Squarefree words with interior disposable factors

Marko Milosevic and Narad Rampersad
Department of Mathematics and Statistics
University of Winnipeg
n.rampersad@uwinnipeg.ca

July 8, 2020

Abstract

We give a partial answer to a problem of Harju by constructing an infinite ternary squarefree word $w$ with the property that for every $k \geq 3312$ there is an interior length-$k$ factor of $w$ that can be deleted while still preserving squarefreeness. We also examine Thue’s famous squarefree word (generated by iterating the map $0 \rightarrow 012$, $1 \rightarrow 02$, $2 \rightarrow 1$) and characterize the positions $i$ for which deleting the symbol appearing at position $i$ preserves squarefreeness.

1 Introduction

The study of squarefree words (words avoiding non-empty repetitions $xx$) is a fundamental topic in combinatorics on words. Thue [17] was the first to construct an infinite squarefree word on three symbols. Recently, Harju [11] defined an interesting class of squarefree word: irreducibly squarefree words. A squarefree word is irreducibly squarefree if the deletion of any letter in the word, other than the first and last letters, produces an occurrence of a square. Harju showed that there exist ternary irreducibly squarefree words of all sufficiently large lengths. Harju’s notion of irreducibly squarefree words was inspired by a similar concept introduced by Grytczuk, Kordulewski, and Niewiadomski [9], who defined extremal squarefree words as follows: a squarefree word is extremal if every possible insertion of a symbol into the word creates an occurrence of a square.

Harju posed three open problems in his paper. We give a partial answer to his third problem here by constructing an infinite squarefree word $w$ with the property that for every $k \geq 3312$ there is an interior (i.e., not a prefix) length-$k$ factor of $w$ that can be deleted while still preserving squarefreeness. We also examine Thue’s famous squarefree word (generated by iterating the map $0 \rightarrow 012$, $1 \rightarrow 02$, $2 \rightarrow 1$) [18] and characterize the positions $i$ for which deleting the symbol appearing at position $i$ preserves squarefreeness.

*The second author is supported by NSERC Discovery Grant 2019-04111.
2 Preliminaries

Let $A$ be a finite alphabet of letters. For a word $w$ over $A$ (i.e., $w \in A^*$), let $|w|$ denote its length. A word $u$ is a factor of $w$, if $w = xuy$ where $x$ and/or $y$ may be empty. If $x$ ($y$, resp.) is empty then $u$ is a prefix (a suffix, resp.) of $w$.

A square is a non-empty word of the form $u^2 = uu$. A finite or infinite word $w$ over $A$ is squarefree if it does not have any square factors. A position $i$ in a squarefree word $w$ is said to be disposable if $w = uvx$ and $i \in A$, $|w| = i$, and the word $uv$ is squarefree. If $w = uxv$ and $uv$ are squarefree, then $x$ is a disposable factor of $w$.

A morphism $h: A^* \to A^*$ is said to be squarefree, if it preserves squarefreeness of words, i.e., if $h(w)$ is squarefree for all squarefree words $w$. A morphism $h: A^* \to A^*$ is uniform if the images $h(a)$ have the same length: $|h(a)| = n$ for all $a \in A$ and for some positive $n$ called the length of $h$.

An infinite word $w$ is a fixed point of a morphism $h$ if $h(w) = w$. This happens if $w$ begins with the letter $a$, and $w$ is obtained by iterating $h$ on the first letter $a$ of $w$: $h(a) = au$ and $w = auh(u)h^2(u) \cdots$. In this case we denote the fixed point $w$ by $h^\omega(a)$.

Let $T$ be the ternary alphabet $T = \{0, 1, 2\}$. Let $\tau: T^* \to T^*$ be the morphism defined by

\[
\tau(0) = 012, \quad \tau(1) = 02 \quad \text{and} \quad \tau(2) = 1. \tag{1}
\]

The word obtained by iterating $\tau$ on $0$ gives the following infinite squarefree word:

\[
vtm = 012021012102012 \cdots (= \tau^\omega(0)).
\]

(Here we follow [4] in using $vtm$, for variant of the Thue–Morse word, to denote this word.) For the next basic result, see [6, 10, 13, 15, 18] and [14]:

Lemma 1. The word $vtm$ is squarefree and it does not contain 010 or 212 as factors.

Note that this lemma implies that $vtm$ does not contain 1021 or 1201 as factors, since the only way these could arise are as factors of $\tau(212) = 1021$ or $\tau(00) = 012012$.

3 Disposable positions in $vtm$

Here is a list of the first few disposable positions in $vtm$:

\[
(0, 2, 12, 18, 44, 50, 60, 66, 76, 82, 108, 114, 140, 146, 172, 178, 188, 194, 204, \ldots)
\]

and here is a list of the first few first differences of the above sequence:

\[
(2, 10, 6, 26, 6, 10, 6, 26, 6, 26, 6, 10, 6, 26, 6, 10, 6, 10, 6, 10, 6, 10, 6, 26, \ldots).
\]

The goal of this section is to give an exact description of these two sequences.

The letter in position 0 of $vtm$ is trivially disposable. The disposability of the 2 in position 2 of $vtm$ is easy to verify. Deleting this letter produces a new word $01w'$, where $w'$ is a squarefree suffix of $vtm$. If a square occurs, it must begin from the first letter 0, or the second letter 1. In either case, a factor 1021 or 010 is found in the first half of the square. Since $vtm$ avoids both factors, this leads to a contradiction, so we conclude that this occurrence of 2 is disposable.
Theorem 2. The second and fourth occurrences of 0 in a factor $\tau(10121) = 0201202102$ of $\text{vtm}$ are disposable in $\text{vtm}$.

Proof. Every disposable letter divides a squarefree word into a squarefree prefix consisting of all letters before the disposable one, and a squarefree suffix consisting of all letters afterwards. For the remainder of this proof, let $u$ denote the aforementioned prefix, and $v$ the corresponding suffix. Then $uv$ is the word obtained by deleting the disposable letter.

We first notice that every factor $\tau(10121)$ is enclosed by $\tau(02)$. That is, $\tau(10121)$ always occurs in the context $\tau(02)\tau(10121)\tau(02) = 01202012021020120120210201202102012020120201202$. After deleting the second 0 in $\tau(10121)$, we have $uv = w'0121021021w''$, where $u = w'012102$ and $v = 12021w''$. Since any potential square would have to start in $u$ and end in $v$, to ensure that every occurrence of this 0 is disposable, we must verify that, no matter which letter in $u$ we start with, no square occurs.

It is clear that 0121021021 (and hence $uv$) is squarefree for all squares of length at most six, and any square in $uv$ of length at least eight that crosses the boundary between $u$ and $v$ contains a factor 212 or 1021 in each half of the square, which is a contradiction, since $\text{vtm}$ avoids 212 and 1021. The only possible exception is a square of the form $xx = 1y1021y102$, where $y \in T^*$ and $x = 1y102$ is both a prefix of $v$ and a suffix of $u$. Now, the square $xx$ was produced by deleting the 0 from $x0x$ which implies that 0 does not immediately follow the square. The only other option is 1. This gives us $xx1 = 1y1021y1021$; a contradiction again since $\text{vtm}$ avoids 1021.

The argument for the fourth 0 in $\tau(10121)$ is similar.

Theorem 3. The first and third occurrences of 2 in a factor $\tau(12101) = 0210201202$ of $\text{vtm}$ are disposable in $\text{vtm}$.

Proof. We proceed in a similar manner to Theorem 2. Let $uv$ be as described in the proof of that theorem. First, we have that $\tau(12101)$ only occurs in the context $\tau(20)\tau(12101)\tau(20) = 1012021020120201202012020120201202$. Deleting the first 2 in $\tau(12101)$ gives us $uv = w'10120102012w''$, where $u = w'10120$ and $v = 102012w''$. Indeed, the factor 10120102012 (and hence $uv$) avoids squares of length at most six. So any potential square in $uv$ must have length at least eight, but then it contains 010 or 1201 in each half of the square, which is not possible, since $\text{vtm}$ does not contain 010 or 1201. However, we must consider the exception $xx = 1y1201y120$, where $y \in T^*$ and $x = 1y120$ is a prefix of $v$ and a suffix of $u$. Then the deletion of 2 from $x2x$ implies that a 1 must occur immediately after the square, giving us $xx1 = 1y1201y1201$, which is a contradiction, since $\text{vtm}$ does not contain 1201.

The argument for the third 2 in $\tau(12101)$ is similar.

Theorems 2 and 3 only show that certain positions in $\text{vtm}$ are disposable and do not indicate which positions are not disposable. We can completely characterize the disposable positions in $\text{vtm}$ using the computer program Walnut [16].
The word vtm is a 2-automatic sequence (see [3]) and is generated by the automaton in Figure 1 (the labels of each state indicate the output associated with that state).

Since vtm is an automatic sequence, we can use Walnut [16] to verify that it has certain combinatorial properties. The following Walnut command computes the set of disposable positions $j$ in vtm. The output automaton is given in Figure 2.

eval dispo_pos "?msd_2 Ai,n (i < j & j < i+2*n) => (Ek i <= k & ((j < i+n & k <= i+n) | (j >= i+n & k < i+n)) & (((j < k | j > k+n) & VTM[k] != VTM[k+n]) | ((k < j & j <= k+n) & VTM[k] != VTM[k+n+1])))"

![Figure 1: 2-DFAO for vtm](image1)

The next command computes the set of values taken by the first difference sequence of the sequence of disposable positions in vtm (excluding the initial position). The output automaton is given in Figure 3.

eval dispo_delta "?msd_2 Ei,j i >=2 & j > i & j = i+1 & $dispo_pos(i) & $dispo_pos(j) & (Ak (i<k & k<j) => ~$dispo_pos(k))"

![Figure 2: dispo_pos output automaton](image2)

The next command computes the set of values taken by the first difference sequence of the sequence of disposable positions in vtm (excluding the initial position). The output automaton is given in Figure 3.

eval dispo_delta "?msd_2 Ei,j i >=2 & j > i & j = i+1 & $dispo_pos(i) & $dispo_pos(j) & (Ak (i<k & k<j) => ~$dispo_pos(k))"

![Figure 2: dispo_pos output automaton](image3)

From Figure 3, we see that the “gaps” between disposable positions (excluding the initial position) in vtm are 6, 10, and 26. We next consider the density of the disposable positions.
Figure 3: dispo\_delta output automaton

Figure 4: Plot of $|D_{vtm}(n)|/n$

in $vtm$. Let $D_w(n)$ denote the set of disposable positions $\leq n + 1$ of an infinite squarefree word $w$. Figure 4 shows a plot of the initial values of $|D_{vtm}(n)|/n$.

We would like to determine the quantity $\lim_{n \to \infty} |D_{vtm}(n)|/n$ if it exists, or failing that, the quantities $\lim\inf_{n \to \infty} |D_{vtm}(n)|/n$ and $\lim\sup_{n \to \infty} |D_{vtm}(n)|/n$. There is a general method for this due to Bell [2]; however, due to the structure of the automaton in Figure 2, we are able to employ simpler techniques.

**Theorem 4.** The density of disposable positions in $vtm$ is

$$\lim_{n \to \infty} |D_{vtm}(n)|/n = 1/12 = 0.08\bar{3}.$$

**Proof.** Let $M$ be the adjacency matrix of the automaton given in Figure 2 restricted to just the states 1 to 8. That is, let $M$ be the $8 \times 8$ matrix whose $ij$-entry is equal to the number
of transitions from state $i$ to state $j$. We have

$$M = \begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}.$$ 

We can verify that $M^5$ is a positive matrix, which implies that $M$ is a *primitive* matrix. The Perron–Frobenius eigenvalue of $M$ is 2, and it follows from the standard Perron–Frobenius theory (see [1, Chapter 8] for a treatment formulated in terms of morphisms rather than automata) that the $i$-th entry of the left eigenvector

$$v = (1/12, 1/24, 1/6, 1/8, 1/4, 1/12, 5/24, 1/24)$$

of $M$, where we have normalized $v$ so that its entries sum to 1, gives the fraction of all input strings that reach state $i$. Since the final states of the automaton in Figure 2 are 2 and 8, it follows that the fraction of all input strings that reach a final state is the sum of the entries of $v$ in positions 2 and 8. We conclude that the frequency of disposable positions in $v_{tm}$ is $1/24 + 1/24 = 1/12 = 0.083$. 

4 Words with longer interior disposable factors

Harju [11, Problem 3] asked if there exists an infinite ternary squarefree word containing interior disposable factors of every length $k \geq 1$. We are unable to prove that this is the case, but we can prove the following weaker result.

**Theorem 5.** There exists an infinite ternary squarefree word containing interior disposable factors of every length $k \geq 3312$.

**Proof.** Let $A = \{0, 1\}$, $B = \{2, 3, 4\}$, and $C = A \cup B$. Fraenkel and Simpson [8] showed that there exists an infinite word over $A$ whose only square factors are 00, 11, and 0101 (see [12] for a simpler construction). Let $x$ be any such infinite word and let $y$ be any infinite squarefree word over $B$. We are going to start by constructing an infinite squarefree word $v$ over $C$ by interleaving the words $x$ and $y$ in a very particular way.

For $n \geq 1$, let $Q(n)$ denote any word of the form $aY_1bY_2cY_3d$ where

- $a, b, c, d \in A$,
- $b \neq c$,
- $abcd \neq 0101$,
- $Y_i \in B^*$, and
We now construct $v$ as follows:

- Taking one symbol at a time, interleave $x$ and $y$ in the following way,
\[ x_0y_0x_1y_1x_2y_2 \cdots x_iy_i+1x_i+1y_i+2y_i+2x_i+3 \cdots \]
until we get 414 occurrences of $Q(1)$.

- Then start taking two symbols at a time from $y$, so at a certain point we have
\[ \cdots x_jy_jy_j+1y_j+1y_j+2y_j+3x_j+2y_j+4y_j+5x_j+3 \cdots \]
until we get 414 occurrences of $Q(2)$.

- Then take three symbols at a time from $y$:
\[ \cdots x_py_qy_q+1y_q+2x_p+1y_q+3y_q+4y_q+5x_p+2 \cdots \]
until we get 414 occurrences of $Q(3)$, etc.

This gives an infinite squarefree word
\[ v = x_0y_0x_1y_1x_2y_2 \cdots x_iy_i+1x_i+1y_i+2y_i+2x_i+3 \cdots \]
\[ \cdots x_jy_jy_j+1y_j+1y_j+2y_j+3x_j+2y_j+4y_j+5x_j+3 \cdots \]
\[ \cdots x_py_qy_q+1y_q+2x_p+1y_q+3y_q+4y_q+5x_p+2 \cdots \]
over the alphabet $C$. The word $v$ therefore has the form
\[ v = \cdots Q(1) \cdots Q(1) \cdots Q(2) \cdots Q(2) \cdots Q(n) \cdots Q(n) \cdots \]
where, for each $n$, there are 414 occurrences of $Q(n)$. Note that although we have shown the $Q(n)$ above as being non-overlapping, they may indeed overlap each other; this poses no problems for the argument below.

Since $y$ is squarefree, the word $v$ is as well. Furthermore, we claim that for any $n \geq 1$, the factor $Y_2$ in any given $Q(n) = aY_1bY_2cY_3d$ can be removed and the resulting word $\hat{v}$ is squarefree. To see this, suppose to the contrary that $\hat{v}$ contains a square $uu$. Since $bc$ does not occur anywhere else in $\hat{v}$, the first $u$ must end at the $b$ in $aY_1bcY_3d$ and the second $u$ must begin at the $c$. Since $b \neq c$, the word $u$ must contain at least two letters from $A$. If $u$ contains exactly two letters from $A$, then $abcd$ is a square in $x$. However, the only square of length 4 in $x$ is 0101, and this contradicts the hypothesis that $abcd \neq 0101$. If $u$ contains more than two letters from $A$, then $x$ contains a square of length $\geq 6$, which is a contradiction.

Let $h_5 : C^* \to \{0, 1, 2\}^*$ be the following 18-uniform morphism given by Brandenburg [5, Theorem 4].

\[
\begin{align*}
0 & \to 0102012012101212 \\
1 & \to 0102012012012012 \\
2 & \to 01020120120121012 \\
3 & \to 01020120120120121012 \\
4 & \to 01020120120120121012
\end{align*}
\]
Brandenburg proved that $h_5$ is a squarefree morphism; i.e., it maps squarefree words to squarefree words. Let $v' = h_5(v)$. The infinite word $v'$ is squarefree, and furthermore, since $h_5$ is a squarefree morphism, we see that any factor $h_5(Y_2)$ occurring in the context $h_5(Q(n)) = h_5(aYibY_2cY_3d)$ can be deleted from $v'$ and the resulting word is also squarefree. Note that $|h_5(Y_2)| = 18n$.

The next step in the construction is to apply the following squarefree multi-valued morphism $g$ from \[2, Theorem 20\]:

$$g(0) = \begin{cases} 01210212012021201210 \\ 012102120201201210 \\ 01210212012021201210 \\ 012102120201201210, \end{cases}$$

$$g(1) = \pi(g(0)),$$ and $g(2) = \pi(g(1))$, where $\pi$ is the permutation $(012)$. Note that the images of each letter have lengths 23, 24, 25, or 26. Consequently, for any $a \in C$, the words in the set $g(h_5(a))$ all have lengths between $18 \times 23 = 414$ and $18 \times 26 = 414 + 54$ and all such lengths are obtained.

Let $w$ be the infinite word in $g(v')$ obtained as follows. When applying $g$ to $v'$, in general, we choose to replace each letter of $v'$ with its image under $g$ of length 23, except in the following situation.

For each $n \geq 1$ and each $i \in \{0, \ldots, \min\{413, 54n\}\}$, when applying $g$ to the $i$-th occurrence of $h_5(Q(n)) = h_5(aYibY_2cY_3d)$, replace $h_5(Y_2)$ with any word $Z$ in $g(h_5(Y_2))$ of length $414n + i$. Since $g$ is a squarefree multi-valued morphism, the word $Z$ is a disposable factor of $w$ of length $414n + i$. The set of lengths of all such disposable factors is therefore

$$L = \bigcup_{n \geq 1} \{414n + i : i \leq \min\{413, 54n\}\} = \{k : k \geq 3312\} \cup \bigcup_{n=1}^{7} \{414n + i : i \leq 54n\}. $$

This completes the proof; we note in conclusion that there are 1792 lengths missing from the set $L$.

Some of the missing lengths from the proof of Theorem 5 can be obtained by the following observation due to Harju: if $w$ is an infinite squarefree word and there is some factor $p$ and some letter $a$ for which we can write $w = apaw'$, then $pa$ is disposable, since the resulting word $aw'$ is a suffix of $w$ and hence is squarefree. Thus, for every such $p$, we can add the length $|pa|$ to the list of lengths of disposable factors of $w$. Note that to explicitly calculate these additional lengths for a word $w$ constructed as described in the proof of Theorem 5, we would have to make some explicit choices for the word $x$ and the word $y$ used in the proof, as well as an explicit rule for choosing the word $Z$ in $g(h_5(Y_2))$.

## 5 Conclusion

An obvious open problem is to completely resolve Harju’s question by improving the construction of Theorem 5 so that there are interior disposable factors of every length. Harju
[11] also stated two other very interesting open problems in his paper, which we have not been able to solve.

Regarding the disposable positions in $vtm$, the main property of $vtm$ that allowed us to identify the disposable positions was that $vtm$ avoids 010 and 212. This places it in one of the three classes of squarefree words characterized by Thue [18]: 1) those avoiding 010 and 212; 2) those avoiding 010 and 020; and, 3) those avoiding 121 and 212. It might be interesting to study disposable positions in words from classes 2) and 3).

We also found that the set of disposable positions in $vtm$ is fairly dense: the density of disposable positions in $vtm$ is 1/12. Let $D_w(n)$ denote the set of disposable positions $\leq n + 1$ of an infinite squarefree word $w$. What is the greatest possible value of $\lim \inf_{n \to \infty} |D_w(n)|/n$ over all infinite ternary squarefree words $w$? Is it achieved by $vtm$?

References

[1] J.-P. Allouche, J. Shallit, *Automatic Sequences: Theory, Applications, Generalizations*, Cambridge, 2003.

[2] J. Bell, “The upper density of an automatic set is rational”. Preprint, [https://arxiv.org/abs/2002.07256](https://arxiv.org/abs/2002.07256).

[3] J. Berstel, “Sur la construction de mots sans carré”, Séminaire de Théorie des Nombres 1978â€“1979, Exp. No. 18, 15 pp., CNRS, Talence, 1979.

[4] F. Blanchet-Sadri, J. Currie, N. Fox, N. Rampersad, “Abelian complexity of fixed point of morphism $0 \mapsto 012$, $1 \mapsto 02$, $2 \mapsto 1$”, INTEGERS 14 (2014), A11.

[5] F.-J. Brandenburg, “Uniformly growing $k$-th power-free homomorphisms”, Theoret. Comput. Sci. 23 (1983), 69–82.

[6] C. Braunholtz, “An infinite sequence of three symbols with no adjacent repeats”, Amer. Math. Monthly 70 (1963), 675–676.

[7] J. Currie, T. Harju, P. Ochem, N. Rampersad, “Some further results on squarefree arithmetic progressions in infinite words”, Theoret. Comput. Sci. 799 (2019), 140–148.

[8] A. S. Fraenkel, J. Simpson, “How many squares must a binary sequence contain?”, Electron. J. Combinatorics 2 (1995) #R2.

[9] J. Grytczuk, H. Kordulewski, A. Niewiadomski, “Extremal squarefree words”. Preprint, [https://arxiv.org/abs/1910.06226](https://arxiv.org/abs/1910.06226).

[10] M. Hall, Jr., “Generators and relations in groups–The Burnside problem”. In *Lectures on Modern Mathematics, Vol. 2*, Wiley, New York, pp. 42–92.

[11] T. Harju, “Disposability in square-free words”. Preprint: [https://arxiv.org/abs/1911.09973](https://arxiv.org/abs/1911.09973) (2020).
[12] T. Harju, D. Nowotka, “Binary words with few squares”, TUCS Technical Report No. 373, January 2006.

[13] S. Istrail, “On irreducible languages and nonrational numbers”, Bull Math. Soc. Sci. Math. R. S. Roumanie (N.S.) 21(69)(3-4) (1977), 301–308.

[14] M. Lothaire, Combinatorics on Words, vol. 17 of Encyclopedia of Mathematics, Addison-Wesley, Reading, MA, 1983.

[15] M. Morse and G. Hedlund, “Unending chess, symbolic dynamics and a problem in semigroups”, Duke Math. J. 11 (1944), 1–7.

[16] H. Mousavi, “Automatic theorem proving in Walnut”. Preprint: https://arxiv.org/abs/1603.06017 (2016). Walnut Prover itself is available at https://github.com/hamousavi/Walnut.

[17] A. Thue, “Über unendliche Zeichenreihen”, Norske Vid. Skrifter I. Mat.-Nat. Kl., Christiana 7 (1906), 1–22.

[18] A. Thue, “Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen”, Norske Vid. Skrifter I. Mat.-Nat. Kl., Christiana 1 (1912), 1–67.