Extremal overlap-free and extremal β-free binary words

Lucas Mol Narad Rampersad* Jeffrey Shallit†

Department of Mathematics and Statistics
The University of Winnipeg
Winnipeg, MB, Canada

{l.mol,n.rampersad}@uwinnipeg.ca

School of Computer Science
University of Waterloo
Waterloo, ON, Canada

shallit@uwaterloo.ca

Submitted: July 9, 2020; Accepted: Nov 12, 2020; Published: Dec 11, 2020

Abstract

An overlap-free (or β-free) word w over a fixed alphabet Σ is *extremal* if every word obtained from w by inserting a single letter from Σ at any position contains an overlap (or a factor of exponent at least β, respectively). We find all lengths which admit an extremal overlap-free binary word. For every "extended" real number β such that $2^+ \leqslant \beta \leqslant 8/3$, we show that there are arbitrarily long extremal β-free binary words.

Mathematics Subject Classifications: 68R15

1 Introduction

Throughout, we use standard definitions and notation from combinatorics on words (see [12]). For every integer $n \geqslant 2$, we let Σ_n denote the alphabet $\{\mathtt{0}, \mathtt{1}, \ldots, \mathtt{n\text{-}1}\}$. The word u is a *factor* of the word w if we can write $w = xuy$ for some (possibly empty) words x, y. A *square* is a word of the form xx, where x is nonempty. An *overlap* is a word of the form $axaxa$, where a is a letter and x is a (possibly empty) word. A word is *square-free* if it contains no square as a factor, and *overlap-free* if it contains no overlap as a factor. Early in the twentieth century, Norwegian mathematician Axel Thue [22, 23] demonstrated that one can construct arbitrarily long square-free words over a ternary

*The work of Narad Rampersad is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number 2019-04111].

†The work of Jeffrey Shallit is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number 2018-04118].

alphabet, and arbitrarily long overlap-free words over a binary alphabet. For an English translation of Thue's work, see [2]. Thue's work is recognized as the beginning of the field of combinatorics on words [3].

Let w be a word over a fixed alphabet Σ. An *extension* of w is a word of the form $w'aw''$, where $a \in \Sigma$, and $w'w'' = w$ for some possibly empty words $w', w'' \in \Sigma^*$. For example, over the English alphabet, the English word `pans` has extensions including the English words `spans`, `plans`, `pawns`, `pants`, and `pansy`. The word w is *extremal square-free* if w is square-free, and every extension of w contains a square. For example, the word

$$\texttt{abcabacbcabcbabcabacbcabc}$$

of length 25 is an extremal square-free word of minimum length over the alphabet $\{\texttt{a}, \texttt{b}, \texttt{c}\}$. The concept of extremal square-free word was recently introduced by Grytczuk et al. [11], who demonstrated that there are arbitrarily long extremal square-free words over a ternary alphabet. Two of the present authors [14] adapted their ideas to find all lengths admitting extremal square-free ternary words.

In this paper, we consider some variations of extremal square-free words, with a focus on the binary alphabet $\Sigma_2 = \{\texttt{0}, \texttt{1}\}$. We begin by considering extremal overlap-free words, as suggested by Grytczuk et al. [11]. For a word w over a fixed alphabet Σ, we say that w is *extremal overlap-free* if w is overlap-free, and every extension of w contains an overlap. For example, the word `0010011011` of length 10 is an extremal overlap-free word of minimum length over Σ_2.

While there is an extremal square-free ternary word of every sufficiently large length, the same cannot be said for extremal overlap-free binary words. Our first main result is the following characterization of the lengths of extremal overlap-free binary words.

Theorem 1. *Let n be a nonnegative number. Then there is an extremal overlap-free word of length n over the alphabet Σ_2 if and only if n is in the set*

$$\mathcal{N} := \{10, 12\} \cup \{2k\colon\ k \geqslant 10\} \cup \left\{2^k + 1\colon\ k \geqslant 5\right\} \cup \left\{3 \cdot 2^k + 1\colon\ k \geqslant 3\right\}.$$

After proving Theorem 1, we consider a more general problem, which we now provide background for. Let $w = w_1 w_2 \cdots w_n$ be a word, where the w_i's are letters. For an integer $p \geqslant 1$, we say that w has *period* p if $w_{i+p} = w_i$ for all i such that $1 \leqslant i \leqslant n - p$. Note that w may have many periods; the minimal period of w is called *the* period of w. The *exponent* of w is the length of w divided by the period of w. For a real number b, the word w is *b-free* if it contains no factor of exponent greater than or equal to b, and the word w is *b^+-free* if it contains no factor of exponent greater than b. So 2-free words are exactly the square-free words, and 2^+-free words are exactly the overlap-free words.

For ease of writing, we unify the notions of b-free word and b^+-free word by considering β-free words, where β belongs to the set of "extended real numbers". Let $\mathbb{R}_{\text{ext}}$ denote the set of extended real numbers, consisting of all real numbers, together with all real numbers with a $+$, where x^+ covers x, and the inequalities $y \leqslant x$ and $y < x^+$ are equivalent. For $\beta \in \mathbb{R}_{\text{ext}}$, we say that w is β-free if no factor of w has exponent greater than or equal to β.

Definition 2. Let w be a word over a fixed alphabet Σ, and let $\beta \in \mathbb{R}_{\text{ext}}$. We say that w is *extremal β-free* if w is β-free, and every extension of w contains a factor of exponent greater than or equal to β.

We consider the following problem.

Problem 3. For which $\beta \in \mathbb{R}_{\text{ext}}$ do there exist arbitrarily long extremal β-free words over Σ_2?

On the affirmative side, by Theorem 1, we know that there are arbitrarily long extremal 2^+-free words over Σ_2. On the negative side, every binary word of length at least 4 contains a square, so it follows that for all $\beta \leqslant 2$, there do not exist arbitrarily long extremal β-free words over Σ_2. We make some further progress on Problem 3 on the affirmative side by establishing the following theorem.

Theorem 4. *Let $\beta \in \mathbb{R}_{ext}$ satisfy $2^+ \leqslant \beta \leqslant 8/3$. Then there are arbitrarily long extremal β-free words over Σ_2.*

We also make the following conjecture.

Conjecture 5. There is some number $\alpha \in \mathbb{R}_{\text{ext}}$ such that for all $\beta \in \mathbb{R}_{\text{ext}}$ satisfying $\beta \geqslant \alpha$, there are no extremal β-free words over Σ_2.

It is possible that Conjecture 5 is true with $\alpha = 8/3^+$, but we have only very weak computational evidence supporting this. If one could show that Conjecture 5 is true with $\alpha = 8/3^+$, then it would completely answer Problem 1.3.

The layout of the remainder of the paper is as follows. We prove Theorem 1 in Section 2 and Section 3. We consider the even lengths in Section 2, and the odd lengths in Section 3. We prove Theorem 4 in Section 4. We conclude with a discussion of some open problems and conjectures over larger alphabets.

2 Extremal overlap-free words of even length

In this section, we characterize the even lengths for which there are extremal overlap-free binary words. Throughout the remainder of the paper, we let $\mu : \Sigma_2^* \to \Sigma_2^*$ denote the *Thue-Morse morphism*, defined by $\mu(\mathtt{0}) = \mathtt{01}$ and $\mu(\mathtt{1}) = \mathtt{10}$. The infinite *Thue-Morse word* $\mathbf{t}$ is the limit of the words obtained by starting with the letter $\mathtt{0}$ and iterating the morphism μ, that is, we have $\mathbf{t} = \mu^\omega(\mathtt{0})$. The Thue-Morse word is the prototypical example of a 2-automatic sequence, and this means that the automatic theorem-proving software `Walnut` [16] can be used to prove results about factors of the Thue-Morse word. We begin with a lemma that is used frequently in the rest of the paper.

Lemma 6. *Let $w \in \Sigma_2^*$ be an overlap-free word of length at least 10, and write $w = w'w''$ with $|w'|, |w''| \geqslant 5$. Then for every letter $a \in \Sigma_2$, the extension $w'aw''$ contains an overlap of period at most 3 (and hence a factor of exponent at least 7/3).*

Proof. It suffices to check the lemma statement for all overlap-free words in Σ_2^* of length exactly 10, which is completed easily by computer. $\square$

Definition 7. A word $w \in \Sigma_2^*$ is called *earmarked* if all of the following conditions are satisfied:

(i) w is overlap-free;

(ii) the length 4 prefix of w is in $\{\texttt{0010}, \texttt{1101}\}$; and

(iii) the length 4 suffix of w is $\texttt{0100}$.

The next lemma explains our interest in earmarked words, as it describes a map that takes every earmarked word of length $n \geqslant 8$ to an extremal overlap-free word of length $2n$. We note that this map was used previously by Cassaigne [4] in counting the overlap-free binary words. In fact, Cassaigne used the family of nine distinct maps defined by applying the Thue-Morse morphism and then leaving alone, complementing (i.e., changing from 0 to 1 or vice versa), or removing the first and last letters. While parts of the proof of the next lemma can be gleaned from the work of Cassaigne, we include the entire proof for completeness.

Lemma 8. *Let u be an earmarked word of length at least 8. Let w be the word obtained from $v = \mu(u)$ by complementing the first and last letters. Then w is both earmarked and extremal overlap-free.*

Proof. Assume that u has prefix 0010; the case that u has prefix 1101 is handled similarly. So we may write $u = \texttt{0010}u'\texttt{0100}$ for some word $u' \in \Sigma_2^*$. It follows that

$$w = \texttt{11011001}\mu(u')\texttt{01100100}$$

So w has length 4 prefix $\texttt{1101}$ and length 4 suffix $\texttt{0100}$.

We now show that w is overlap-free. First of all, since u and μ are overlap-free, we see that v is overlap-free. Now suppose that w contains the overlap x. Since v is overlap-free, we see that x must be either a prefix or a suffix of w. Assume that x is a prefix of w; the case that x is a suffix of w is handled similarly. Since the word $\texttt{11011}$ may only appear as a prefix or a suffix of an overlap-free word, we conclude that the period of x is at most 4. But by inspection, there is no such overlap in w.

Finally, we show that w is extremal overlap-free. By Lemma 6, it suffices to check that every extension of w of the form $w'aw''$, where $w = w'w''$, $a \in \Sigma_2$, and either $|w'| \leqslant 4$ or $|w''| \leqslant 4$, contains an overlap. We complete this check by inspection. $\square$

Lemma 9. *Let $n \geqslant 10$ be an integer satisfying $n \not\equiv 0 \pmod 4$. Then there is an earmarked word of length n.*

Proof. We use the automatic theorem-proving software `Walnut` [16] to show that the Thue-Morse word $\mathbf{t}$ contains a factor u of length $n-4$ such that the word $u\texttt{0100}$ is earmarked. The interested reader can verify our results in `Walnut`; the complete code that we used

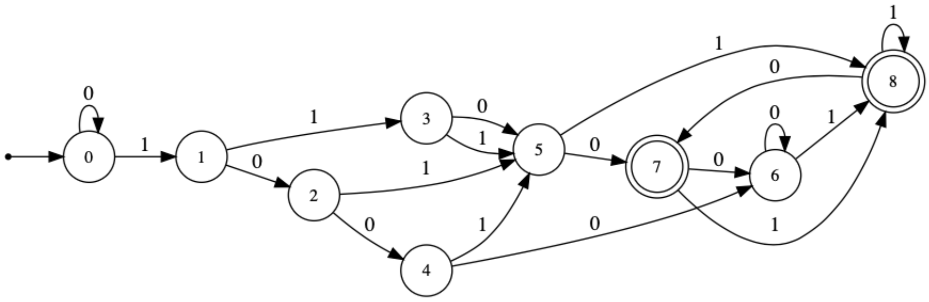

Figure 1: The automaton accepting those $(n)_2$ for which the Thue-Morse word contains a factor v of length $n-4$ such that $v\texttt{0100}$ is earmarked.

can be found in Appendix A. We essentially adapt the predicates used by Clokie, Gabric, and Shallit [5, Theorem 1].

First, we create a predicate $\texttt{overlap}(i, n, p, s)$ which evaluates to $\texttt{true}$ if the word $u\texttt{0100}$ contains an overlap of period p with $p \geqslant 1$ beginning at index $i - s$, where $u = \mathbf{t}[s..s + n - 5]$. We use a straightforward modification of the method described by Clokie, Gabric, and Shallit [5, Proof of Theorem 1] to do so. Next, we create a predicate $\texttt{earmarked}(n, s)$ which evaluates to $\texttt{true}$ if the word $u\texttt{0100}$ defined above is earmarked:

$$\begin{aligned}&(n \geqslant 8) \wedge (\mathbf{t}[s..s+3] \in \{\texttt{0010}, \texttt{1101}\})\\&\quad \wedge (\forall i, p\ ((p \geqslant 1) \wedge (i \geqslant s) \wedge (i - s + 2p < n)) \Rightarrow \neg(\texttt{overlap}(i, n, p, s)))\end{aligned}$$

Finally, the predicate

$$\texttt{testEarmarked}(n) := \exists s\ \texttt{earmarked}(n, s)$$

evaluates to *true* if there is some length $n-4$ factor v of the Thue-Morse word such that $v\texttt{0100}$ is earmarked. The automaton for $\texttt{testEarmarked}(n)$ is shown in Figure 1. By inspection, this automaton accepts all integers $n \geqslant 10$ such that $n \not\equiv 0 \pmod 4$. $\square$

Lemma 10. *Let $n \geqslant 10$ be an integer that is not a power of two. Then there is an earmarked word of length n.*

Proof. By Lemma 9, we may assume that $n \equiv 0 \pmod 4$. Since n is not a power of two, we may write $2^k < n < 2^{k+1}$ for some $k \geqslant 3$. We proceed by induction on k. If $k \leqslant 4$, then $n \in \{12, 20, 24, 28\}$. It is easily verified by computer that the following words (found by computer search) are earmarked:

Length 12: `001001100100`
Length 20: `00100110100101100100`
Length 24: `110110010110100101100100`
Length 28: `1101100110100101101001100100`

So we may assume that $k \geqslant 5$. Let $m = n/2$. Note that m is not a power of two, and that $10 < 2^{k-1} < m < 2^k$. If $m \not\equiv 0 \pmod 4$, then there is an earmarked word of length m by Lemma 9. If $m \equiv 0 \pmod 4$, then there is an earmarked word of length m by the induction hypothesis. So either way, there is an earmarked word of length m. By Lemma 8, there is an earmarked word of length $2m = n$. □

Corollary 11. *Let $n \geqslant 20$ be an even integer that is not a power of two. Then there is an extremal overlap-free word of length n.*

Proof. By Lemma 10, there is an earmarked word u of length $m = n/2$. By Lemma 8, the word w of length n obtained from $\mu(u)$ by complementing the first and last letters is extremal overlap-free. □

Lemma 12. *For every integer $k \geqslant 5$, there is an extremal overlap-free word of length 2^k.*

Proof. Let $s = \mathtt{0}\mu(\mathtt{0011001})\mathtt{1} = (\mathtt{00101101})^2$. We claim that the word $\mu^\ell(s)$ is extremal overlap-free for every integer $\ell \geqslant 1$. Since s has length 16, the word $\mu^\ell(s)$ has length $2^{\ell+4}$, and hence the theorem statement follows.

Fix $\ell \geqslant 1$, and let $w = \mu^\ell(s)$. If $\ell \leqslant 2$, then we verify that w is extremal overlap-free by computer, so we may assume that $\ell \geqslant 3$. First note that s is overlap-free, and hence w is overlap-free. It remains to show that every extension of w contains an overlap. Consider an extension $w'aw''$ of w, where $w = w'w''$ and $a \in \Sigma_2$. By Lemma 6, we may assume that $|w'| \leqslant 4$ or $|w''| \leqslant 4$. We consider several cases.

Case I: $|w'| = 0$. Note that w begins with the squares $\mu^\ell(00)$ and $\mu^\ell(s)$. If ℓ is even, then $\mu^\ell(00)$ ends with a $\mathtt{0}$, and $\mu^\ell(s)$ ends with a $\mathtt{1}$. If ℓ is odd, then $\mu^\ell(00)$ ends with a $\mathtt{1}$, and $\mu^\ell(s)$ ends with a $\mathtt{0}$. So either way, the extensions $\mathtt{0}w$ and $\mathtt{1}w$ both contain an overlap.

Case II: $1 \leqslant |w'| \leqslant 4$. Since $\ell \geqslant 3$, we see that w has prefix $\mu^3(\mathtt{0}) = \mathtt{01101001}$. If $|w'| = 1$, then the extension $w'\mathtt{0}w'' = \mathtt{0}w$ contains an overlap by Case I, and the extension $w'\mathtt{1}w''$ contains the overlap $\mathtt{111}$. So we may assume that $2 \leqslant |w'| \leqslant 4$. By inspection, the extension $w'aw''$ contains an overlap of period at most 3.

Case III: $|w''| = 0$. Note that $\mu^\ell(s)$ and $\mu^\ell(\mathtt{101101})$ are square suffixes of $\mu^\ell(s)$ which begin in $\mathtt{0}$ and $\mathtt{1}$, respectively. So both of the extensions $w\mathtt{0}$ and $w\mathtt{1}$ contain an overlap.

Case IV: $1 \leqslant |w''| \leqslant 4$. Since $\ell \geqslant 3$, we see that w has suffix $\mu^3(\mathtt{0}) = \mathtt{01101001}$ if ℓ is even, and suffix $\mu^3(\mathtt{1}) = \mathtt{10010110}$ if ℓ is odd. Either way, the remainder of the proof is similar to that of Case II. □

Proposition 13. *Let n be a nonnegative even number. Then there is an extremal overlap-free word of length n over the alphabet Σ_2 if and only if $n \in \mathcal{N}$.*

Proof. If $n \in \{0, 2, 4, 6, 8, 14, 16, 18\}$, then an exhaustive backtracking search shows that no extremal overlap-free word of length n exists over Σ_2. The words $\mathtt{0010011011}$ and $\mathtt{001001100100}$, of lengths 10 and 12, respectively, are extremal overlap-free. So suppose that $n \geqslant 20$. If n is a power of two, then there is an extremal overlap-free word of length n by Corollary 12. If n is not a power of two, then there is an extremal overlap-free word of length n by Lemma 11. □

3 Extremal overlap-free words of odd length

In this section, we characterize the odd lengths for which there are extremal overlap-free binary words. We need two classical results from the theory of overlap-free binary words. The first is the so-called *factorization theorem* of Restivo and Salemi [20] (see also [1, Proposition 1.7.5(a)]).

Theorem 14. *Let $x \in \{\mathtt{0},\mathtt{1}\}^*$ be overlap-free. Then there exist $u, v \in \{\varepsilon, \mathtt{0}, \mathtt{1}, \mathtt{00}, \mathtt{11}\}$ and an overlap-free word y such that $x = u\mu(y)v$. Furthermore, this factorization is unique if $|x| \geqslant 7$.*

Words u and v are *conjugates* if there exist words x and y such that $u = xy$ and $v = yx$, i.e., if they are cyclic shifts of one another. Let $w \in \Sigma^*$. The *circular word* formed from w is the set of all conjugates of w. Thue [2, Proposition 2.13 (Satz 13)] characterized the circular overlap-free binary words, which also yields a characterization of the overlap-free binary squares (see also the work of Shelton and Soni [21]).

Define

$$A = \{\mathtt{00}, \mathtt{11}, \mathtt{010010}, \mathtt{101101}\}$$

and

$$\mathcal{A} = \bigcup_{k \geqslant 0} \mu^k(A).$$

Theorem 15. *The overlap-free binary squares are the conjugates of the words in $\mathcal{A}$.*

Remark 16. From Theorems 14 and 15, we deduce that if vv is an overlap-free binary square of length greater than 6, then vv can be written in exactly one of the following two forms: $vv = \mu(zz)$ or $vv = \overline{a}\mu(z)a$ for some $a \in \{\mathtt{0},\mathtt{1}\}$ and some $z \in \{\mathtt{0},\mathtt{1}\}^*$.

Proposition 17. *Let u be an extremal overlap-free binary word of odd length. Then either $|u| = 2^k + 1$ or $|u| = 3 \cdot 2^k + 1$ for some k.*

Proof. By Theorem 14, we can, without loss of generality, consider two possible forms for u: either $u = \mu(y)a$ or $u = bb\mu(y)a$ for some $a, b \in \{\mathtt{0},\mathtt{1}\}$. If u is extremal overlap-free, then both ua and $u\overline{a}$ end in overlaps. Consequently, the word u ends in at least two distinct squares. Let vv be the longest square suffix of u.

Suppose first that $|vv| > 6$. By Remark 16, we see that $vv = \overline{a}\mu(z)a$ for some word z. If vv is a proper factor of $\mu(y)a$, then vv is preceded by a in u; however, since v ends with a, the word avv is an overlap in u, which is a contradiction. We conclude that $u = \overline{a}\overline{a}\mu(z)a = \overline{a}vv$, and hence that either $|u| = 2^k + 1$ or $|u| = 3 \cdot 2^k + 1$ for some k, as required.

Now consider the case $|vv| \leqslant 6$. Since u ends in two distinct squares, these squares are both conjugates of words in $A \cup \{\mathtt{0101}\}$, and, since one must be a suffix of the other, we observe that the only possibilities for these two squares are aa and $\overline{a}aa\overline{a}aa$. However, $\overline{a}aa\overline{a}aa$ is not a suffix of a word of either the form $\mu(y)a$ or the form $bb\mu(y)a$. This contradiction completes the proof. $\square$

The proof of Lemma 17 tells us that any extremal overlap-free word of odd length can be obtained from an overlap-free square by adding a single letter at either the beginning or the end. This led us to the constructions of overlap-free words of odd length given in the next two lemmas.

Lemma 18. *For every integer $k \geqslant 5$, there is an extremal overlap-free word of length $2^k + 1$.*

Proof. Fix $k \geqslant 5$. Let $u = (\mathtt{011})^{-1}\mu^{k-1}(\mathtt{00})\mathtt{011}$. Note that u is a conjugate of $\mu^{k-1}(\mathtt{00})$. In particular, we have that u is a square of length 2^k, and by Theorem 15, we see that u is overlap-free. We claim that the word $v = \mathtt{0}u$ is extremal overlap-free. We first show that v is overlap-free. Since u is overlap-free, it suffices to show that no prefix of v is an overlap. Since v has prefix $\mathtt{00100}$, which never appears again in v, it suffices to check that u does not begin with an overlap of period at most 4, which is easily done by inspection.

It remains to show that every extension of v contains an overlap. Consider an extension $v'av''$ of v, where $v = v'v''$ and $a \in \Sigma_2$. By Lemma 6, we may assume that $|v'| \leqslant 4$ or $|v''| \leqslant 4$. First suppose that $|v'| \leqslant 4$. Note that v has prefix $\mathtt{0}(\mathtt{011})^{-1}\mu^4(\mathtt{0}) = \mathtt{00100110010110}$. By inspection, the extension $v'av''$ contains an overlap of period at most 4. Now suppose that $|v''| \leqslant 4$. Since u is a square with first letter $\mathtt{0}$, and u ends in the square $\mathtt{11}$, the extension va contains an overlap. Thus we may assume that $1 \leqslant |v''| \leqslant 4$. If k is even, then v has suffix $\mu^4(\mathtt{0})\mathtt{011}$, and by inspection, the extension $v'av''$ contains an overlap of period at most 6. If k is odd, then v has suffix $\mu^4(\mathtt{1})\mathtt{011}$, and by inspection, the extension $v'av''$ contains an overlap of period at most 6. ☐

Lemma 19. *For every integer $k \geqslant 3$, there is an extremal overlap-free word of length $3 \cdot 2^k + 1$.*

Proof. Fix $k \geqslant 3$. Let $u = (\mathtt{011})^{-1}\mu^{k-1}(\mathtt{010010})\mathtt{011}$. Note that u is a conjugate of $\mu^{k-1}(\mathtt{010010})$. In particular, we have that u is a square of length $3 \cdot 2^k$, and by Theorem 15, we see that u is overlap-free. We claim that the word $v = \mathtt{0}u$ is extremal overlap-free. The remainder of the proof is strictly analogous to the proof of Lemma 18, and is omitted. ☐

We now prove the analogue of Proposition 13 for odd n.

Proposition 20. *Let n be a nonnegative odd number. Then there is an extremal overlap-free word of length n over the alphabet Σ_2 if and only if $n \in \mathcal{N}$.*

Proof. ($\Leftarrow$) Let $n \in \mathcal{N}$. Since n is odd, we must have either $n = 2^k + 1$ for some $k \geqslant 5$, or $n = 3 \cdot 2^k + 1$ for some $k \geqslant 3$. In the former case, there is an extremal overlap-free word of length n by Lemma 18, and in the latter case, there is an extremal overlap-free word of length n by Lemma 19.

($\Rightarrow$) Suppose that there is an extremal overlap-free word of length n over the alphabet $\{0, 1\}$. By Proposition 17, we must have $n = 2^k + 1$ or $n = 3 \cdot 2^k + 1$ for some k. By exhaustive computer search, there is no extremal overlap-free word of length $2^k + 1$ for $k \leqslant 4$, and no extremal overlap-free word of length $3 \cdot 2^k + 1$ for $k \leqslant 2$. Thus, we conclude that $n \in \mathcal{N}$. ☐

Together, Proposition 13 and Proposition 20 give Theorem 1.

4 Extremal β-free binary words

This section is devoted to the proof of Theorem 4. Another definition facilitates our proof method.

Definition 21. Let w be a word over a fixed alphabet Σ, and let $\alpha, \beta \in \mathbb{R}_{\text{ext}}$ satisfy $1 < \alpha \leqslant \beta$. We say that w is (α, β)-*extremal* if w is α-free, and every extension of w contains a factor of exponent greater than or equal to β.

If w is (α, β)-extremal, then for any $\gamma \in \mathbb{R}_{\text{ext}}$ such that $\alpha \leqslant \gamma \leqslant \beta$, the word w is extremal γ-free. Thus, the following result immediately implies Theorem 4.

Proposition 22. *All of the following hold.*

(a) *There are arbitrarily long* $(2^+, 7/3)$*-extremal binary words.*

(b) *There are arbitrarily long* $(7/3^+, 17/7)$*-extremal binary words.*

(c) *There are arbitrarily long* $(17/7^+, 5/2)$*-extremal binary words.*

(d) *There are arbitrarily long* $(5/2^+, 18/7)$*-extremal binary words.*

(e) *There are arbitrarily long* $(18/7^+, 8/3)$*-extremal binary words.*

We prove the first part of Proposition 22 now.

Proof of Proposition 22(a). Let u be a factor of the Thue-Morse word of the form $\mathtt{011}v\mathtt{110}$, where v is a nonempty word. Note that there are arbitrarily long words of this form. We claim that the word $x = \mathtt{00}\mu^2(\mathtt{11}v\mathtt{11})\mathtt{00}$ is $(2^+, 7/3)$-extremal.

First we show that x is 2^+-free (or in other words, overlap-free). Since u is a factor of the Thue-Morse word, we have that u, and hence $\mu^2(u)$, are overlap-free. Since the word $\mu^2(u)$ contains the word $\mathtt{0}\mu^2(\mathtt{11}v\mathtt{11})\mathtt{0}$ as a factor, any overlap contained in x must be either a prefix or a suffix of x. Suppose without loss of generality that x contains an overlap z as a prefix. Since the factor $\mathtt{00100}$ does not appear in the Thue-Morse word, this factor appears only as a prefix and a suffix of x. So z must have period at most 4. But this is impossible by inspection.

It remains to show that every extension of x contains a factor of exponent at least 7/3. Consider an extension $x'ax''$ of x, where $x = x'x''$ and $a \in \Sigma_2$. By Lemma 6, we may assume that $|x'| \leqslant 4$ or $|x''| \leqslant 4$. First suppose that $|x'| \leqslant 4$. Note that x has prefix $\mathtt{00}\mu^2(\mathtt{11}) = \mathtt{0010011001}$. By inspection, the extension $x'ax''$ contains a factor of exponent at least 7/3. The case that $|x''| \leqslant 4$ is handled by a symmetric argument. $\square$

One of the main tools that we use to prove Proposition 22 parts (b)-(e) is the following extension of a lemma due to Ochem [17, Lemma 2.1]. A morphism $f : \Sigma^* \to \Delta^*$ is called q-*uniform* if $|f(a)| = q$ for all $a \in \Sigma$, and is called *synchronizing* if for any $a, b, c \in \Sigma$ and $u, v \in \Delta^*$, if $f(ab) = uf(c)v$, then either $u = \varepsilon$ and $a = c$, or $v = \varepsilon$ and $b = c$.

Lemma 23. *Let $a, b \in \mathbb{R}$ satisfy $1 < a < b$. Let $\alpha \in \{a, a^+\}$ and $\beta \in \{b, b^+\}$. Let $h\colon \Sigma^* \to \Delta^*$ be a synchronizing q-uniform morphism. If $h(w)$ is β-free for every α-free word w such that*

$$|w| \leqslant \max\left\{\frac{2b}{b-a}, \frac{2(q-1)(2b-1)}{q(b-1)}\right\},$$

then $h(z)$ is β-free for every α-free word $z \in \Sigma^*$.

Proof. Suppose that there is an α-free word w such that the word $W = h(w)$ contains a factor of exponent greater than or equal to β, and assume without loss of generality that w is a shortest word satisfying this property. We will show that $|w| \leqslant \max\left\{\frac{2b}{b-a}, \frac{2(q-1)(2b-1)}{q(b-1)}\right\}$, which gives the theorem statement.

Let X be a factor of W of exponent greater than or equal to β. Let P be the period of X, and write $X = UV$, where $|U| = P$. Since X has period P, we can also write $X = VU'$ for some word $U' \in \Delta^*$. Let $R = |V|$. Then we have $\frac{P+R}{P} \geqslant b$, or equivalently $P \leqslant \frac{R}{b-1}$.

First suppose that $R \leqslant 2q - 2 = 2(q-1)$. Then we have

$$|X| = P + R \leqslant \tfrac{R}{b-1} + R = \frac{Rb}{b-1} \leqslant \frac{2(q-1)b}{b-1}.$$

By the minimality of w, we must have $|w| \leqslant \frac{|X|-2}{q} + 2$. Putting this together with the above bound on $|X|$, we find $|w| \leqslant \frac{2(q-1)(2b-1)}{q(b-1)}$.

Now suppose that $R \geqslant 2q - 1$. Write $V = V_1 h(v) V_2$ for some word $v \in \Sigma^*$, where the word V_1 is a proper suffix of a block of h, and the word V_2 is a proper prefix of a block of h. Let $r = |v|$. Since $R \geqslant 2q-1$, we must have $r \geqslant 1$. Further, since $|V_1|, |V_2| < q$, we have $R < qr + 2q$. Similarly, write $X = X_1 h(x) X_2$ for some word $x \in \Sigma^*$, where the word X_1 is a proper suffix of a block of h, and the word X_2 is a proper prefix of a block of h. Since $X = UV = VU'$, and since h is synchronizing, it must be the case that $X_1 = V_1$ and $X_2 = V_2$. It follows that we may write $x = uv = vu'$ for some words $u, u' \in \Sigma^*$, i.e., the word x has period $|u|$. Let $p = |u|$. Note that $h(u) = V_1^{-1} U V_1$, so $|h(u)| = |U| = P$, and hence $qp = P$.

Since w is α-free, we must have $\frac{p+r}{p} \leqslant a$, or equivalently $r \leqslant (a-1)p$. Now

$$qp = P \leqslant \frac{R}{b-1} < \frac{qr+2q}{b-1} \leqslant q \cdot \frac{r+2}{b-1} \leqslant q \cdot \frac{(a-1)p+2}{b-1},$$

from which we conclude that $p < \frac{(a-1)p+2}{b-1}$, or equivalently, that $p < \frac{2}{b-a}$. Finally, by the minimality of w, we must have

$$|w| \leqslant 2 + p + r \leqslant 2 + ap < 2 + \frac{2a}{b-a} = \frac{2b}{b-a}.$$

We conclude in either case that $|w| \leqslant \max\left\{\frac{2b}{b-a}, \frac{2(q-1)(2b-1)}{q(b-1)}\right\}$, as desired. □

We are now ready to prove the remaining parts of Proposition 22. We use the following terminology in the proof. Let w be a word over a fixed alphabet Σ. A *left extension* of w is a word of the form aw, where $a \in \Sigma$. A *right extension* of w is a word of the form wa, where $a \in \Sigma$. An *internal extension* of w is a word of the form $w'aw''$, where $|w'|, |w''| \geqslant 1$, we have $a \in \Sigma$, and $w'w'' = w$.

Proof of Proposition 22(b). Let $u \in \Sigma_3^*$ be a square-free word of length at least 3, and write $u = avb$, where $a, b \in \Sigma_3$. Define $f : \Sigma_3^* \to \Sigma_2^*$ by

$$
\begin{aligned}
f(0) &= \mathtt{001011001101100100110100110110010011} \\
f(1) &= \mathtt{001011001101100100110110010110010011} \\
f(2) &= \mathtt{001011001101100101100100110110010011}.
\end{aligned}
$$

Let $r = \mathtt{1100110010011}$ and $s = \mathtt{001001}$. We claim that the word $w = rf(v)s$ is $(7/3^+, 17/7)$-extremal.

First of all, we verify the following statements by computer for every letter $c \in \Sigma_3$:

- Every internal extension of the word $f(c)$ contains a factor of exponent at least $17/7$.
- Every left extension and every internal extension of the word $rf(c)$ contains a factor of exponent at least $17/7$.
- Every right extension and every internal extension of the word $f(c)s$ contains a factor of exponent at least $17/7$.

It now follows easily that every extension of the word $w = rf(v)s$ contains a factor of exponent at least $17/7$. The only extensions of w not checked above are those obtained by inserting a letter between two blocks of f. Since every block of f begins in $\mathtt{00}$ and ends in $\mathtt{11}$, every such extension contains a cube.

It remains to show that w is $7/3^+$-free. We first show that $f(u)$ is $7/3^+$-free. Note that f is 36-uniform, and we verify by computer that f is synchronizing. Thus, by Lemma 23, it suffices to check that $f(x)$ is $7/3^+$-free for every square-free word $x \in \Sigma_3^*$ such that $|x| \leqslant 14$, which we verify by computer. Note that every block of $f(u)$ has prefix $s' = \mathtt{0010}$ and suffix $r' = \mathtt{0110010011}$. So $f(u)$ contains the word $w' = r'f(v)s'$, and hence w' is $7/3^+$-free. Note that $s = s'\mathtt{01}$ and $r = \mathtt{110}r'$, so

$$w = rf(v)s = \mathtt{110}r'f(v)s'\mathtt{01}.$$

Suppose that w contains a factor z of exponent greater than $7/3$. Then z begins at one of the first three letters of w, or ends at one of the last two letters of w. Suppose first that z begins at one of the first three letters of w. We claim that the factor $t_r = \mathtt{00110010011}$, which occurs starting at the third letter of w, occurs only once in w. To establish this claim, we verify the following by computer:

- For every $c \in \Sigma_3$, the word t_r occurs exactly once in the word $rf(c)$, and does not occur in the word $f(c)s$.

- For every square-free word $y \in \Sigma_3^*$ of length 2, the word t_r does not occur in $f(y)$.

So we see that the period of z is at most 13. However, this possibility is ruled out by computer check. So we may assume that z ends at one of the last two letters of w. By a computer check similar to the one used for t_r, we verify that the factor $t_s = \texttt{001001100100}$, which occurs ending at the second last letter of w, occurs only once in w. So again, we see that the period of z is at most 13. This possibility is ruled out by computer check. $\square$

The proofs of Proposition 22(c)–(e) are similar to the proof of Proposition 22(b), so we describe the construction used in each of these proofs below, but omit the remaining details.

Proof of Proposition 22(c). Let $u \in \Sigma_3^*$ be a square-free word of length at least 3, and write $u = avb$, where $a, b \in \Sigma_3$. Define $f : \Sigma_3^* \to \Sigma_2^*$ by

$$\begin{aligned} f(\texttt{0}) &= \texttt{001001100101100100110010110100110010110011011} \\ f(\texttt{1}) &= \texttt{001001100101100110100101100110110010110011011} \\ f(\texttt{2}) &= \texttt{001001100101100110100110110011010010110011011}. \end{aligned}$$

Let $r = \texttt{00110110011011}$ and $s = \texttt{00100110010011}$. By a method similar to the one used in the proof of Proposition 22(b), one can show that the word $w = rf(v)s$ is $(17/7^+, 5/2)$-extremal. $\square$

Proof of Proposition 22(d). Let $u \in \Sigma_3^*$ be a square-free word of length at least 3, and write $u = avb$, where $a, b \in \Sigma_3$. Define $f : \Sigma_3^* \to \Sigma_2^*$ by

$$\begin{aligned} f(\texttt{0}) &= \texttt{0011011001001100101100110110010011} \\ f(\texttt{1}) &= \texttt{0011011001001101001101100110010011} \\ f(\texttt{2}) &= \texttt{0011011001001101100110100110010011}. \end{aligned}$$

Let $r = \texttt{00110110011011001010011}$ and $s = \texttt{00110101100100110010011}$. By a method similar to the one used in the proof of Proposition 22(b), one can show that the word $w = rf(v)s$ is $(5/2^+, 18/7)$-extremal. $\square$

Proof of Proposition 22(e). Let $u \in \Sigma_3^*$ be a square-free word of length at least 3, and write $u = avb$, where $a, b \in \Sigma_3$. Define $f : \Sigma_3^* \to \Sigma_2^*$ by

$$\begin{aligned} f(\texttt{0}) &= \texttt{0011011001001100101100110110010011} \\ f(\texttt{1}) &= \texttt{0011011001001101100110100110010011} \\ f(\texttt{2}) &= \texttt{0011011001101001100100110110010011}. \end{aligned}$$

Let $r = \texttt{01101100110110011001010011}$ and $s = \texttt{00110101100110010011001001}$. By a method similar to the one used in the proof of Proposition 22(b), one can show that the word $w = rf(v)s$ is $(18/7^+, 8/3)$-extremal. $\square$

5 Conclusion

We close with a discussion of some related problems over larger alphabets. First of all, we have the following general problem which subsumes Problem 3.

Problem 24. Let $n \geqslant 2$ be an integer. For which $\beta \in \mathbb{R}_{\text{ext}}$ do there exist arbitrarily long extremal β-free words over Σ_n?

For every integer $n \geqslant 2$, let B_n denote the set of all $\beta \in \mathbb{R}_{\text{ext}}$ such that there exist arbitrarily long extremal β-free words over Σ_n. While it seems plausible that B_n is an interval for every n, it is not immediately obvious to us that this is the case.

We note that Dejean's theorem gives us a partial answer to Problem 24. The *repetition threshold* for n letters, denoted $\mathrm{RT}(n)$, is defined by

$$\mathrm{RT}(n) = \inf\{b \in \mathbb{R}\colon \text{ there are arbitrarily long } b\text{-free words over } \Sigma_n\}.$$

Dejean's theorem, originally conjectured by Dejean [10], and confirmed through the work of many authors [10, 8, 7, 9, 19, 6, 15, 18], states that

$$\mathrm{RT}(n) = \begin{cases} 2, & \text{if } n = 2; \\ 7/4, & \text{if } n = 3; \\ 7/5, & \text{if } n = 4; \\ n/(n-1), & \text{if } n \geqslant 5. \end{cases}$$

In fact, for every $n \geqslant 2$, it is known that there are only finitely many $\mathrm{RT}(n)$-free words over n letters, but infinitely many $\mathrm{RT}(n)^+$-free words over n letters. Thus, if there are arbitrarily long extremal β-free words over Σ_n, then $\beta > \mathrm{RT}(n)$.

Conjecture 25. For every $n \geqslant 2$, there are arbitrarily long extremal $\mathrm{RT}(n)^+$-free words over Σ_n.

We define the *extremal repetition threshold* over n letters, denoted $\mathrm{ERT}(n)$, by

$$\mathrm{ERT}(n) = \sup\left\{b \in \mathbb{R}\colon \text{ there are arbitrarily long extremal } b^+\text{-free words over } \Sigma_n\right\}.$$

By Theorem 4, we know that $\mathrm{ERT}(2) \geqslant 8/3$. From the work of Grytczuk et al. [11], we know that $\mathrm{ERT}(3) \geqslant 2$. It may be the case that $\mathrm{ERT}(2) = 8/3$ and $\mathrm{ERT}(3) = 2$, but we have only weak computational evidence supporting this conjecture.

If Conjecture 25 is true, then $\mathrm{ERT}(n) \geqslant \mathrm{RT}(n)$ for every $n \geqslant 2$. We conjecture further that $\mathrm{ERT}(n)$ is finite for every $n \geqslant 2$. In fact, we make the following stronger conjecture, which subsumes Conjecture 5.

Conjecture 26. Let $n \geqslant 2$ be an integer. Then there is some number $\alpha_n \in \mathbb{R}_{\text{ext}}$ such that for all $\beta \in \mathbb{R}_{\text{ext}}$ satisfying $\beta \geqslant \alpha_n$, there are no extremal β-free words over Σ_n.

Finally, we submit the following problem, which appears to be quite difficult.

Problem 27. For every $n \geqslant 2$, find $\mathrm{ERT}(n)$ and the smallest number α_n for which Conjecture 26 holds (if the conjecture is true). It is possible that we have $\alpha_n = \mathrm{ERT}(n)^+$ for every n.

Acknowledgements

The authors wish to thank Trevor Clokie for helpful discussions, and the anonymous referee for a careful reading of the paper and helpful comments.

References

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

[2] J. Berstel, Axel Thue's papers on repetitions in words: A translation, *Publications du LaCIM (Université du Québec à Montréal)*, vol. 20, 1995.

[3] J. Berstel and D. Perrin, The origins of combinatorics on words, *European J. Combin.* **28** (2007), 996-1022.

[4] J. Cassaigne, Counting overlap-free binary words, in *Proc. 10th Annual Symposium on Theoretical Aspects of Comput. Sci. (STACS 1993)*, Lect. Notes in Comput. Sci., Vol. 665, Springer, 1993, pp. 216–225.

[5] T. Clokie, D. Gabric, and J. Shallit, Circularly squarefree words and unbordered conjugates: A new approach, in *Proc. 8th Intl. Conf. Words 2019 (WORDS 2019)*, Lect. Notes in Comput. Sci., Vol. 11682, Springer, 2019, pp. 133–144.

[6] A. Carpi, On Dejean's conjecture over large alphabets, *Theoret. Comput. Sci.* **385** (2007), 137–151.

[7] J. D. Currie and N. Rampersad, Dejean's conjecture holds for $N \geqslant 27$, *RAIRO - Theor. Inform. Appl.* **43** (2009), 775–778.

[8] J. D. Currie and N. Rampersad, Dejean's conjecture holds for $n \geqslant 30$, *Theoret. Comput. Sci.* **410** (2009), 2885–2888.

[9] J. D. Currie and N. Rampersad, A proof of Dejean's conjecture, *Math. Comp.* **80** (2011), 1063–1070.

[10] F. Dejean, Sur un théorème de Thue, *J. Combin. Theory Ser. A* **13** (1972), 90–99.

[11] J. Grytczuk, H. Kordulewski, and A. Niewiadomski, Extremal square-free words, *Electron. J. Combin.* **27** (2020), #P1.48.

[12] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, 2002.

[13] M. Mohammad-Noori and J. D. Currie, Dejean's conjecture and Sturmian words, *European J. Combin.* **28** (2007), 876–890.

[14] L. Mol and N. Rampersad, Lengths of extremal square-free ternary words, preprint, 2020. Available at `https://arxiv.org/abs/2001.11763`.

[15] J. Moulin-Ollagnier, Proof of Dejean's conjecture for alphabets with 5, 6, 7, 8, 9, 10, and 11 letters, *Theoret. Comput. Sci.* **95** (1992), 187–205.

[16] H. Mousavi, Automatic theorem proving in Walnut, preprint, 2016. Available at `https://arxiv.org/abs/1603.06017`.

[17] P. Ochem, A generator of morphisms for infinite words, *RAIRO – Theoret. Inform. Appl.* **40** (2006), 427–441.

[18] J. J. Pansiot, A propos d'une conjecture de F. Dejean sur les répétitions dans les mots, *Discrete Appl. Math.* **7** (1984), 297–311.

[19] M. Rao, Last cases of Dejean's conjecture, *Theoret. Comput. Sci.* **412** (2011), 3010–3018.

[20] A. Restivo and S. Salemi, Overlap-free words on two symbols, in *Automata on infinite words*, Lect. Notes in Comput. Sci., Vol. 192, Springer, 1985, 198–206.

[21] R. O. Shelton and R. P. Soni, Chains and fixing blocks in irreducible binary sequences, *Discrete Math.* **54** (1985), 93–99.

[22] 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.

[23] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.

Appendix A

The free software `Walnut` used in the proof of Lemma 9 is available at https://github.com/hamousavi/Walnut, and a manual for its use is [16]. The complete `Walnut` code used in the proof of Lemma 9 is given below.

```
def overlap "(n >=8) & (1 <= p) & (s <= i) & (i+2*p < s+n)                     &
    (Aj ((j>=i)&(j<i+p+1)  &               (j+p < s+n-4) ) =>  T[j]  = T[j+p]) &
    (Aj ((j>=i)&(j<i+p+1)  &               (j+p = s+n-4) ) =>  T[j]  =  @0)    &
    (Aj ((j>=i)&(j<i+p+1)  & (j < s+n-4) & (j+p = s+n-3) ) =>  T[j]  =  @1)    &
    (Aj ((j>=i)&(j<i+p+1)  & (j < s+n-4) & (j+p = s+n-2) ) =>  T[j]  =  @0)    &
    (Aj ((j>=i)&(j<i+p+1)  & (j < s+n-4) & (j+p = s+n-1) ) =>  T[j]  =  @0)    &
    (Aj ~((j>=i)&(j<i+p+1) & (j = s+n-4) & (j+p = s+n-3)))                     &
    (Aj ~((j>=i)&(j<i+p+1) & (j = s+n-3) & (j+p = s+n-2)))                     &
    (Aj ~((j>=i)&(j<i+p+1) & (j = s+n-3) & (j+p = s+n-1)))":

def earmarked "(n>=8) &
    (((T[s]= @0) & (T[s+1]= @0) & (T[s+2]= @1) & (T[s+3]= @0)) |
    ((T[s]= @1) & (T[s+1]= @1) & (T[s+2]= @0) & (T[s+3]= @1)))    &
    (Ai,p ((1 <= p) & (s <= i) & (i+2*p < s+n)) => ~($overlap(i,n,p,s)))":

def testEarmarked "Es $earmarked(n,s)":
```