

# NO EXTREMAL SQUARE-FREE WORDS OVER LARGE ALPHABETS

Letong Hong<sup>1</sup> and Shengtong Zhang<sup>2</sup>

<sup>1,2</sup>*Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02142, U.S.A.*  
*clhong@mit.edu, stzh1555@mit.edu*

Submitted: Sep 12, 2021; Accepted: Feb 23, 2022; Published: Mar 31, 2022

© The authors. Released under the CC BY license (International 4.0).

**Abstract.** A word is *square-free* if it does not contain any *square* (a word of the form  $XX$ ), and is *extremal square-free* if it cannot be extended to a new square-free word by inserting a single letter at any position. Grytczuk, Kordulewski, and Niewiadomski proved that there exist infinitely many ternary extremal square-free words. We establish that there are no extremal square-free words over any alphabet of size at least 17.

**Keywords.** Combinatorics on words, square-free words, extremal words

**Mathematics Subject Classifications.** 05A05, 05D10, 68R15

## 1. Introduction

A *word* is a finite sequence of letters over a finite alphabet. A *factor* of a word is a subword of it consisting of consecutive letters. A *square* is a nonempty word of the form  $XX$  (examples: “couscous”, “hotshots”, “murmur”). A word is *square-free* if it does not contain a square as a factor (examples: “abracadabra”, “bonobo”, “squares”; non-examples: “entente”, “referee”, “tartar”). It is easy to check that there are no binary square-free words of length more than 3. Thue showed in 1906 [8] that there are arbitrarily long ternary square-free words (see [1]). His work is considered to be the beginning of research in combinatorics on words [2].

Recently, Grytczuk, Kordulewski, and Niewiadomski [4] introduced the study of *extremal square-free words*.

**Definition 1.1.** An *extension* of a finite word  $W$  is a word  $W' = W_1xW_2$ , where  $x$  is a single letter and  $W_1, W_2$  are (possibly empty) words such that  $W = W_1W_2$ . An *extremal square-free word*  $W$  is a square-free word such that none of its extensions is square-free.

The only binary extremal square-free words are 010 and 101. Via a delicate construction, Grytczuk et al. showed in [4] that there exist infinitely many ternary extremal square-free words. Grytczuk, Kordulewski, and Pawlik also raised several open problems concerning larger alphabet

sizes ([4], [5]), including nonexistence of extremal square-free words over an alphabet of size 4. Mol and Rampersad [6] then classified all possible lengths of extremal ternary square-free words.

**Conjecture 1.2** ([4], [6]). There exists no extremal square-free word over a finite alphabet of size at least 4.

To the authors' knowledge, Conjecture 1.2 is open for any finite alphabet. Using ideas of Ter-Saakov and Zhang in [7] and some new observations, our main result confirms their conjecture for alphabets of size at least 17.

**Theorem 1.3.** *For any integer  $k \geq 17$ , there exists no extremal square-free word over an alphabet of size  $k$ .*

In [4] and [5], Grytczuk, Kordulewski, Niewiadomskim and Pawlik also introduced and discussed the notion of *nonchalant words*. The sequence of nonchalant words  $G_i$  is generated recursively by the following greedy procedure. Fix a total ordering on the alphabet.  $G_0$  is the empty word, and  $G_{i+1} = G'_i x G''_i$  is a square-free extension of  $G_i$ , where  $G_i = G'_i G''_i$  with  $G''_i$  being the shortest possible suffix of  $G_i$  and  $x$  being the smallest possible letter such that  $G_{i+1}$  is square-free. Theorem 1.3 partially affirmatively answers Conjecture 14 and 15 in [4] for nonchalant words.

**Corollary 1.4.** *For any integer  $k \geq 17$ , the sequence of nonchalant words over a fixed alphabet of size  $k$  converges to an infinite word.*

## 2. Proof of Theorem 1.3

For a word  $W$  of length  $n$ , we number the letters in  $W$  from left to right as letter 1, 2,  $\dots$ ,  $n$ , and let  $W[i]$  be the letter  $i$  in  $W$ . We refer to the space between the letter  $i$  and the letter  $i + 1$  as gap  $i$ , and call the first and last gap 0 and  $n$ . For  $0 < a < b \leq n + 1$ , we define the factor  $W[a, b)$  as the subword of  $W$  consisting of letters  $a, a + 1, \dots, b - 1$ .

**Definition 2.1.** Let  $W$  be any word. Let  $W +_b c$  denote the word formed by inserting the letter  $c$  at gap  $b$ . For a positive integer  $a$  and a non-negative integer  $b$  with  $a \leq b + 1$ , a positive integer  $\ell$  and a letter  $c$ , we say the quadruple  $(a, \ell, b, c)$  is *square-completing in  $W$*  if the factor  $(W +_b c)[a, a + \ell)$  and the factor  $(W +_b c)[a + \ell, a + 2\ell)$  of  $W +_b c$  are the same word.

Define the *sign* of the quadruple to be 1 if  $b \leq a + \ell - 2$ , and  $-1$  if  $b \geq a + \ell - 1$ . The sign indicates whether the new letter we inserted at gap  $b$  lies in the factor  $(W +_b c)[a, a + \ell)$  or  $(W +_b c)[a + \ell, a + 2\ell)$ .

We now demonstrate two key propositions, then use them to prove Theorem 1.3.

**Proposition 2.2.** *Let  $W$  be a square-free word, and suppose  $(a, L, b, c)$  and  $(a', L, b', c')$  are square-completing quadruples in  $W$  with the same sign. Then one of the following holds:*

1.  $|a - a'| \geq L - 1$ ;
2.  $b = b'$  and  $c = c'$ .

*Proof.* Suppose to the contrary that neither (1) nor (2) is satisfied. Then  $|a - a'| < L - 1$ . By symmetry, we can assume the sign of both quadruples is 1, that is  $b \leq a + L - 2$  and  $b' \leq a' + L - 2$ . We argue by two cases on whether  $b = b'$  or not.

**Case 1.**  $b \neq b'$ . Without loss of generality, assume  $b < b'$ . We do additional case work on whether  $b' = a' - 1$  or  $b' \geq a'$ , i.e. whether the inserted letter at gap  $b'$  is at the start of the square in  $W +_{b'} c'$  or not.

First we handle the case  $b' \geq a'$ . We first show that it is impossible for  $b = a + L - 2$ . If  $b = a + L - 2$ , then we have

$$W[a, a+L-1] = (W +_b c)[a, a+L-1] = (W +_b c)[a+L, a+2L-1] = W[a+L-1, a+2L-2],$$

and we have found a square in  $W$ , which is a contradiction. Hence, we have

$$b \leq a + L - 3.$$

Furthermore, by the assumption that  $|a - a'| < L - 1$  we have

$$a' \leq a + L - 2.$$

Therefore, if we let  $i = \max(a', b + 1)$ , then we have  $i + 1 \leq a + L - 1$ , so

$$W[i] = (W +_b c)[i + 1] = (W +_b c)[i + 1 + L] = W[i + L].$$

On the other hand, as we assumed  $b < b'$  and  $a' \leq b'$ , we have  $i \leq b'$ . Thus we have

$$W[i] = (W +_{b'} c')[i] = (W +_{b'} c')[i + L] = W[i - 1 + L].$$

Thus we conclude that

$$W[i + L] = W[i - 1 + L].$$

So we have found a square in  $W$ , which is a contradiction.

Then we handle the case  $b' = a' - 1$ . In this case, we have

$$c' = (W +_{b'} c')[a'] = (W +_{b'} c')[a' + L] = W[a' + L - 1].$$

Note that as  $b' > b$ , we have  $a' + L - 1 > b$ , so

$$c' = W[a' + L - 1] = (W +_b c)[a' + L].$$

As  $a' + L > b + L + 1 \geq a + L$ , and  $a' + L \leq a + 2L - 1$ , we find that  $(W +_b c)[a' + L]$  is a letter in  $(W +_b c)[a + L, a + 2L - 1]$ . Therefore,

$$c' = (W +_b c)[a' + L] = (W +_b c)[a'].$$

Since  $a' = b' + 1 \geq b + 2$ , we get

$$c' = (W +_b c)[a'] = W[a' - 1].$$

However, this implies that

$$\begin{aligned} W[a' - 1, a' + L - 1] &= (W +_{b'} c')[a', a' + L] \\ &= (W +_{b'} c')[a' + L, a' + 2L] \\ &= W[a' + L - 1, a' + 2L - 1], \end{aligned}$$

so we have found a square in  $W$ , which is a contradiction.

**Case 2.**  $b = b'$ . We know  $(W +_b c)[b+1] = c$  and  $(W +_b c)[a, a+L] = (W +_b c)[a+L, a+2L]$ , so  $(W +_b c)[b+1+L] = c$ . This implies

$$W[b+L] = c.$$

The exact same logic also gives  $W[b'+L] = c'$ . As  $b = b'$ , we conclude that  $c = c'$ , which contradicts our assumption that (2) is not satisfied.  $\square$

**Proposition 2.3.** *Let  $W$  be a square-free word, and suppose  $(a, \ell, b, c)$  and  $(a', \ell', b', c')$  are square-completing quadruples in  $W$  with the same sign. Then one of the following holds:*

1. one of  $a, b, \ell$  differs by at least  $\frac{1}{5}L - 2$  from the corresponding  $a', b', \ell'$ , where  $L = \max(\ell, \ell')$ ;
2.  $b = b'$  and  $c = c'$ .

*Proof.* Suppose to the contrary that neither (1) nor (2) is satisfied. Then we have

$$\ell, \ell' \in [4L/5 + 2, L], \quad |b - b'| < \frac{1}{5}L - 2 \quad \text{and} \quad |a - a'| < \frac{1}{5}L - 2.$$

The case when  $\ell' = \ell = L$  follows from Proposition 2.2, so we only need to prove the proposition when  $\ell' \neq \ell$ . By symmetry, we can assume that  $L = \ell' > \ell$ , and that the sign of both quadruples is 1, that is,  $b \in [a - 1, a + \ell - 2]$  and  $b' \in [a' - 1, a' + \ell' - 2]$ . We argue by two cases on the quantity  $M = \max(b, b')$ .

**Case 1.**  $M \leq a + \frac{3L}{5}$ . Then, consider the word  $W[M+1, M+1+\ell'-\ell]$ . We know that the factor  $(W +_b c)[a, a+\ell]$  and the factor  $(W +_b c)[a+\ell, a+2\ell]$  of  $W +_b c$  are the same word. As  $M+1 > b$ , we have

$$W[M+1, M+1+\ell'-\ell] = (W +_b c)[M+2, M+2+\ell'-\ell].$$

On the other hand, we have

$$(M+2) + (\ell' - \ell) \leq \left(a + \frac{3L}{5} + 2\right) + \frac{L}{5} \leq a + \ell.$$

Therefore,  $(W +_b c)[M+2, M+2+\ell'-\ell]$  is a factor of  $(W +_b c)[a, a+\ell]$ , so it is equal to the corresponding factor of  $(W +_b c)[a+\ell, a+2\ell]$ . More precisely,

$$(W +_b c)[M+2, M+2+\ell'-\ell] = (W +_b c)[M+2+\ell, M+2+\ell'].$$

Thus we have

$$W[M+1, M+1+\ell'-\ell] = W[M+1+\ell, M+1+\ell'].$$

Similarly, since

$$a' \leq b' + 1 < M + 2$$

and

$$M + 2 + \ell' - \ell \leq a + \frac{4L}{5} + 2 \leq a' + L = a' + \ell',$$

we have  $(W +_b c)[M + 2, M + 2 + \ell' - \ell]$  is a factor of  $(W +_b c)[a', a' + \ell']$ , so we conclude that

$$(W +_b c)[M + 2, M + 2 + \ell' - \ell] = (W +_b c)[M + 2 + \ell', M + 2 + 2\ell' - \ell].$$

Thus we have

$$W[M + 1, M + 1 + \ell' - \ell] = W[M + 1 + \ell', M + 1 + 2\ell' - \ell].$$

But then we have

$$W[M + 1 + \ell, M + 1 + \ell'] = W[M + 1 + \ell', M + 1 + 2\ell' - \ell],$$

and we have found a square in  $W$ , which is a contradiction.

**Case 2.**  $M = \max(b, b') > a + \frac{3L}{5}$ . In this case, as  $|b - b'| \leq \frac{L}{5} - 2$ , we have  $\min(b, b') > a + \frac{2L}{5} + 2$ , and therefore  $\min(b, b') > \max(a, a') + \frac{L}{5} + 4$ . Let  $A = \max(a, a')$ . Then we note that

$$A + \ell' - \ell \leq A + \frac{L}{5} - 2 < \min(b, b').$$

So we conclude that

$$W[A, A + \ell' - \ell] = (W +_b c)[A, A + \ell' - \ell]$$

and

$$W[A, A + \ell' - \ell] = (W +_{b'} c')[A, A + \ell' - \ell].$$

As  $\min(b, b') \leq b < a + \ell$ , we have  $(W +_b c)[A, A + \ell' - \ell]$  is a factor of  $(W +_b c)[a, a + \ell]$ . So we conclude that

$$(W +_b c)[A, A + \ell' - \ell] = (W +_b c)[A + \ell, A + \ell'] = W[A + \ell - 1, A + \ell' - 1].$$

Similarly, because  $(W +_{b'} c')[A, A + \ell' - \ell]$  is a factor of  $(W +_{b'} c')[a, a' + \ell']$ , and

$$A + \ell' - 1 \geq a' + \ell' - 1 \geq b' + 1,$$

we conclude that

$$(W +_{b'} c')[A, A + \ell' - \ell] = (W +_{b'} c')[A + \ell', A + 2\ell' - \ell] = W[A + \ell' - 1, A + 2\ell' - \ell - 1].$$

Then we have

$$W[A + \ell - 1, A + \ell' - 1] = W[A + \ell' - 1, A + 2\ell' - \ell - 1],$$

and we have found a square in  $W$ , which is a contradiction. □

**Corollary 2.4.** *Let  $W$  be a square-free word of length  $n$ . Let  $\mathcal{A}$  be a set of square-completing quadruples  $(a, \ell, b, c)$  such that no two elements of  $\mathcal{A}$  share the same  $(b, c)$ . For each  $L \geq 2$ , define*

$$\mathcal{A}_L = \mathcal{A} \cap \{(a, L, b, c) : a, b \in \mathbb{Z}, c \text{ any letter}\}.$$

Then for any  $L \geq 2$ , we have

$$|\mathcal{A}_L| \leq \frac{2n}{L-1}.$$

Furthermore, for any  $L \geq 300$ , we have

$$\sum_{\ell=L}^{2L-1} |\mathcal{A}_\ell| \leq \frac{320n}{L}.$$

*Proof.* To prove the first proposition for  $L \geq 2$ , note that for a given sign  $\epsilon \in \{-1, 1\}$ , and any two quadruples  $(a, L, b, c)$  and  $(a', L, b', c')$  in  $\mathcal{A}_L$  with sign  $\epsilon$ , by Proposition 2.2 we must have  $|a - a'| \geq L - 1$ . Thus over all quadruples in  $\mathcal{A}_L$  with sign  $\epsilon$ , the  $a$ 's must be spaced at least  $L - 1$  apart, and must be in the range  $[1, n - 2L + 2]$ . Therefore, there are at most

$$\frac{n - 2L + 2}{L - 1} + 1 \leq \frac{n}{L - 1}$$

such quadruples. Given there are two possible signs, the total number of quadruples in  $\mathcal{A}_L$  is at most  $\frac{2n}{L-1}$ .

To prove the second statement, we can divide the range  $[1, n - 2L + 2]$  into at most

$$\frac{n - 2L + 2}{L/6} + 1 \leq \frac{6n}{L}$$

intervals of length  $\frac{L}{6}$ . For each such interval  $I = [x, x + \frac{L}{6})$ , define

$$\mathcal{B}_I = \{(a, \ell, b, c) \in \mathcal{A} : \ell \in [L, 7L/6), a \in I, (a, \ell, b, c) \text{ has sign } 1\}.$$

Assume  $(a, \ell, b, c)$  and  $(a', \ell', b', c')$  are two distinct quadruples in  $\mathcal{B}_I$ . Note that

$$|\ell - \ell'| \leq \frac{L}{6} < \frac{L}{5} - 2,$$

and

$$|a - a'| \leq \frac{L}{6} < \frac{L}{5} - 2.$$

Thus by Proposition 2.3, we must have  $b, b'$  spaced at least  $L/5 - 2$  apart. Furthermore, for each quadruple  $(a, \ell, b, c)$  in  $\mathcal{B}_I$ , the  $b$ 's are restricted to the interval  $[x - 1, x + \frac{4L}{3})$  due to the quadruples having sign 1. Thus the size of  $\mathcal{B}_I$  is upper bounded by

$$\frac{\frac{4L}{3} + 1}{\frac{L}{5} - 2} + 1.$$

For  $L \geq 300$ , we can verify that

$$\frac{\frac{4L}{3} + 1}{\frac{L}{5} - 2} + 1 < 8,$$

which implies

$$|\mathcal{B}_I| \leq 7.$$

Symmetrically, if we let

$$\mathcal{C}_I = \{(a, \ell, b, c) \in \mathcal{A} : \ell \in [L, 7L/6), a \in I, (a, \ell, b, c) \text{ has sign } -1\},$$

then

$$|\mathcal{C}_I| \leq 7.$$

Summing over all the intervals, we conclude that

$$\sum_{L \leq \ell < 7L/6} |\mathcal{A}_\ell| = \sum_I (|\mathcal{B}_I| + |\mathcal{C}_I|) \leq 14 \cdot \frac{6n}{L}.$$

Analogously, for any non-negative integer  $i$ , we have

$$\sum_{(7/6)^i L \leq \ell < (7/6)^{i+1} L} |\mathcal{A}_\ell| \leq \frac{14 \cdot 6n}{(7/6)^i L}.$$

Summing over  $i \in \{0, 1, 2, 3, 4\}$ , we obtain

$$\sum_{L \leq \ell < 2L} |\mathcal{A}_\ell| \leq \frac{320n}{L},$$

as desired. □

*Proof of Theorem 1.3.* Let  $W$  be any extremal square-free word of length  $n$  on an alphabet of size  $k$ . Then for any gap  $0 \leq b \leq n$  and any letter  $c$  not equal to the two letters adjacent to gap  $b$ , there exists some  $a$  and  $\ell \geq 2$  such that  $(a, \ell, b, c)$  is a square-completing quadruple in  $W$ . Let  $\mathcal{A}$  be the set consisting of one such quadruple for each choice of  $(b, c)$ . On one hand, by construction we have

$$|\mathcal{A}| \geq (k - 2)(n + 1).$$

On the other hand, by Corollary 2.4, we have

$$\begin{aligned} |\mathcal{A}| &= \sum_{\ell=2}^{\infty} |\mathcal{A}_\ell| \\ &= \sum_{\ell=2}^{319} |\mathcal{A}_\ell| + \sum_{j=0}^{\infty} \sum_{\ell \in [320 \cdot 2^j, 320 \cdot 2^{j+1})} |\mathcal{A}_\ell| \\ &\leq \sum_{\ell=2}^{319} \frac{2n}{\ell - 1} + \sum_{j=0}^{\infty} \frac{320n}{320 \cdot 2^j} < 14.7n. \end{aligned}$$

Thus we conclude that  $k < 17$ , as desired. □

*Proof of Corollary 1.4.* We showed that the number of square-completing quadruples  $(a, \ell, b, c)$  with  $\ell \geq 2$  such that no two elements share the same  $(b, c)$  in a square-free word  $W$  of length  $n$  is less than  $14.7n$ . Thus, the number of ways to insert a letter into  $W$  such that the result is no longer square-free is less than  $16.7n$ . Therefore, if the alphabet size is at least 17, then it is possible to insert a letter into the latter  $\frac{16.7}{17}$  of any square-free word  $W$  such that the result is square-free. Therefore, for any positive integers  $N > 0$  and  $i > 57N$ , the length of  $G'_i$  is at least  $i - \frac{16.7}{17}i > N$ , so the length  $N$  prefix of  $G_i$  and  $G_{i+1}$  is the same. So the prefix of  $\{G_i\}$  will stabilize and the sequence converges to an infinite limit word.  $\square$

## Acknowledgements

The authors thank Prof. Joseph Gallian, Prof. Jarosław Grytczuk, Prof. Lucas Mol, Prof. Bartłomiej Pawlik, Benjamin Przybocki and the anonymous referees for their invaluable comments to the paper.

## References

- [1] Jean Berstel. Axel Thue's papers on repetitions in words: a translation. *Publications du LaCIM*, Vol. 20, Université du Québec a Montréal, 1995.
- [2] Jean Berstel and Dominique Perrin. The origins of combinatorics on words, *European Journal of Combinatorics*, 28:996–1022, 2007. doi:10.1016/j.ejc.2005.07.019.
- [3] Roger Charles Entringer, Douglas E. Jackson, and J. A. Schatz. On nonrepetitive sequences. *Journal of Combinatorial Theory, Series A*, 16:159–164, 1974. doi:10.1016/0097-3165(74)90041-7.
- [4] Jarosław Grytczuk, Hubert Kordulewski, and Artur Niewiadomski. Extremal square-free words. *Electronic Journal of Combinatorics*, 27(1):P1.48, 2020. doi:10.37236/9264.
- [5] Jarosław Grytczuk, Hubert Kordulewski, and Bartłomiej Pawlik. Square-free extensions of words. *Journal of Integer Sequences*, 24, 2021, No. 8, Art. 21.8.7.
- [6] Lucas Mol and Narad Rampersad. Lengths of extremal square-free ternary words. *Contributions to Discrete Mathematics*, Vol. 16 No. 1 (2021). doi:10.11575/cdm.v16i1.69831.
- [7] Natalya Ter-Saakov and Emily Zhang. Extremal Pattern-Avoiding Words. *arXiv preprint*. URL: <https://arxiv.org/abs/2009.10186>.
- [8] Axel Thue. Über unendliche zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, 7:1–22, 1906.