Cubefree Binary Words Avoiding Long Squares

Narad Rampersad, Jeffrey Shallit, and Ming-wei Wang
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
CANADA
nrampersad@math.uwaterloo.ca
shallit@graceland.uwaterloo.ca
m2wang@math.uwaterloo.ca

Abstract

Entringer, Jackson, and Schatz conjectured in 1974 that every infinite cubefree binary word contains arbitrarily long squares. In this paper we show this conjecture is false: there exist infinite cubefree binary words avoiding all squares xx with $|x| \geq 4$, and the number 4 is best possible. However, the Entringer-Jackson-Schatz conjecture is true if "cubefree" is replaced with "overlap-free".

1 Introduction

Let Σ be a finite nonempty set, called an *alphabet*. We consider finite and infinite words over Σ. The set of all finite words is denoted by Σ^*. The set of all infinite words (that is, maps from $\mathbb{N}$ to Σ) is denoted by Σ^ω.

A *morphism* is a map $h : \Sigma^* \to \Delta^*$ such that $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma^*$. A morphism may be specified by providing the image words $h(a)$ for all $a \in \Sigma$. If $h : \Sigma^* \to \Sigma^*$ and $h(a) = ax$ for some letter $a \in \Sigma$, then we say that h is *prolongable* on a, and we can then iterate h infinitely often to get the fixed point $h^\omega(a) := a\, x\, h(x)\, h^2(x)\, h^3(x) \cdots$.

A *square* is a nonempty word of the form xx, as in the English word `murmur`. A *cube* is a nonempty word of the form xxx, as in the English sort-of-word `shshsh`. An *overlap* is a word of the form $axaxa$, where x is a possibly empty word and a is a single letter, as in the English word `alfalfa`.

It is well-known and easily proved that every word of length 4 or more over a two-letter alphabet contains a square as a subword. However, Thue proved in 1906 [4] that there exist infinite words over a three-letter alphabet that contain no squares; such words are said to *avoid* squares or be *squarefree*. Thue also proved that the word $\mu^\omega(0) = 0110100110010110\cdots$ is overlap-free (and hence cubefree); here μ is the morphism sending $0 \to 01$ and $1 \to 10$.

Entringer, Jackson, and Schatz [2] proved that while squares cannot be avoided over a two-letter alphabet, arbitrarily long squares can. More precisely, they proved that there exist infinite binary words with no squares of length ≥ 3, and that the number 3 is best possible. Later, this result was improved by Fraenkel and Simpson [3], who proved that there exist infinite binary words where the only squares are 00, 11, and 0101.

Entringer, Jackson, and Schatz conjectured in 1974 that any infinite cubefree word over $\{0, 1\}$ contains arbitrarily long squares [2, Conjecture B, p. 163]. In this paper we show that this conjecture is false; there exist infinite cubefree binary words with no squares xx with $|x| \geq 4$. The number 4 is best possible. Further, we show that the Entringer-Jackson-Schatz conjecture is true if the word "cubefree" is replaced with "overlap-free".

2 A cubefree word without arbitrarily long squares

In this section we disprove the conjecture of Entringer, Jackson, and Schatz. First we prove the following result.

Theorem 1 *There is a squarefree infinite word over* $\{0, 1, 2, 3\}$ *with no occurrences of the subwords* 12*,* 13*,* 21*,* 32*,* 231*, or* 10302*.*

Proof. Let the morphism h be defined by

$$\begin{aligned} 0 &\to 0310201023 \\ 1 &\to 0310230102 \\ 2 &\to 0201031023 \\ 3 &\to 0203010201 \end{aligned}$$

Then we claim the fixed point $h^\omega(0)$ has the desired properties.

First, we claim that if $w \in \{0, 1, 2, 3\}^*$ then $h(w)$ has no occurrences of 12, 13, 21, 32, 231, or 10302. For if any of these words occur as subwords of

$h(w)$, they must occur within some $h(a)$ or straddling the boundary between $h(a)$ and $h(b)$, for some single letters a, b. They do not; this easy verification is left to the reader.

Next, we prove that if w is any squarefree word over $\{0, 1, 2, 3\}$ having no occurrences of 12, 13, 21, or 32, then $h(w)$ is squarefree.

We argue by contradiction. Let $w = a_1 a_2 \cdots a_n$ be a squarefree string such that $h(w)$ contains a square, i.e., $h(w) = xyyz$ for some $x, z \in \{0, 1, 2, 3\}^*$, $y \in \{0, 1, 2, 3\}^+$. Without loss of generality, assume that w is a shortest such string, so that $0 \leq |x|, |z| < 10$.

Case 1: $|y| \leq 20$. In this case we can take $|w| \leq 5$. To verify that $h(w)$ is squarefree, it therefore suffices to check each of the 49 possible words $w \in \{0, 1, 2, 3\}^5$ to ensure that $h(w)$ is squarefree in each case.

Case 2: $|y| > 20$. First, we establish the following result.

Lemma 2 *(a) Suppose $h(ab) = th(c)u$ for some letters $a, b, c \in \{0, 1, 2, 3\}$ and strings $t, u \in \{0, 1, 2, 3\}^*$. Then this inclusion is trivial (that is, $t = \epsilon$ or $u = \epsilon$) or u is not a prefix of $h(d)$ for any $d \in \{0, 1, 2, 3\}$.*

(b) Suppose there exist letters a, b, c and strings s, t, u, v such that $h(a) = st$, $h(b) = uv$, and $h(c) = sv$. Then either $a = c$ or $b = c$.

Proof.

(a) This can be verified with a short computation. In fact, the only a, b, c for which the equality $h(ab) = th(c)u$ holds nontrivially is $h(31) = th(2)u$, and in this case $t = 020301$, $u = 0102$, so u is not a prefix of any $h(d)$.

(b) This can also be verified with a short computation. If $|s| \geq 6$, then no two distinct letters share a prefix of length 6. If $|s| \leq 5$, then $|t| \geq 5$, and no two distinct letters share a suffix of length 5.

■

For $i = 1, 2, \ldots, n$ define $A_i = h(a_i)$. Then if $h(w) = xyyz$, we can write

$$h(w) = A_1 A_2 \cdots A_n = A_1' A_1'' A_2 \cdots A_{j-1} A_j' A_j'' A_{j+1} \cdots A_{n-1} A_n' A_n''$$

where

$$\begin{aligned}
A_1 &= A_1' A_1'' \\
A_j &= A_j' A_j'' \\
A_n &= A_n' A_n'' \\
x &= A_1' \\
y &= A_1'' A_2 \cdots A_{j-1} A_j' = A_j'' A_{j+1} \cdots A_{n-1} A_n' \\
z &= A_n'',
\end{aligned}$$

where $|A_1''|, |A_j''| > 0$. See Figure 1.

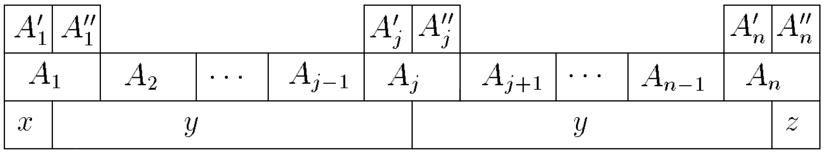

Figure 1: The string $xyyz$ within $h(w)$

If $|A_1''| > |A_j''|$, then $A_{j+1} = h(a_{j+1})$ is a subword of $A_1'' A_2$, hence a subword of $A_1 A_2 = h(a_1 a_2)$. Thus we can write $A_{j+2} = A_{j+2}' A_{j+2}''$ with

$$A_1'' A_2 = A_j'' A_{j+1} A_{j+2}'.$$

See Figure 2.

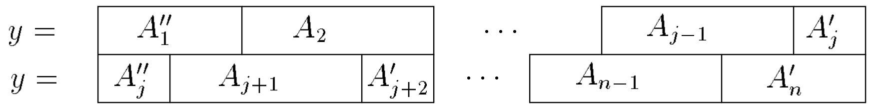

Figure 2: The case $|A_1''| > |A_j''|$

But then, by Lemma 2 (a), either $|A_j''| = 0$, or $|A_1''| = |A_j''|$, or A_{j+2}' is a not a prefix of any $h(d)$. All three conclusions are impossible.

If $|A_1''| < |A_j''|$, then $A_2 = h(a_2)$ is a subword of $A_j'' A_{j+1}$, hence a subword of $A_j A_{j+1} = h(a_j a_{j+1})$. Thus we can write $A_3 = A_3' A_3''$ with

$$A_1'' A_2 A_3' = A_j'' A_{j+1}.$$

See Figure 3.

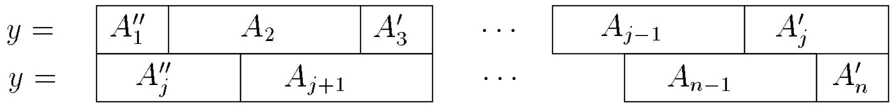

Figure 3: The case $|A_1''| < |A_j''|$

By Lemma 2 (a), either $|A_1''| = 0$ or $|A_1''| = |A_j''|$ or A_3' is not a prefix of any $h(d)$. Again, all three conclusions are impossible.

Therefore $|A_1''| = |A_j''|$. Hence $A_1'' = A_j''$, $A_2 = A_{j+1}$, …, $A_{j-1} = A_{n-1}$, and $A_j' = A_n'$. Since h is injective, we have $a_2 = a_{j+1}, \ldots, a_{j-1} = a_{n-1}$. It also follows that $|y|$ is divisible by 10 and $A_j = A_j' A_j'' = A_n' A_1''$. But by Lemma 2 (b), either (1) $a_j = a_n$ or (2) $a_j = a_1$. In the first case, $a_2 \cdots a_{j-1} a_j = a_{j+1} \cdots a_{n-1} a_n$, so w contains the square $(a_2 \cdots a_{j-1} a_j)^2$, a contradiction. In the second case, $a_1 \cdots a_{j-1} = a_j a_{j+1} \cdots a_{n-1}$, so w contains the square $(a_1 \cdots a_{j-1})^2$, a contradiction.

It now follows that the infinite word

$$h^\omega(0) = 03102010230203010201031023010203102010230201031023\cdots$$

is squarefree and contains no occurrences of 12, 13, 21, 32, 231, or 10302. ■

Theorem 3 *Let* $\mathbf{w}$ *be any infinite word satisfying the conditions of Theorem 1. Define a morphism* g *by*

$$\begin{aligned} 0 &\rightarrow 010011 \\ 1 &\rightarrow 010110 \\ 2 &\rightarrow 011001 \\ 3 &\rightarrow 011010 \end{aligned}$$

Then $g(\mathbf{w})$ *is a cubefree word containing no squares* xx *with* $|x| \geq 4$*.*

Before we begin the proof, we remark that all the words 12, 13, 21, 32, 231, 10302 must indeed be avoided, because

$$\begin{aligned} g(12) &\quad \text{contains the squares } (0110)^2,\ (1100)^2,\ (1001)^2 \\ g(13) &\quad \text{contains the square } (0110)^2 \\ g(21) &\quad \text{contains the cube } (01)^3 \\ g(32) &\quad \text{contains the square } (1001)^2 \\ g(231) &\quad \text{contains the square } (10010110)^2 \\ g(10302) &\quad \text{contains the square } (100100110110)^2. \end{aligned}$$

Proof. The proof parallels the proof of Theorem 1. Let $w = a_1 a_2 \cdots a_n$ be a squarefree string, with no occurrences of 12, 13, 21, 32, 231, or 10302. We first establish that if $g(w) = xyyz$ for some $x, z \in \{0,1,2,3\}^*$, $y \in \{0,1,2,3\}^+$, then $|y| \leq 3$. Without loss of generality, assume w is a shortest such string, so $0 \leq |x|, |z| < 6$.

Case 1: $|y| \leq 12$. In this case we can take $|w| \leq 5$. To verify that $g(w)$ contains no squares yy with $|y| \geq 4$, it suffices to check each of the 41 possible words $w \in \{0,1,2,3\}^5$.

Case 2: $|y| > 12$. First, we establish the analogue of Lemma 2.

Lemma 4 *(a) Suppose $g(ab) = tg(c)u$ for some letters $a, b, c \in \{0,1,2,3\}$ and strings $t, u \in \{0,1,2,3\}^*$. Then this inclusion is trivial (that is, $t = \epsilon$ or $u = \epsilon$) or u is not a prefix of $g(d)$ for any $d \in \{0,1,2,3\}$.*

(b) Suppose there exist letters a, b, c and strings s, t, u, v such that $g(a) = st$, $g(b) = uv$, and $g(c) = sv$. Then either $a = c$ or $b = c$, or $a = 2$, $b = 1$, $c = 3$, $s = 0110$, $t = 01$, $u = 0101$, $v = 10$.

Proof.

(a) This can be verified with a short computation. The only a, b, c for which $g(ab) = tg(c)u$ holds nontrivially are

$$\begin{aligned} g(01) &= 010\ g(3)\ 110 \\ g(10) &= 01\ g(2)\ 0011 \\ g(23) &= 0110\ g(1)\ 10. \end{aligned}$$

But none of 110, 0011, 10 are prefixes of any $g(d)$.

(b) If $|s| \geq 5$ then no two distinct letters share a prefix of length 5. If $|s| \leq 3$ then $|t| \geq 3$, and no two distinct letters share a suffix of length 3. Hence $|s| = 4$, $|t| = 2$. But only $g(2)$ and $g(3)$ share a prefix of length 4, and only $g(1)$ and $g(3)$ share a suffix of length 2.

■

The rest of the proof is exactly parallel to the proof of Theorem 1, with the following exception. When we get to the final case, where $|y|$ is divisible by 6, we can use Lemma 4 to rule out every case except where $x = 0101$, $z = 01$, $a_1 = 1$, $a_j = 3$, and $a_n = 2$. Thus $w = 1\alpha 3\alpha 2$ for some string $\alpha \in \{0,1,2,3\}^*$. This special case is ruled out by the following lemma:

Lemma 5 *Suppose $\alpha \in \{0, 1, 2, 3\}^*$, and let $w = 1\alpha 3\alpha 2$. Then either w contains a square, or w contains an occurrence of one of the subwords 12, 13, 21, 32, 231, or 10302.*

Proof. This can be verified by checking (a) all strings w with $|w| \leq 4$, and (b) all strings of the form $w = abcw'de$, where $a, b, c, d, e \in \{0, 1, 2, 3\}$ and $w' \in \{0, 1, 2, 3\}^*$. (Here w' may be treated as an indeterminate.) ■

It now remains to show that if w is squarefree and contains no occurrence of 12, 13, 21, 32, 231, or 10302, then $g(w)$ is cubefree. If $g(w)$ contains a cube yyy, then it contains a square yy, and from what precedes we know $|y| \leq 3$. It therefore suffices to show that $g(w)$ contains no occurence of 0^3, 1^3, $(01)^3$, $(10)^3$, $(001)^3$, $(010)^3$, $(011)^3$, $(100)^3$, $(101)^3$, $(110)^3$. The longest such string is of length 9, so it suffices to examine the 16 possibilities for $g(w)$ where $|w| = 3$. This is left to the reader.

The proof of Theorem 3 is now complete. ■

Corollary 6 *If g and h are defined as above, then*

$$g(h^\omega(0)) = 0100110110100101100100110110010100110101100100110110010110 10\cdots$$

is cubefree, and avoids all squares xx with $|x| \geq 4$.

3 The constant 4 is best possible

It is natural to wonder if the constant 4 in Corollary 6 can be improved. It cannot, as the following theorem shows.

Theorem 7 *Every binary word of length ≥ 30 contains a cube or a square xx with $|x| \geq 3$.*

Proof. This may be proved purely mechanically. More generally, let $P \subset \Sigma^*$ be a set of subwords to be avoided. We create and traverse a certain tree T, as follows. The root of the tree is labeled ϵ. If a node is labeled x and contains no subword in P, then it has children labeled xa for each $a \in \Sigma$; otherwise it is a leaf of T. This tree is infinite if and only if there is an infinite word avoiding the elements of P.

If T is finite, then the height of T gives the length l such that every word of length l or greater contains an element of P. The tree can be created and traversed using a queue and breadth-first search.

If the set P is symmetric under renaming of the letters—as it is in this case—we may further improve the procedure by labeling the root with any particular letter, say 0. When we run this procedure on the statement of the theorem, we obtain a tree with 289 leaves, the longest being of length 30. The unique string of length 29 starting with 0 and avoiding cubes and squares xx with $|x| \geq 3$ is 00110010100110101100101001100. ■

4 Overlap-free words contain arbitrarily long squares

It is also natural to wonder if a result like Corollary 6 holds if "cubefree" is replaced with "overlap-free". It does not, as the following result shows.

Theorem 8 *Any infinite overlap-free word over* $\{0, 1\}$ *contains arbitrarily long squares.*

Proof. By [1, Lemma 3] we know that if $\mathbf{x}$ is an overlap-free infinite word over $\{0, 1\}$, then there exist a word $u \in \{\epsilon, 0, 1, 00, 11\}$ and an overlap-free infinite word $\mathbf{y}$ such that $\mathbf{x} = u\mu(\mathbf{y})$, where μ is the Thue-Morse morphism. By iterating this theorem, we get that every overlap-free infinite word must contain $\mu^n(0)$ for arbitrarily large n; hence contains arbitrarily long squares. ■

5 Acknowledgments

We thank Jean-Paul Allouche for helpful discussions.

References

[1] J.-P. Allouche, J. Currie, and J. Shallit. Extremal infinite overlap-free binary words. *Electronic J. Combinatorics* **5** (1998), #R27.

[2] R. C. Entringer, D. E. Jackson, and J. A. Schatz. On nonrepetitive sequences. *J. Combin. Theory. Ser. A* **16** (1974), 159–164.

[3] A. S. Fraenkel and J. Simpson. How many squares must a binary sequence contain? *Electronic J. Combinatorics* **2** (1995), #R2.

[4] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.