Bounded Number of Squares in Infinite Repetition-Constrained Binary Words

Golnaz Badkobeh[1] and Maxime Crochemore[1,2]

[1] King's College London, London, UK
[2] Université Paris-Est, France
Maxime.Crochemore@kcl.ac.uk

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 sketch of the new 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.

Keywords: combinatorics on words, repetitions, word morphisms.

1 Introduction

Repetitions in words is a basic question in Theoretical Informatics, certainly because it is related to many applications although it is first been studied by Thue at the beginning of the twentieth century [10] 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 helps 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 findings is that an infinite binary word can avoid factors with an exponent larger than 2, called 2^+-powers. This has been extended by Dejean [2] 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 [8] and references therein).

Another constraint is considered by Fraenkel and Simpson [3]: 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. Indeed all factors of exponent at least 2 occurring in their word should be considered, which adds 2 cubes. Their proof uses a pair of morphisms, one morphism to get an infinite string by iteration, the other morphism to produce the final translation on the binary alphabet. Their result has been proved with different pairs of morphism by Rampersad et al. [7] (the first morphism is uniform), by Harju and Nowotka [4] (the second morphism accepts any infinite square-free word), and by Badkobeh et al. [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

Golnaz Badkobeh, Maxime Crochemore: Bounded Number of Squares in Infinite Repetition-Constrained Binary Words, pp. 161–166 .
Proceedings of PSC 2010, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04597-8 © Czech Technical University in Prague, Czech Republic

is the smallest possible while containing the smallest number 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 [5] that if an infinite binary 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 [9] has built an infinite binary word avoiding $7/3^+$-powers and all squares of period at least 7. His word contains more than 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 corresponds to a uniquely-decipherable code but admits a unique decoding on the words produced by the first.

2 Repetitions in binary words

A word is a sequence of letters drawn from a finite alphabet. We consider the binary alphabet $B = \{\mathtt{0}, \mathtt{1}\}$, 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. Indeed, it is 3 if it contains no square (e.g. 010), 7 if it contains 1 square (e.g. 0001000), and 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 $\mathbf{f} = f(\mathtt{a})^\infty$ is square-free (see [6, Chapter 2]). It can additionally be checked that all square-free words of length 3 occur in $\mathbf{f}$ except aba and cbc. The morphism h_0 is from A_3^* to B^* and defined by

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

This morphism is not uniform but the three codewords form a uniquely-decipherable code. Then the above 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 [5], the exponent has to go up to 7/3 to allow the property.

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

Theorem 2. *There exist 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 $\texttt{0}^2$, $\texttt{1}^2$, $(\texttt{01})^2$, $(\texttt{10})^2$, $(\texttt{001})^2$, $(\texttt{010})^2$, $(\texttt{011})^2$, $(\texttt{100})^2$, $(\texttt{101})^2$, $(\texttt{110})^2$, $(\texttt{01101001})^2$, $(\texttt{10010110})^2$, and the two words `0110110` and `1001001` of exponent 7/3.

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 $A_6^* = \{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}, \texttt{f}\}^*$ to itself by:

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

It can be shown that the morphism is weakly square-free in the sense that $\mathbf{g} = g^\infty(\texttt{a})$ 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 $g(\texttt{cf}) = \texttt{eabdfabdf}$ contains the square $(\texttt{abdf})^2$. Moreover there is no known characterisation of weakly square-free morphisms defined on more than three letters (unless of course if only three letters occur in the infinite word).

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 `abac` in $g(w)$, for $w \in A_6^*$, uniquely corresponds to an occurrence of `a` in w.

Lemma 3. *The set of doublets occurring in* $\mathbf{g}$ *is*

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

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

Lemma 4.
The set of triplets in $g^\infty(\texttt{a})$ *is*

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

Proof. Triplets appear in the images of a letter of a doublet. Found in images of one letter are: `aba`, `abd`, `ace`, `bab`, `bac`, `bdf`, `eab`, `fba`. The images of doublets occurring in $\mathbf{g}$, in set D of Lemma 3, contain the extra triplets: `acb`, `bda`, `cba`, `cea`, `dab`, `dfb`. □

To prove the infinite word $g^\infty(\mathtt{a})$ is square-free first we discard squares containing less than four occurrences of the word $g(\mathtt{a}) = \mathtt{abac}$, Then squares containing at least four. The word abac is chosen because its occurrences in $g^\infty(\mathtt{a})$ correspond to $g(\mathtt{a})$ only, so they are used to synchronise the parsing of the word according to the codewords $g(a)$'s.

Lemma 5. *No square in $g^\infty(\mathtt{a})$ can contain less than four occurrences of* abac*.*

Proof. Assume by contradiction that a square ww in $g^\infty(\mathtt{a})$ 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 $g^\infty(\mathtt{a})$ that belongs to the set $\mathtt{a}((A_6 \setminus \{\mathtt{a}\})^*\mathtt{a})^4$. Since two consecutive occurrences of a in $g^\infty(\mathtt{a})$ are separated by a string of length at most 4 (the largest such string is indeed bdfb as a consequence of Lemma 3), 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 6. *No square in $g^\infty(\mathtt{a})$ can contain at least four occurrences of* abac*.*

Table 1. Gaps: words between consecutive occurrences of abac in $g^\infty(\mathtt{a})$. They are images of gaps between consecutive occurrences of a.

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

Proof (Sketch). The complete proof is by contradiction: let k be the maximal integer k for which $g^k(\mathtt{a})$ is square-free and let ww be a square occurring in $g^{k+1}(\mathtt{a})$. 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. □

Corollary 7. *The infinite word $g^\infty(\mathtt{a})$ 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 square-free words 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 $g(\mathtt{ae}) = \mathtt{10011} \cdot \mathtt{0110} =$

$\texttt{1001} \cdot \texttt{10110} = g(\texttt{fd})$. However, parsing the word $h(y)$ when y is a factor of $g^{\infty}(\texttt{a})$ is unique due to the absence of some doublets in it (see Lemma 3). For example fd does not occur, which induces the unique parsing of 100110110 as $\texttt{10011} \cdot \texttt{0110}$.

Proposition 8. *The infinite word* $\mathbf{h} = h(g^{\infty}(\texttt{a}))$ *contains no factor of exponent larger than* 7/3. *It contains the* 12 *squares* 0^2, 1^2, $(01)^2$, $(10)^2$, $(001)^2$, $(010)^2$, $(011)^2$, $(100)^2$, $(101)^2$, $(110)^2$, $(01101001)^2$, $(10010110)^2$ *only. Words 0110110 and 1001001 are the only factors with an exponent larger than* 2.

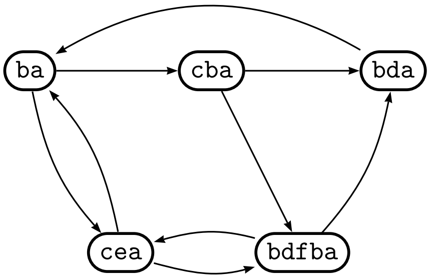

Figure 1. Graph showing immediate successors of gaps in the word $g^{\infty}(\texttt{a})$: a suffix of it following an occurrence of a is the label of an infinite path.

The proof is far beyond this extended abstract. It is based on the fact that occurrences of 10011 in **h** identify occurrences of a in **g** and on the unique parsing mentioned above. It proceeds by considering several cases according to the gaps between consecutive occurrences of a, which leads to analyse paths in the graph of Figure 1.

5 Conclusion

The constraint on the number squares imposed on binary words slightly differs from the constraint considered by Shallit [9]. The squares occurring in his word have period smaller than 8. Our word contains less squares but their maximal period is 8. Indeed it is impossible to have both constrains simultaneously for an infinite binary strings.

Looking at repetitions in words on larger alphabets, the subject introduces a new type of threshold, that we call the *bounded-repetitions threshold.* For the alphabet of a letters, it is defined as the smallest rational number t_a for which there exist an infinite word avoiding t^{+}-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 $t_2 = 7/3$. Values for larger alphabets remains to explore.

References

1. G. Badkobeh and M. Crochemore: *An infinite binary word containing only three distinct squares*, 2010, submitted.
2. F. Dejean: *Sur un théorème de Thue.* J. Comb. Theory, Ser. A, 13(1) 1972, pp. 90–99.
3. A. S. Fraenkel and J. Simpson: *How many squares must a binary sequence contain?* Electr. J. Comb., 2 1995.
4. T. Harju and D. Nowotka: *Binary words with few squares.* Bulletin of the EATCS, 89 2006, pp. 164–166.

5. J. Karhumäki and J. Shallit: *Polynomial versus exponential growth in repetition-free binary words.* J. Comb. Theory, Ser. A, 105(2) 2004, pp. 335–347.
6. M. Lothaire, ed., *Combinatorics on Words*, Cambridge University Press, second ed., 1997.
7. N. Rampersad, J. Shallit, and M. wei Wang: *Avoiding large squares in infinite binary words.* Theor. Comput. Sci., 339(1) 2005, pp. 19–34.
8. M. Rao: *Last cases of Dejean's conjecture*, in WORDS 2009, A. Carpi and C. de Felice, eds., University of Salerno, Italy, 2009.
9. J. Shallit: *Simultaneous avoidance of large squares and fractional powers in infinite binary words.* Int'l. J. Found. Comput. Sci., 15 2004, pp. 317–327.
10. A. Thue: *Über unendliche Zeichenreihen.* Norske vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana, 7 1906, pp. 1–22.