Extremal Square-free Words

Jarosław Grytczuk * Hubert Kordulewski
Artur Niewiadomski

Faculty of Mathematics and Information Science
Warsaw University of Technology
00-662 Warsaw, Poland

j.grytczuk@mini.pw.edu.pl, apostatajulian@gmail.com,
artur.niewiadomski.93@gmail.com

Submitted: Jan 4, 2020; Accepted: Jan 29, 2020; Published: Mar 6, 2020

Abstract

A word is *square-free* if it does not contain nonempty factors of the form XX. In 1906 Thue proved that there exist arbitrarily long square-free words over a 3-letter alphabet. We consider a new type of square-free words with an additional property. A square-free word is called *extremal* if it cannot be extended to a new square-free word by inserting a single letter at any position. We prove that there exist infinitely many square-free extremal words over a 3-letter alphabet. Some parts of our construction relies on computer verifications. It is not known if there exist any extremal square-free words over a 4-letter alphabet.

Mathematics Subject Classifications: 05A05, 05D10

1 Introduction

A *square* is a nonempty word of the form XX. For instance,

$$\mathtt{aa}, \mathtt{abab}, \mathtt{abcabc}, \mathtt{abacabac}$$

are examples of squares. A word is *square-free* if it does not contain a square as a *factor* (a subword consisting of consecutive letters). It is easy to check that there are no binary square-free words of length more than 4. However, there exist ternary square-free words of any length, as proved by Thue in [9] (see [3]). This result is the starting point of Combinatorics on Words, a wide discipline with lots of exciting problems, deep results, and important applications (see [1, 2, 4, 6, 7, 8]).

*Supported by the Polish National Science Center grant 2015/17/B/ST1/02660.

In this paper we propose a new problem of extremal nature in this area. Let $\mathbb{A}$ be a fixed alphabet and let W be a finite word over $\mathbb{A}$. An *extension* of W is any word of the form $W'xW''$, where $x \in \mathbb{A}$ and $W = W'W''$. A square-free word W is called *extremal* over $\mathbb{A}$ if there is no square-free extension of W. For instance, the word

$$H = \mathtt{abcabacbcabcbabcabacbcabc}$$

is the shortest extremal word over alphabet $\mathbb{A} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Our main result asserts that there exist infinitely many such words.

Theorem 1. *There exist infinitely many extremal square-free words over a* 3*-letter alphabet.*

The proof is by recursive construction whose validity is partially based on computer verifications. We will give it in section 2. In the final section we state some open problems.

2 Proof of the main result

We start with a general result on which our construction is based. Consider a finite directed graph D on the set of vertices $V = \{v_1, v_2, \ldots, v_n\}$. Suppose that each vertex v_i is labeled with some word $B_i = f(v_i)$ over a fixed alphabet $\mathbb{A}$. We will refer to these words B_i as *blocks*.

A *walk* in D is any sequence $W = w_1 w_2 \cdots w_t$, with $w_i \in V$, such that (w_i, w_{i+1}) is a directed edge of D for every $i = 1, 2, \ldots, t-1$. Every walk $W = w_1 w_2 \cdots w_t$ generates in a natural way the word $f(W) = f(w_1)f(w_2)\cdots f(w_t)$ over alphabet $\mathbb{A}$ by concatenating blocks corresponding to consecutive vertices w_i in W. More formally, one may consider f as a homomorphism from the monoid V^* to the monoid $\mathbb{A}^*$ defined by the substitution $f(v_i) = B_i$.

A walk is *square-free* if it is a square-free word over alphabet V. We say that a digraph D is a *Thue digraph* if for every square-free walk W, the word $f(W)$ is also square-free (as a word over $\mathbb{A}$). Let $S(D)$ denote the set of all words over $\mathbb{A}$ derived as images of any square-free walks in D. So, a digraph D is a Thue digraph if $S(D)$ contains only square-free words. The result below gives sufficient conditions for this property.

Theorem 2. *Let D be a digraph on the set of vertices $V = \{v_1, v_2, \ldots, v_n\}$ labeled with some blocks $B_i = f(v_i)$ over alphabet $\mathbb{A}$. Then D is a Thue digraph if the following conditions are satisfied:*

(1) For every square-free walk $W = w_1 w_2 w_3$, the word $f(W)$ is also square-free.

(2) No block B_i is a factor of another block B_j (unless $i = j$). In particular, blocks B_i are pairwise different.

(3) For every pair of distinct blocks B_i and B_j, $i \neq j$, and any factorizations $B_i = XX'$ and $B_j = YY'$, none of the words XY' nor $X'Y$ can be equal to any block B_k, unless $B_k = B_i = X$ or $B_k = B_j = Y$.

Proof. Suppose on the contrary that a square XX appears in some word $f(W)$, where $W = w_1w_2\cdots w_t$ is a square-free walk in D. Assume also that W is a shortest such walk. So, we may write (see Figure 1):

$$f(W) = PP'f(w_2)\cdots f(w_j)QQ'f(w_{j+2})\cdots f(w_{t-1})RR' = PXXR',$$

where $f(w_1) = PP'$,$f(w_{j+1}) = QQ'$, $f(w_t) = RR'$, and

$$X = P'f(w_2)\cdots f(w_j)Q = Q'f(w_{j+2})\cdots f(w_{t-1})R = X.$$

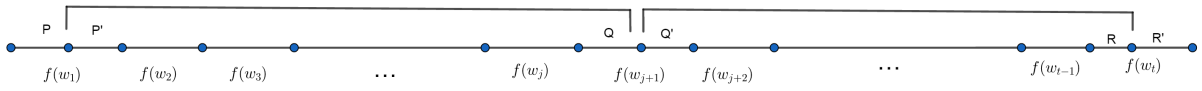

Figure 1: A square in $f(W)$.

By condition (1), the walk W has at least four vertices, hence, at least one part of the square must contain a full occurrence of some block. With no loss of generality we may assume that this happens in the left part. Also we may assume that this part contains as many blocks as the other part.

Let $q \in \{1, 2, \ldots, j\}$ be the smallest index such that $w_q \neq w_{j+q}$. There must be at least one such index since otherwise the walk W would contain the square

$$w_1w_2\cdots w_jw_1w_2\cdots w_j,$$

contradicting our assumption. We distinguish two cases.

If $q > 1$, then either $f(w_q)$ is a prefix of $f(w_{j+q})$ or the other way around, which contradicts condition (2).

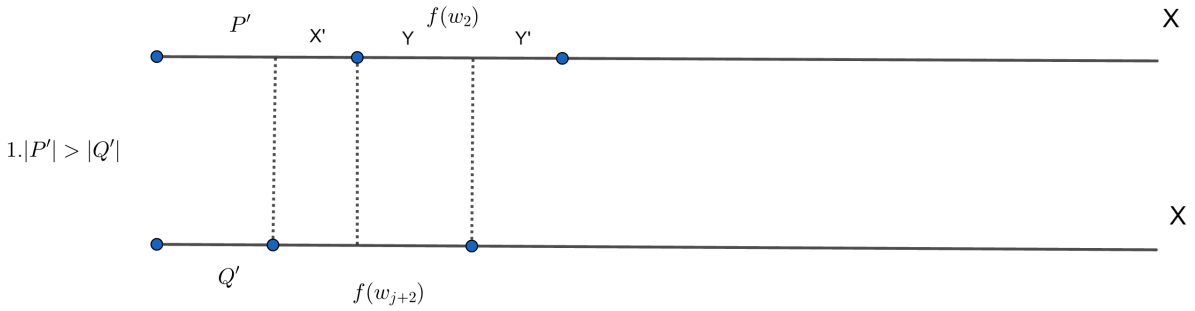

Figure 2: Comparing blocks in the square XX (case 1).

If $q = 1$, then $f(w_1) \neq f(w_{j+1})$ and we consider two cases. First suppose that the words P' and Q' have different lengths, and assume that P' is longer than Q'. Then we may write $P' = Q'X'$, where X' is a nonempty suffix of the block $f(w_1)$ (see Figure 2). Now, the

block $f(w_{j+2})$ must end before $f(w_2)$ since otherwise $f(w_2)$ would be contained in $f(w_{j+2})$, contradicting condition (2). So, we may write $f(w_{j+2}) = X'Y$, where $f(w_2) = YY'$. This contradicts condition (3). If Q' is longer than P', then the reasoning is similar.

Suppose now that the words P' and Q' have equal length, which means that $P' = Q'$. This implies that all pairs of corresponding inner blocks in the left and the right part of the square XX must be equal (otherwise one of them would be included in the other, contradicting condition (2) (see Figure 3)). This implies that $t = 2j + 1$ and $w_i = w_{j+i}$ for all $i = 2, 3, \ldots, j$, and the walk W can be written as

$$W = w_1 w_2 \cdots w_j w_{j+1} w_2 \cdots w_j w_t = w_1 Z w_{j+1} Z w_t.$$

In consequence, we get that also $Q = R$ (see Figure 3), which means that $f(w_{j+1}) = QQ' = RP'$. If $f(w_1) = f(w_t)$, then we have $P = R = Q$ and $P' = R' = Q'$, which implies that $f(w_1) = f(w_{j+1}) = f(w_t)$. If $f(w_1) \neq f(w_t)$, then by condition (3) it follows that either $f(w_{j+1}) = f(w_1)$ or $f(w_{j+1}) = f(w_t)$. In both cases we get that $w_{j+1} = w_1$ or $w_{j+1} = w_t$ which gives a square in the walk W. This completes the proof. □

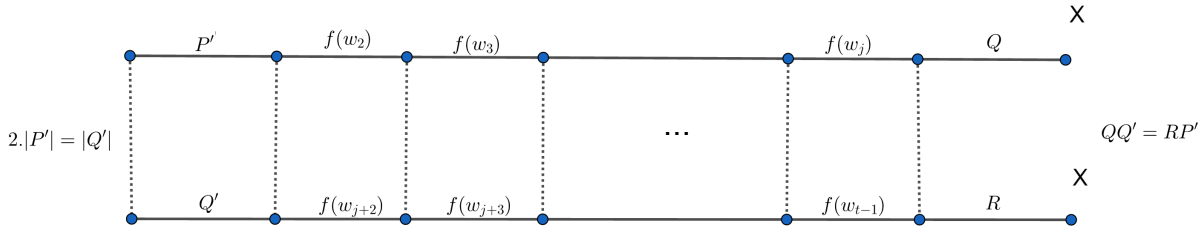

Figure 3: Comparing blocks in the square XX (case 2).

Using the above result we may now prove Theorem 1. First we will construct a Thue digraph on the set of 12 vertices together with the set of 12 blocks defined as follows. Consider the following square-free word

$$N = \texttt{abacbabcabacbcacbabcabacabcbabcabacbcabcb}.$$

This word is nearly extremal in the following sense. A square-free word W is called *nearly extremal* if it has at most two square-free extensions, and these extensions may have only the form xW or Wy, where $x, y \in \mathbb{A}$. The following lemma can be verified by a computer.

Lemma 3. *The word N is nearly extremal and the only square-free extensions of N are $\texttt{c}N$ and $N\texttt{a}$.*

It is clear that each word obtained from N by a permutation of the alphabet and by reversal is also nearly extremal. Let us denote the six words corresponding to the six permutations of the alphabet as:

$$N, N_{\texttt{ab}}, N_{\texttt{ac}}, N_{\texttt{bc}}, N_{\texttt{abc}}, N_{\texttt{acb}},$$

where indices denote nontrivial cycles of these permutations. Let us also denote reversals of the above six words by:

$$\tilde{N}, \tilde{N}_{\mathtt{ab}}, \tilde{N}_{\mathtt{ac}}, \tilde{N}_{\mathtt{bc}}, \tilde{N}_{\mathtt{abc}}, \tilde{N}_{\mathtt{acb}}.$$

Now we may define a digraph D_N as depicted in Figure 4. Its vertices are labeled by the above 12 nearly extremal words. It can be checked that for each directed edge of D_N the corresponding concatenation of blocks gives a square-free word. Moreover, the following lemma was verified by a computer.

Lemma 4. *The digraph D_N from Figure 4 is a Thue digraph.*

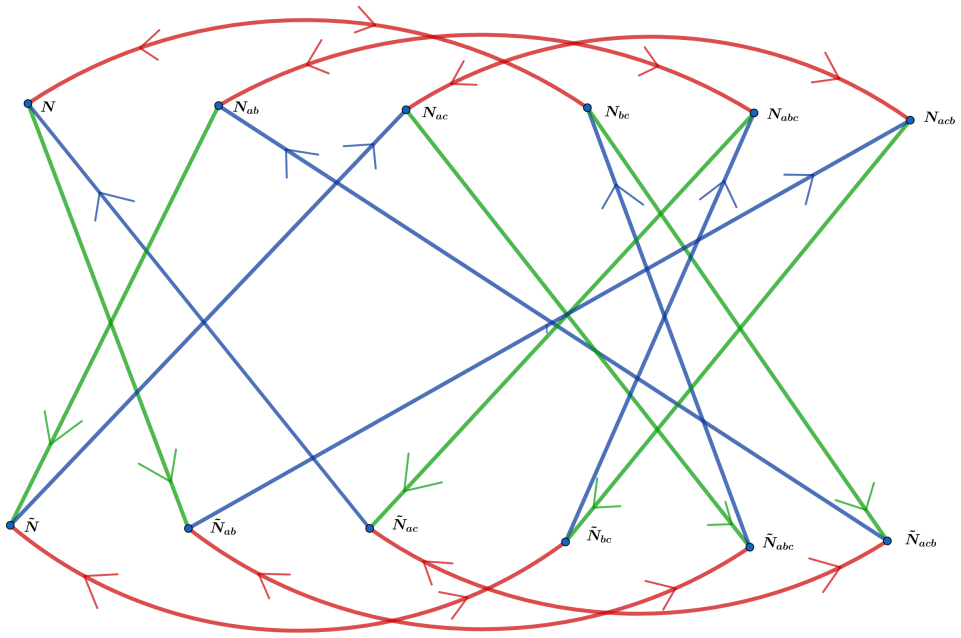

Figure 4: The digraph D_N.

Recall that the set $S(D_N)$ consists of all words derived as homomorphic images of all square-free walks in D_N.

Corollary 5. *All words in $S(D_N)$ are square-free and nearly extremal.*

Proof. By Lemma 4, the set $S(D_N)$ consists only of square-free words. Moreover, every word in $S(D_N)$ is a concatenation of blocks that are nearly extremal words. Thus it cannot be extended at any inner position of a block. By the same reason it cannot be extended by inserting a letter between any two blocks. Indeed, for any two consecutive blocks B_1B_2 occurring in some word from $S(D_N)$ there is only one letter x such that B_1x is square-free, and this letter must be the first letter of the next block B_2. Hence, for any letter y, the word B_1yx must contain a square. □

We are going to prove that the set $S(D_N)$ is infinite. We will need the following general lemmas.

Lemma 6. *Let $T = a_1a_2\cdots a_s$ be a square free word over alphabet $\mathbb{A} = \{1,2,3\}$, and let V_1, V_2, V_3 be three pairwise disjoint alphabets. Then any word of the form $W = W_1W_2\cdots W_s$ is square-free, where W_i is any word over alphabet V_{a_i} consisting of pairwise distinct letters.*

Proof. Indeed, the word W can be seen as an image of T in a multi-substitution, where for each letter $i \in \mathbb{A}$ we may put any word over V_i with pairwise distinct letters. Such substitutions obviously preserve square-freenes since alphabets V_i are pairwise disjoint. □

Lemma 7. *Let D be a digraph whose vertices can be partitioned into three sets V_1, V_2, and V_3 so that the following property holds:*

(∗) For every pair $i, j \in \{1,2,3\}$ and any vertex $v \in V_i$, there is a directed path $P = u_1u_2\cdots u_t$ such that $u_1 = v$, $\{u_2, \ldots, u_{t-1}\} \subseteq V_i$, and $u_t \in V_j$.

Then there exists arbitrarily long square-free walks in D.

Proof. Let us take any square-free word $T = a_1a_2\cdots a_s$ over the alphabet $\mathbb{A} = \{1,2,3\}$. Let $v = v_1$ be any vertex in V_{a_1}. Let P_1 be a directed path satisfying condition $(*)$, starting at v_1 and ending at some vertex $v_2 \in V_{a_2}$. Now take a similar path P_2 starting from v_2 and ending at some vertex $v_3 \in V_{a_3}$. And so on, until we arrive to some vertex $v_s \in V_{a_s}$. In this way, we obtain a walk

$$W = P_1'P_2'\cdots P_s',$$

where $P_i' = P_i - \{v_{i+1}\}$ and $P_s' = v_s$. By Lemma 6, the walk W is square-free. □

Lemma 8. *There exist arbitrarily long square-free walks in the digraph D_N starting and ending at the vertex labeled with the word N.*

Proof. It is not hard to check that the following partition of $V(D_N)$ satisfies condition $(*)$ of Lemma 7 (see Figure 5):

$$V_1 = \{N, N_{\mathtt{bc}}, \tilde{N}, \tilde{N}_{\mathtt{bc}}\}, V_2 = \{N_{\mathtt{ab}}, N_{\mathtt{abc}}, \tilde{N}_{\mathtt{ab}}, \tilde{N}_{\mathtt{abc}}\}, V_3 = \{N_{\mathtt{ac}}, N_{\mathtt{acb}}, \tilde{N}_{\mathtt{ac}}, \tilde{N}_{\mathtt{acb}}\}.$$

So, by Lemma 7, there exist arbitrarily long square-free walks in D_N. We need to show that they may start and end at the vertex N. To see this take a sufficiently long square-free word T over the alphabet $\{1,2,3\}$ of the form $T = 1U1$ such that the word $T' = T231$ is also square-free. Now, we may construct a square free walk W along T like in the proof of Lemma 7, starting from the vertex N and ending at some vertex v in V_1. If $v = N$ we are done. If not, then we need to extend the walk W slightly. If $v = N_{\mathtt{bc}}$, then we make just one step to reach N directly. If $v = \tilde{N}_{\mathtt{bc}}$, then we have to go first to $N_{\mathtt{abc}}$ in V_2, next to $\tilde{N}_{\mathtt{ac}}$ in V_3, and then jump to N from there (see Figure 5). This gives a square-free walk, since T' is square-free. Finally, if $v = \tilde{N}$, then we move first to $\tilde{N}_{\mathtt{bc}}$ and then repeat the previous three steps from there. □

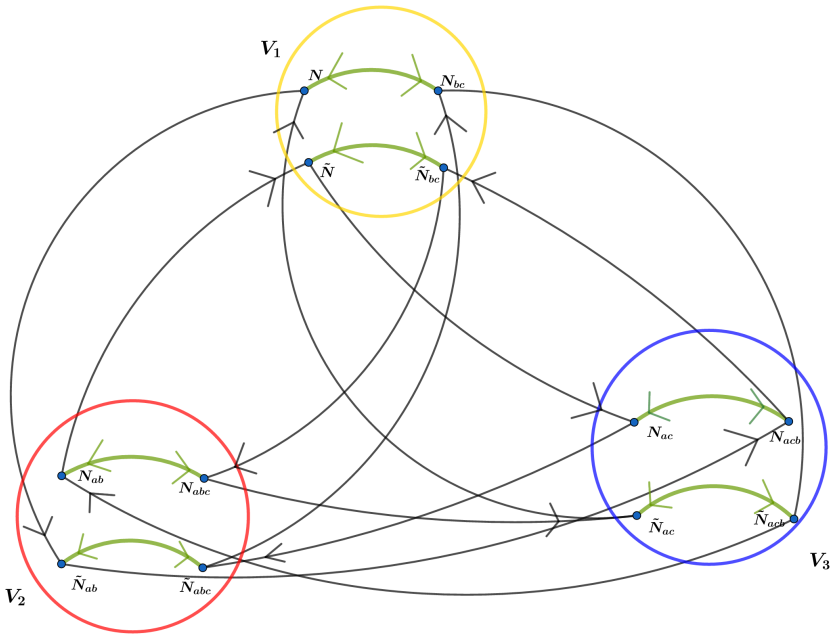

Figure 5: The digraph D_N with vertex partition into three sets.

Corollary 5 and Lemma 8 give immediately the following conclusion.

Corollary 9. *There exist infinitely many nearly extremal square-free words over a* 3*-letter alphabet.*

To prove the assertion of Theorem 1 we need to modify slightly the digraph D_N. The idea is to use two special words:

$$Q = \texttt{cbacbcabacbabcabacbcabcbacbc}$$

and

$$R = \texttt{acabcacbabcabacbcabcbacabacbcabcb}.$$

The following lemma can be checked by a computer.

Lemma 10. *The words QN and NR are square-free and each have only one square-free extension, namely $QN\texttt{a}$ and $\texttt{c}NR$.*

Now we may add two new vertices to our digraph D_N with labels Q and R, and join Q to N and N to R by directed edges. Denote this modified digraph as D_N^*. The following lemma can also be verified by a computer.

Lemma 11. *The digraph D_N^* is a Thue digraph.*

To construct extremal words of length exceeding any given constant it is enough to take a sufficiently long square-free walk W in D_N^* starting at Q and ending in R. Such a walk exists by Lemma 8 and Lemma 11. The word $E = f(W)$ corresponding to this walk will have the form $E = QNYNR$, where NYN is a nearly extremal word by Corollary 5. Hence, by Lemma 10, the word E is extremal. This completes the proof of Theorem 1.

3 Discussion

A natural question is whether an analogue of Theorem 1 holds for larger alphabets. Notice that if we have four letters at a disposal, then there are two potential possibilities for extension of a square-free word at every inner position. Actually, our computer experiment failed in finding extremal words over four letters of length up to 100. Perhaps there are no such words at all. On the other hand, it is known [2] that every square-free word (over any alphabet) is a prefix of a *maximal* square-free word, that is, a word non-extendable by attaching a single letter at the beginning or at the end.

Conjecture 12. There are no extremal square-free words over 4-letter alphabet.

The conjecture states, in other words, that every square-free word over four letters can be extended to a new square-free word by inserting a single letter at some position.

The concept of extremal words can be considered for any fixed *pattern* (or even for any property of words which is monotone on factors). To state a general conjecture on extremal words we recall briefly basic notions of pattern avoidance (see [2, 5, 8]).

Let $\mathbb{V}$ be an alphabet of variables. A *pattern* $P = p_1 p_2 \dots p_r$, with $p_i \in \mathbb{V}$, is any nonempty word over $\mathbb{V}$. A word W *realizes* a pattern P if it can be split into nonempty factors $W = W_1 W_2 \dots W_r$ so that $W_i = W_j$ if and only if $p_i = p_j$, for all $i, j = 1, 2, \dots, r$. A word W *avoids* a pattern P if no factor of W realizes P. For instance, a square-free word avoids a pattern $P = xx$. A pattern P is *avoidable* if there exist arbitrarily long words over some finite alphabet avoiding P. A complete characterizations of avoidable patterns was provided independently by Zimin [10] and Bean, Ehrenfeucht and McNulty [2].

Now, given a fixed pattern P, we may define *extremal P-avoiding* words analogously as in the case of squares. The following conjecture seems plausible.

Conjecture 13. For every avoidable pattern P there exists a constant $k = k(P)$ such that the set of extremal P-avoiding words over k-letter alphabet is finite.

We conclude the paper with another related question. Consider the following *greedy* way of generating square-free words. Given a fixed ordered alphabet $\mathbb{A}$, we start with the first letter from $\mathbb{A}$ and continue by inserting at the rightmost position of the actual word the earliest possible letter so that the new word is square-free. For instance, for the alphabet $\mathbb{A} = \{\mathtt{1}, \mathtt{2}, \mathtt{3}\}$ this greedy procedure starts with the following sequence of square-free words:

$$\mathtt{1}, \mathtt{12}, \mathtt{121}, \mathtt{1213}, \mathtt{12131}, \mathtt{121312}, \mathtt{1213121}, \mathtt{12131231}.$$

The last word was obtained by inserting $\mathtt{2}$ at the penultimate position of the previous word.

We conjecture that the above procedure never stops. To state it formally, let us define recursively a sequence of *nonchalant words* G_i over the alphabet $\mathbb{A}_k = \{\mathtt{1}, \mathtt{2}, \dots, \mathtt{k}\}$ by putting $G_1 = \mathtt{1}$, and letting $G_{i+1} = G_i' x G_i''$ to be a square-free extension of G_i such that G_i'' is the shortest possible suffix of G_i and $x \in \mathbb{A}_k$ is the earliest possible letter.

Conjecture 14. The sequence of nonchalant words over $\mathbb{A}_k$ is infinite for every $k \geqslant 3$.

The results of a computer experiment for $k = 3$ supports this conjecture; a nonchalant word of length 5000 was obtained by the above greedy procedure. Moreover, the algorithm never moved back by more that 15 positions. Therefore the following conjecture seems plausible.

Conjecture 15. The sequence of nonchalant words over $\mathbb{A}_k$ converges to an infinite word for every $k \geqslant 3$.

Here are the first 70 terms of the presumably infinite limit word for $k = 3$:

1213123132123121312313231213123212312131231321231213123212312132123132 ...

Acknowledgements

We would like to thank Lucas Mol and Jeffrey Shallit for useful remarks and suggestions.

References

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

[2] D. R. Bean, A. Ehrenfeucht, G. F. McNulty, Avoidable patterns in strings of symbols, Pacific J. Math. 85 (1979) 261–294.

[3] J. Berstel, Axel Thue's papers on repetitions in words: a translation, Publications du LaCIM, vol. 20, Université du Québec a Montréal, 1995.

[4] J. Berstel, D. Perrin, The origins of combinatorics on words, Europ. J. Combin. 28 (2007) 996–1022.

[5] J. Currie, Pattern avoidance: themes and variations, Theoret. Comput. Sci. 339 (2005) 7–18.

[6] J. Grytczuk, Thue type problems for graphs, points, and numbers, Discrete Math. 308 (2008) 4419–4429.

[7] M. Lothaire, Combinatorics on Words, Addison-Wesley, Reading, MA, 1983.

[8] M. Lothaire, Algebraic Combinatorics on Words, Cambridge University Press, Cambridge, UK, 2002.

[9] A. Thue, Über unendliche Zeichenreichen, Norske Vid. Selsk. Skr., I Mat. Nat. Kl., Christiania 7 (1906) 1–22.

[10] A. I. Zimin, Blocking sets of terms, Mat. Sb. 119 (1982) 363–375. Translated in Sb. Math. 47 (1984) 353–364.