt{c}).$$



The sign XX shows that the particular branch of the trie terminates because either a square occurs or the sequence is not a factor of $\mathbf{g}$. The sign X on the other hand represents the termination of a particular branch as a consequence of the discontinuation of the corresponding branch in the other trie. If we continue these tries we will have:

$$\mathtt{ce}\,\underbrace{\mathtt{abac\,babd\,fbace\,abdf\,babd\,abac\,eabdf\,bace\,abac\,babd\,abac\,eabdf\,\ldots ba}}$$

$$\mathtt{bd}\,\underbrace{\mathtt{abac\,babd\,fbace\,abdf\,babd\,abac\,eabdf\,bace\,abac\,babd\,abac\,eabdf\,\ldots ba}}\,\mathtt{c}$$

which is the image of

$$\mathtt{eabdf\,bace}\,\underbrace{\mathtt{abac\ldots abac}}\,\mathtt{babd\,fbace}\,\underbrace{\mathtt{abac\ldots abac}}\,\mathtt{e}$$

itself image of

$$\mathtt{ce}\,\underbrace{\mathtt{a\ldots a}}\,\mathtt{bd}\,\underbrace{\mathtt{a\ldots a}}\,\mathtt{c}$$

so we have the same situation as at the starting point; but U is shorter in this case, therefore if we continue this process we should have

$$\mathtt{ce\,abac\,babd\,fbace\,abdf\,babd\,abac\,babd\,fbace\,abdf\,bace\,a}$$

but $\mathtt{abdf\,bace}$ is the image of $\mathtt{fe}$ that is not in D (Lemma 1).

3. $u_1v_1 = \texttt{0100101100}$ corresponds to $h(\texttt{cb})$.

   The word $\texttt{acba}$ is a factor of $g(\texttt{ab})$ only, so U has a prefix $\texttt{abd}$ and a suffix $\texttt{aba}$:

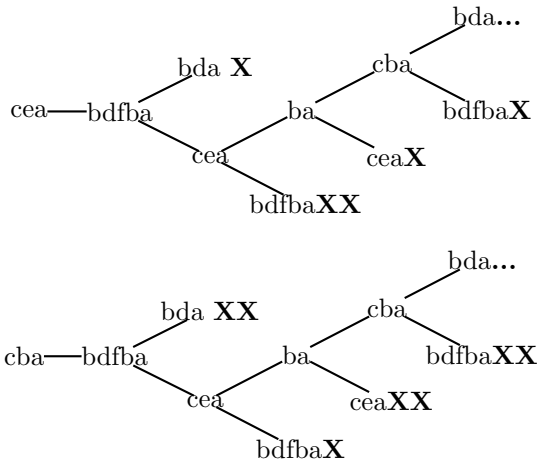
   $$v_0 \underbrace{(h(\texttt{abd})\dots h(\texttt{aba}))}\, h(\texttt{cb}) \underbrace{(h(\texttt{abd})\dots h(\texttt{aba}))}\, u_2$$

   The word u_2 comes from or has a prefix $h(\texttt{c})$. If the letter after it is $\texttt{b}$, we have the square $U\texttt{cb}U\texttt{cb}$.

   Otherwise u_2 comes from or has a prefix $h(\texttt{ce})$. If v_0 comes from or has a suffix $h(\texttt{b})$ then we have the square $\texttt{b}U\texttt{cb}U\texttt{c}$.

   Therefore the letter before U is $\texttt{e}$ preceded by $\texttt{c}$, i.e. the string before the left U is $\texttt{ce}$:

   $$h(\texttt{ce}) \underbrace{(h(\texttt{abd})\dots h(\texttt{aba}))}\, h(\texttt{cb}) \underbrace{(h(\texttt{abd})\dots h(\texttt{aba}))}\, h(\texttt{ce}).$$



   Now we have the same situation as in the previous case

   $$h(g(\texttt{ce})) \underbrace{(h(g(\texttt{abac}))\dots h(g(\texttt{ba})))}\, h(g(\texttt{bd})) \underbrace{(h(g(\texttt{abac}))\dots h(g(\texttt{ba})))}\, h(g(\texttt{c})).$$

4. $u_1v_1 = \texttt{010010110}$ corresponds to $h(\texttt{ce})$ only.

   Before $\texttt{c}$ is always $\texttt{ba}$ (Lemma 2) and after $\texttt{e}$ is $\texttt{ab}$ (Lemma 2), so $\texttt{ab}$ is a prefix of U and $\texttt{ba}$ is a suffix of U:

   $$v_0 \underbrace{(h(\texttt{ab})\dots h(\texttt{ba}))}\, h(\texttt{ce}) \underbrace{(h(\texttt{ab})\dots h(\texttt{ba}))}\, u_2.$$

   (i): u_2 belongs to $h(\texttt{cb})$ since we cannot have $U\texttt{ce}U\texttt{ce}$ and the letter after $\texttt{c}$ is $\texttt{b}$ or $\texttt{e}$ (Lemma 1):

   $$v_0 \underbrace{(h(\texttt{ab})\dots h(\texttt{ba}))}\, h(\texttt{ce}) \underbrace{(h(\texttt{ab})\dots h(\texttt{ba}))}\, h(\texttt{cb})$$

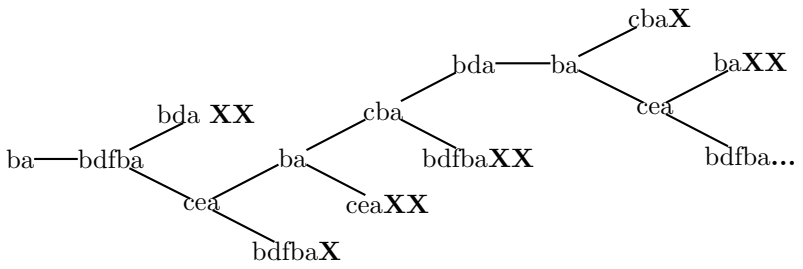
The letter before bacb is a so:

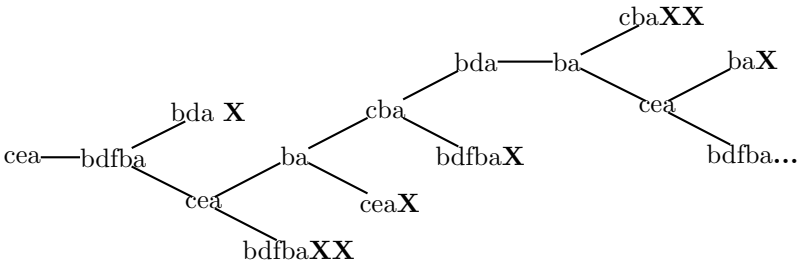
$$v_0 \underbrace{(h(\mathtt{ab})\dots h(\mathtt{aba}))} h(\mathtt{ce}) \underbrace{(h(\mathtt{ab})\dots h(\mathtt{aba}))} h(\mathtt{cb}).$$

NOTE: U is not aba since abaceabacb is not a factor of $g^k(\mathtt{a})$.

Now abace is a prefix of the image of ac so U has a prefix abdf and the word before it is either ce or b; the first choice gives the square ce$U$ce$U$ and the second choice:

$$h(\mathtt{b}) \underbrace{(h(\mathtt{abdf})\dots h(\mathtt{aba}))} h(\mathtt{ce}) \underbrace{(h(\mathtt{abdf})\dots h(\mathtt{aba}))} h(\mathtt{cb}).$$





Now if we continue the above tries we get:

$$\mathtt{b}\,\underbrace{\mathtt{abd\,fbace\,abac\,babd\,abac\,eabdf\,babd\,abac\,babd\,fbace\,abdf\,ba\dots ba}}$$

$$\mathtt{ce}\,\underbrace{\mathtt{abdf\,bace\,abac\,babd\,abac\,eabdf\,babd\,abac\,babd\,fbace\,abdf\,ba\dots ba}}\,\mathtt{cb}$$

which is the image of

$$\mathtt{bd}\,\underbrace{\mathtt{abac\,babd\,fbace\,abdf\dots ba}}\,\mathtt{ce}\,\underbrace{\mathtt{abac\,babd\,fbace\,abdf\dots ba}}\,\mathtt{b}.$$

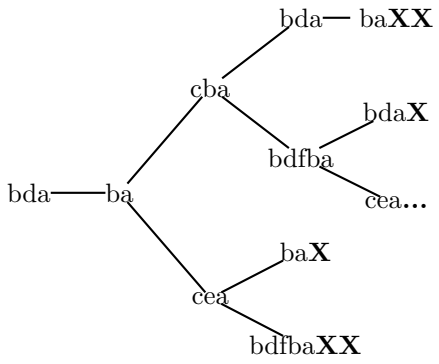
This is the same situation as the next case and we will see that after going one step back it brings us back to this case again. Now we are exactly in the same situation as at the beginning except that the length of the word $X = \mathtt{abdf}\dots\mathtt{a}$ is shorter than U. Repeating this process enough times we should see that the word

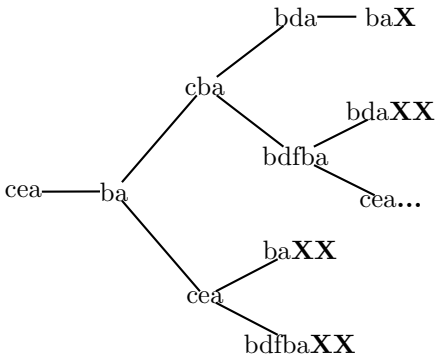
$$\mathtt{babd\,fbace\,abac\,babd\,abac\,eabdf\,bace\,abac\,babd\,aba}$$

which is the image of bdabaceaba, is not a factor of $g^k(\mathtt{a})$.

(ii): u_2 belongs to $h(\mathtt{b})$ (the LCP of $h(\mathtt{c})$ and $h(\mathtt{b})$ is 01) so v_0 must have a suffix 0010110, which belongs to $h(\mathtt{bd})$ because if it belongs to $h(\mathtt{ce})$ then $\mathtt{ce}U\mathtt{ce}U$ is a square.

$$h(\mathtt{bd})\,\underbrace{(h(\mathtt{ab})\dots h(\mathtt{ba}))}\,h(\mathtt{ce})\,\underbrace{(h(\mathtt{ab})\dots h(\mathtt{ba}))}\,h(\mathtt{b}).$$





Continuing this trie we have

$$\mathtt{bd}\,\underbrace{\mathtt{abac\,babd\,fbace\,a\dots ba}}\,\mathtt{ce}\,\underbrace{\mathtt{abac\,babd\,fbace\,a\dots ba}}\,\mathtt{bd}.$$

This is factor of $g(\mathtt{b}\,\underbrace{\mathtt{abdf\dots a}}\,\mathtt{ce}\,\underbrace{\mathtt{abdf\dots a}}\,\mathtt{cb})$ which is the previous case.

5. $u_1v_1 = \mathtt{0110010110100101100}$ corresponds to $h(\mathtt{bdfb})$ only. This case is dealt with the same method.

$$u_0\,\underbrace{(h(\mathtt{a})\dots h(\mathtt{a}))}\,h(\mathtt{bdfb})\,\underbrace{(h(\mathtt{a})\dots h(\mathtt{a}))}\,u_2.$$

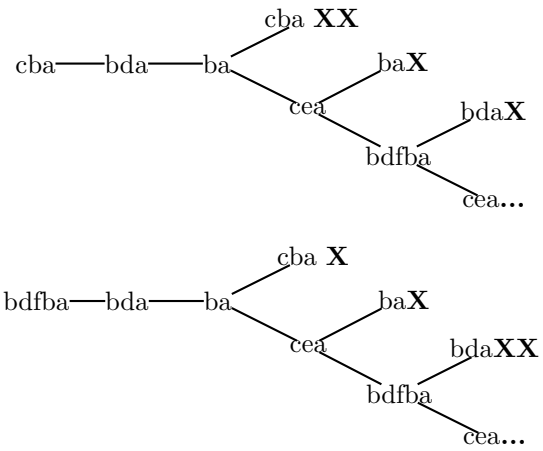
If u_2 belongs to $h(\mathtt{c})$, the LCP of $h(\mathtt{c})$ and $h(\mathtt{b})$ is 01 so u_0 must have a suffix 10010110100101100, therefore u_0 belongs to $h(\mathtt{bdfb})$. But $\mathtt{bdfb}U\mathtt{bdfb}U$

is a square and a factor of $g^k(a)$; a contradiction, so u_2 belongs to or has a prefix $h(\mathtt{b})$. We have two choices here.

(i): the next word after the right occurrence of U is $\mathtt{ba}$. The LCP of $h(\mathtt{bd})$ and $h(\mathtt{ba})$ is $\mathtt{10}$, u_0 has suffix of $\mathtt{110100101100}$, so it either belongs to $h(\mathtt{dfb})$ or $h(\mathtt{acb})$. The first case gives that $\mathtt{dbf}U\mathtt{bdbf}U\mathtt{b}$ is a square and a factor of $g^k(a)$, a contradiction. So u_0 belongs to $h(\mathtt{acb})$:

$$h(\mathtt{acb})\underbrace{(h(\mathtt{abda})\dots h(\mathtt{a}))}\,h(\mathtt{bdf\ b})\underbrace{(h(\mathtt{abda})\dots h(\mathtt{a}))}\,h(\mathtt{ba}).$$

Prefixes and suffixes of U are determined only by looking at D and T.



We have:

$$\mathtt{abac\ babd}\underbrace{\mathtt{abac\ eabdf\ bace}\dots\mathtt{abac}}\mathtt{\ babd\ fbace}$$

$$\mathtt{abdf\ babd}\underbrace{\mathtt{abac\ eabdf\ bace}\dots\mathtt{abac}}\mathtt{\ babd\ fbace\ abac}$$

which is the image of

$$\mathtt{ab}\underbrace{\mathtt{ace}\dots\mathtt{a}}\mathtt{\ bdfb}\underbrace{\mathtt{ace}\dots\mathtt{a}}\mathtt{\ bda}.$$

Now this is the next case so if we go back enough steps we should see that the length of U decreases and at the end we get

$$\mathtt{ac\ babd\ abac\ eabdf\ babd\ abac\ eaba}$$

but this is not a factor of $g^k(a)$, a contradiction.

(ii): the word after U is $\mathtt{bd}$. Now here the only possible letter after $\mathtt{abd}$ is $\mathtt{a}$ since if it is $\mathtt{f}$ it is a prefix of $\mathtt{fb}$ so we have $U\mathtt{bdfb}U\mathtt{bdfb}$, a contradiction. As the LCP of $h(\mathtt{bdfb})$ and $h(\mathtt{bda})$ is $\mathtt{01100101101001}$ u_0 must have a suffix $\mathtt{01100}$ so it can belong to $h(\mathtt{ab})$ or $h(\mathtt{acb})$.

(I):

$$h(\mathtt{ab})\underbrace{(h(\mathtt{a})\dots h(\mathtt{a}))}h(\mathtt{bdfb})\underbrace{(h(\mathtt{a})\dots h(\mathtt{a}))}h(\mathtt{bda}).$$

Only using D, T and the Figure 1 we can continue building U,

$$h(\mathtt{ab})\underbrace{(h(\mathtt{ace})\dots h(\mathtt{ba}))}h(\mathtt{bdfb})\underbrace{(h(\mathtt{acea})\dots h(\mathtt{ba}))}h(\mathtt{bda}).$$

Continuing further we get:

$$h(\mathtt{abac\,eabdf\,babd}\underbrace{\mathtt{abac}\dots\mathtt{abac}}\mathtt{babd\,fbace\,abdf\,babd}\underbrace{\mathtt{abac}\dots\mathtt{abac}}\mathtt{babda}).$$

This is the image of

$$h(g(\mathtt{acb}\underbrace{\mathtt{a}\dots\mathtt{a}}\mathtt{bdfb}\underbrace{\mathtt{a}\dots\mathtt{a}}\mathtt{ba}))$$

and we are back to the case above.

(II):

$$h(\mathtt{acb})\underbrace{(h(\mathtt{a})\dots h(\mathtt{a}))}h(\mathtt{bdfb})\underbrace{(h(\mathtt{a})\dots h(\mathtt{a}))}h(\mathtt{bda}).$$

Using the same method we build the word U:

$$\mathtt{ac\,b}\underbrace{\mathtt{abd}\dots\mathtt{ba}}\mathtt{bd\,fb}\underbrace{\mathtt{ace}\dots\mathtt{ba}}\mathtt{bd\,a}.$$

Here we cannot go further as U cannot have `abd` nor `ace` as prefixes at the same time.

An occurrence of $h(\mathtt{a})$ in u_1v_1. Looking at Figure 1, the image of the concatenation of two connected nodes (distance 1 arrow) are the possibilities for $u_1v_1h(\mathtt{a})$, but note that the second period of the square must start within $h(a)$, starting point of the arrow, otherwise it is one of the cases above. If the lengths of both nodes are larger than 2 then by unique parsing we are bound to have a square in $g^k(a)$ and get a contradiction. So we have to consider only the four cases where one of the nodes is `ba`:

1. $u_1v_1 = h(\mathtt{bacb}) = 01100\mathbf{10011}0100101100$, so u_2 must have a prefix $h(\mathtt{b})$ and u_0 a suffix of $h(\mathtt{cb})$, before `cb` is always `a`, so $acbUbacbUb$ is a square in $g^k(a)$.

2. $u_1v_1 = h(\mathtt{bace}) = 01100\mathbf{10011}010010110$, so u_2 must have a prefix $h(\mathtt{b})$ and u_0 a suffix $h(\mathtt{ce})$, before `ce` is always `a`, so $aceUbaceUb$ is a square in $g^k(a)$.

3. $u_1v_1 = h(\mathtt{ceab}) = 0100101101\mathbf{0011}01100$, so u_2 must have a prefix of $h(\mathtt{ce})$ and u_0 a suffix of $h(\mathtt{b})$, after `ce` is always `a`, so $bUceabUcea$ is a square in $g^k(a)$.

4. $u_1v_1 = h(\texttt{bdab}) = 01100101101001101100$, so using tries as before shows that after enough backward iteration we should have

$$\texttt{fbace abdf babd abac babd abac eabdf babd abac babd}$$

which contains a square.

In all cases the conclusion is that we get a square in $g^k(\texttt{a})$, a contradiction with the definition of k. This completes the proof of Proposition 2. ■

Theorem 2 follows immediately from Proposition 2.

5 Conclusion

The constraint on the number of squares imposed on binary words slightly differs from the constraint considered by Shallit [10]. The squares occurring in his word have period smaller than 7. Our word contains less squares but their maximal period is 8.

Looking at repetitions in words on larger alphabets, the subject introduces a new type of threshold, that we call the *finite-repetitions threshold* (FRt). For the alphabet of a letters, $\mathrm{FRt}(a)$ is defined as the smallest rational number for which there exists an infinite word avoiding $\mathrm{FRt}(a)^+$-powers and containing a finite number of r-powers, where r is Dejean's repetitive threshold. Karhumäki and Shallit results as well as ours show that $\mathrm{FRt}(2) = 7/3$. Our result additionally proves that the associated minimal number of squares is 12.

Computation shows that the maximal length of $(7/4)^+$-free ternary word with only one $7/4$-repetition is 102. This leads us state the following conjecture, which has been tested up to length 20000.

Conjecture 1 *The finite-repetitions threshold of 3-letter alphabet is* $\frac{7}{4}$ *and the associated number of* $\frac{7}{4}$*-powers is* 2.

Values for larger alphabets remain to be explored.

References

[1] G. Badkobeh and M. Crochemore. An infinite binary word containing only three distinct squares. 2010. Submitted.

[2] J. D. Currie and N. Rampersad. A proof of Dejean's conjecture. *Math. Comput.*, 80(274):1063–1070, 2011.

[3] F. Dejean. Sur un théorème de Thue. *J. Comb. Theory, Ser. A*, 13(1):90–99, 1972.

[4] A. S. Fraenkel and J. Simpson. How many squares must a binary sequence contain? *Electr. J. Comb.*, 2, 1995.

[5] T. Harju and D. Nowotka. Binary words with few squares. *Bulletin of the EATCS*, 89:164–166, 2006.

[6] J. Karhumäki and J. Shallit. Polynomial versus exponential growth in repetition-free binary words. *J. Comb. Theory, Ser. A*, 105(2):335–347, 2004.

[7] M. Lothaire, editor. *Combinatorics on Words*. Cambridge University Press, second edition, 1997.

[8] N. Rampersad, J. Shallit, and M. Wei Wang. Avoiding large squares in infinite binary words. *Theor. Comput. Sci.*, 339(1):19–34, 2005.

[9] M. Rao. Last cases of Dejean's conjecture. *Theor. Comput. Sci.*, 412(27):3010–3018, 2011.

[10] J. Shallit. Simultaneous avoidance of large squares and fractional powers in infinite binary words. *Intl. J. Found. Comput. Sci*, 15(2):317–327, 2004.

[11] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana*, 7:1–22, 1906.