NON-REPETITIVE COMPLEXITY OF INFINITE WORDS

JEREMY NICHOLSON AND NARAD RAMPERSAD

Abstract. The non-repetitive complexity function of an infinite word \( x \) (first defined by Moothathu) is the function of \( n \) that counts the number of distinct factors of length \( n \) that appear at the beginning of \( x \) prior to the first repetition of a length-\( n \) factor. We examine general properties of the non-repetitive complexity function, as well as obtain formulas for the non-repetitive complexity of the Thue-Morse word, the Fibonacci word and the Tribonacci word.

1. Introduction

For any infinite word \( x \), there is an associated complexity function \( c_x \) defined as follows: the quantity \( c_x(n) \) is the number of distinct factors of length \( n \) that appear in the word \( x \). Properties of the complexity function for various classes of infinite words have been extensively studied [6, Chapter 4]. Several variants of the complexity function have been introduced and studied, such as palindrome complexity [1] or abelian complexity [17]. In this paper we study the non-repetitive complexity function, which was first introduced by Moothathu [15].

We define the non-repetitive complexity function \( \text{nrc}_x(n) \) for an infinite word \( x \) by

\[
\text{nrc}_x(n) = \max\{m \in \mathbb{N} : x_i \cdots x_{i+n-1} \neq x_j \cdots x_{j+n-1} \text{ for every } 0 \leq i < j \leq m - 1\}.
\]

In other words \( \text{nrc}_x(n) \) is the maximum number of length-\( n \) factors that we see when reading \( x \) from left to right prior to the first repeated occurrence of a length-\( n \) factor. Note that this is not quite the quantity that Moothathu called “non-repetitive complexity”; rather, he defined the non-repetitive complexity of \( x \) to be the quantity

\[
\limsup_{n \to \infty} \frac{\log \text{nrc}_x(n)}{n}
\]

(by analogy with the definition of topological entropy). Nevertheless, in this paper we will refer to the function \( \text{nrc}_x \), as defined above, as the non-repetitive complexity function. Also, although Moothathu introduced the concept, he did not explicitly compute this function for any particular infinite words.

The non-repetitive complexity also bears some resemblance to the quantity \( R'_x(n) \), which is the length of the shortest prefix of \( x \) that contains at least one occurrence of every length-\( n \) factor of \( x \) [2]. There is also a connection (which we shall make use of later) to the concept of a word with grouped factors, which was studied by Cassaigne [9].

In the remainder of this paper, we will give some general properties of the non-repetitive complexity function in comparison to the usual complexity function. We will also give explicit
formulas for the non-repetitive complexity of some of the classical infinite words, namely, the Thue-Morse word (m), the Fibonacci word (f) and the Tribonacci word (t). Finally, we examine the possible range of values that the non-repetitive complexity function can take for squarefree or overlap-free words. We attempt to construct squarefree and overlap-free words with slowly growing non-repetitive complexity functions. This is somewhat similar to the notion of a highly repetitive word [16].

2. Preliminaries

Let \( \Sigma \) denote a finite alphabet and let \( \Sigma^* \) denote the set of finite words over \( \Sigma \). Let \( \{0, 1\} \) be the alphabet in the case of the Thue–Morse and Fibonacci words and let the alphabet be \( \{0, 1, 2\} \) for the Tribonacci word. If \( \theta : \Sigma^* \to \Sigma^* \) is a morphism, then \( \theta^r(u) \) for a non-negative integer \( r \) and a word \( u \) is obtained by applying the morphism \( \theta \) to \( u \) \( r \) times (we define \( \theta^0(u) = u \)). By convention, we denote the string of length 0 by \( \epsilon \). If \( x \) is a word (finite or infinite) we let \( x[i \ldots j] \) denote the factor of \( x \) of length \( j - i + 1 \) that starts at position \( i \) in \( x \). We denote the length of any finite word \( u \) by \( |u| \). For any letter \( a \), we denote the number of occurrences of \( a \) in \( u \) by \( |u|_a \).

If \( \theta : \Sigma^* \to \Sigma^* \) is a morphism, then the adjacency matrix associated with \( \theta \) is the matrix \( M \) with rows and columns indexed by elements of \( \Sigma \) such that the \( ij \) entry of \( M \) equals \( |\theta(i)|_j \).

A square is a non-empty word of the form \( xx \), and a cube is a non-empty word of the form \( xxx \). More generally, for any real number \( \alpha \geq 2 \), an \( \alpha \)-power is a non-empty prefix \( u \) of some infinite word \( xxx \cdot \cdot \cdot \) satisfying \( |u|/|x| \geq \alpha \). An overlap is a word of the form \( axaxa \), where \( a \) is a letter and \( x \) is a word (possibly empty). A word is squarefree (resp. cubefree, \( \alpha \)-power-free, overlap-free) if none of its factors are squares (resp. cubes, \( \alpha \)-powers, overlaps). A palindrome is a word that equals its reversal.

Let \( \mu \) be the Thue–Morse morphism defined by \( \mu(0) = 01, \mu(1) = 10 \). Clearly \( |\mu(u)| = 2|u| \) for any factor \( u \) of \( m \). We define the Thue–Morse word as \( m = \mu^\infty(0) \). If \( u = x_1x_2 \cdot \cdot \cdot x_s \) is a word over \( \{0, 1\} \) for some positive integer \( s \), then we define \( \overline{u} \) by \( \overline{u} = y_1y_2 \cdot \cdot \cdot y_s \) where \( y_i = 1 - x_i \).

Let \( x \) be a finite or infinite word. A factor \( v \) of \( x \) is left special (resp. right special) if there are distinct letters \( a \) and \( b \) such that \( va \) and \( vb \) (resp. \( av \) and \( bv \)) are factors of \( x \). A factor \( v \) of \( x \) is bispecial if it is both left special and right special. An infinite word is Sturmian if it contains exactly \( n + 1 \) factors of length \( n \) for every \( n \geq 0 \). A Sturmian word is standard if each of its prefixes is left special.

Let \( \phi \) be the Fibonacci morphism defined by \( \phi(0) = 01, \phi(1) = 0 \). We define the Fibonacci word as \( f = \phi^\infty(0) \). We define \( f_k = \phi^k(0) \). We define the Fibonacci sequence as \( F_0 = 1, F_1 = 2 \) and \( F_k = F_{k-1} + F_{k-2} \) for \( k \geq 2 \). Note that \( |f_k| = F_k \) and that \( f_k = f_{k-1}f_{k-2} \) (that is, \( f_k \) is the concatenation of \( f_{k-1} \) with \( f_{k-2} \)). Also note that the Fibonacci word is a standard Sturmian word.

Let \( \sigma \) be the Tribonacci morphism defined by \( \sigma(0) = 01, \sigma(1) = 02, \sigma(2) = 0 \). We define the Tribonacci word as \( t = \sigma^\infty(0) \). We define \( t_k = \sigma^k(0) \). We define the Tribonacci sequence as \( T_0 = 1, T_1 = 2, T_2 = 4 \) and \( T_k = T_{k-1} + T_{k-2} + T_{k-3} \) for \( k \geq 3 \). Also, we define \( t_{-1} = 2 \) and \( T_{-1} = 1 \). Note that \( |t_k| = T_k \) and that \( t_k = t_{k-1}t_{k-2}t_{k-3} \). We define \( D_k = t_{k-1}t_{k-2} \cdot \cdot \cdot t_2t_1t_0 \) for \( k \geq 1 \). By convention, we define \( D_0 = \epsilon \).
3. Some general properties of non-repetitive complexity

Recall that the complexity function $c_w(n)$ satisfies $c_w(n) > n$ for any aperiodic word $w$. This is not necessarily true for the nonrepetitive complexity function. Nevertheless, the nonrepetitive complexity must grow at least linearly for any aperiodic word $w$.

**Theorem 1.** Let $w$ be an infinite word and let $\varphi$ be the golden ratio. Then

$$\limsup_{n \to \infty} \frac{nrc_w(n)}{n} < \frac{1}{1 + \varphi^2}$$

if and only if $w$ is ultimately periodic.

**Proof.** One direction is clear. For the other direction, let $\epsilon < 1/(1 + \varphi^2)$ and suppose that there exists $N$ such that $nrc_w(n) < \epsilon n$ for all $n \geq N$. Suppose further that $N$ satisfies $\lfloor (1 + \varphi^2)\epsilon(N + 1) \rfloor < N$. For each $n \geq N$, there exist $i_n$ and $j_n$ satisfying $0 \leq i_n < j_n \leq \epsilon n$ such that $w[i_n \ldots i_n + n - 1] = w[j_n \ldots j_n + n - 1]$. Define $p_n = j_n - i_n$ and note that $w[i_n \ldots i_n + n - 1]$ has period $p_n \leq \epsilon n$. Define discrete intervals $I_n = [i_n + \lfloor \varphi^2 p_n \rfloor, i_n + n]$. For every $i \in I_n$, the prefix $w[0 \ldots i - 1]$ ends with a $\varphi^2$-power. Moreover, since

$$i_{n+1} + \lfloor \varphi^2 p_{n+1} \rfloor \leq \epsilon(n + 1) + \lfloor \varphi^2(n + 1) \rfloor \leq \lfloor (1 + \varphi^2)\epsilon(n + 1) \rfloor < n \leq i_n + n,$$

the intervals $I_n$ and $I_{n+1}$ overlap. Consequently, we have $\bigcup_{n \geq N} I_n = [i_N + \lfloor \varphi^2 p_N \rfloor, \infty]$. Thus, for every $i \geq i_N + \lfloor \varphi^2 p_N \rfloor$, the prefix $w[0 \ldots i - 1]$ ends with a $\varphi^2$-power. Mignosi, Restivo, and Salemi [14, Theorem 2] showed that this implies that $w$ is ultimately periodic, as required. \qed

This result gives an interesting new characterization of ultimate periodicity. Later (Theorems 6 and 12) we shall compute the non-repetitive complexity function for the Thue–Morse word $m$ and the Fibonacci word $f$. These results imply

$$\limsup_{n \to \infty} \frac{nrc_m(n)}{n} = 3$$

and

$$\limsup_{n \to \infty} \frac{nrc_f(n)}{n} = 1.$$ 

One may therefore reasonably wonder if the constant $1/(1 + \varphi^2)$ is optimal in Theorem 1 or if it could perhaps be replaced by 1.

Next we show that there are infinite words whose non-repetitive complexity is maximal. First, recall that for any alphabet of size $q$ and any $n$ there exists a (non-cyclic) $q$-ary de Bruijn sequence of order $n$, that is, a word of length $q^n + n - 1$ that contains every $q$-ary word of length $n$ as a factor. A cyclic $q$-ary de Bruijn sequence of order $n$ is a word $B_n$ of length $q^n$ that contains every $q$-ary word of length $n$ as a circular factor. Here by circular factor we mean a factor of some cyclic shift of $B_n$.

**Proposition 2.**

(a) Over any alphabet of size $q \geq 3$ there exists an infinite word $w$ satisfying

$$nrc_w(n) = q^n$$

for all $n \geq 1$. 

3
(b) Over a binary alphabet there exists an infinite word $w$ satisfying
\[ \text{ncr}_w(2n) = 2^{2n} \]
for all $n \geq 1$.

Proof. This is a consequence of a result of Becher and Heiber [5]. They showed that over alphabets of size at least three any (non-cyclic) de Bruijn sequence of order $n$ can be extended to a de Bruijn sequence of order $n + 1$. Taking the limit of such extensions gives an infinite word with the desired property. Curiously, over a binary alphabet, de Bruijn sequences of order $n$ cannot be extended to order $n + 1$, but can be extended to give de Bruijn sequences of order $n + 2$. □

Next we explore the relationship (if any) between the factor complexity and non-repetitive complexity functions. The next results shows that there are infinite words with maximal factor complexity but only linear non-repetitive complexity.

**Proposition 3.** Let $B_n$ denote a cyclic $q$-ary de Bruijn sequence of order $n$ starting with $n$ 0’s. Then
\[ x = 0^{q^1} B_1 0^{q^2} B_2 0^{q^3} B_3 \cdots \]
is an infinite word with complexity $q^n$ and nonrepetitive complexity $\leq 4n$ for $n \geq 1$.

Proof. Since $B_n$ contains every $q$-ary word of length $n$ as a circular factor, having at least $n - 1$ 0’s follow each $B_n$ ensures that every $q$-ary word of length $n$ shows up in $x$. Thus $x$ has complexity $q^n$ for all positive $n$. The factor of length $n < q^{q^k}$ starting at the first position of the factor $0^{q^k}$ consists of $n$ 0’s. The factor of length $n$ starting at the second 0 of $0^{q^k}$ also consists of $n$ 0’s. It follows that if $n < q^{q^k}$, then $\text{ncr}_x(n)$ must be less or equal to the length of the prefix of $x$ ending just before the second 0 of the $0^{q^k}$ substring. That length is $(q^{q^{k-1}} + q^{q^{k-2}} + \cdots + q^1) + (q^{k-1} + q^{k-2} + \cdots + q^1) + 1$ for $k \geq 2$. It follows that if $q^{q^{k-1}} \leq n < q^{q^k}$, then

\[
\begin{align*}
\text{ncr}_x(n) &\leq \sum_{i=1}^{k-1} q^{q^i} + \sum_{i=1}^{k-1} q^i + 1 \\
&\leq \sum_{i=1}^{q^{k-1}} q^{q^i} + \sum_{i=1}^{k-1} q^i + 1 \\
&\leq \frac{q - q^{q^{k-1}+1}}{1 - q} + \frac{q - q^k}{1 - q} + 1 \\
&\leq \left(\frac{q}{q - 1}\right)(q^{q^{k-1}} + q^{k-1} - 2) + 1 \\
&\leq 2(2n - 2) + 1 \\
&\leq 4n.
\end{align*}
\]

Since $\text{ncr}_x(1) = \text{ncr}_x(2) = \text{ncr}_x(3) = 1$, the proof is now complete. □

The previous result showed that there can be a dramatic difference between the behaviours of the factor complexity function and the non-repetitive complexity function. Next we show
what kind of separation is possible for these two functions when we restrict our attention to pure morphic words. It is well-known that pure morphic words have $O(n^2)$ factor complexity \cite{[3]}. Define the morphism $\phi$ by $\phi(0) = 001$, $\phi(1) = 1$ and let $x = \phi^\infty(0)$. It is known that $x$ has $\Theta(n^2)$ factor complexity (see, for instance, \cite{[6]} Example 4.7.67).

**Lemma 4.** For all $k \geq 0$, the word $x$ has the prefix $zz$, where $|z| = 2^{k+1} - 1$.

**Proof.** Since $x$ begins with 00, it begins with $\phi^k(0)\phi^k(0)$ for all $k \geq 0$. Thus we may take $z = \phi^k(0)$. It remains to show that $|z| = 2^{k+1} - 1$.

Let $M = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$ be the adjacency matrix associated with $\phi$. Then an easy induction shows that $M^k = \begin{bmatrix} 2^k & 0 \\ 2^k - 1 & 1 \end{bmatrix}$ for $k \geq 0$. Now we have

$$|z| = |\phi^k(0)| = |\phi^k(0)|_0 + |\phi^k(0)|_1 = 2^k + 2^k - 1 = 2^{k+1} - 1,$$

as required. $\square$

**Proposition 5.** $\text{nrc}_x(n) < 2n$ for $n \geq 1$.

**Proof.** From Lemma 4, we have that if $n \leq 2^{k+1} - 1$, then $\text{nrc}_x(n) \leq 2^{k+1} - 1$. It follows that if $2^k - 1 < n \leq 2^{k+1} - 1$, then $\text{nrc}_x(n) \leq 2^{k+1} - 1 - 2(2^k) - 1 \leq 2n - 1 < 2n$. $\square$

### 4. Non-repetitive complexity of the Thue–Morse word

We now begin to compute explicitly the non-repetitive complexity functions for some of the classical infinite words, beginning with the Thue–Morse word. Recall that the Thue–Morse word is the word $m = \mu^\infty(0)$, where $\mu(0) = 01$ and $\mu(1) = 10$.

**Theorem 6.** If $2^{k-1} < n \leq 2^k$ for $k \geq 1$, then $\text{nrc}_m(n) = 3(2^{k-1})$.

The following is well-known.

**Lemma 7.** If $A$ is a prefix of $m$ of length $2^k$ for $k \geq 0$, then $\mu(A) = A\overline{A}$.

**Lemma 8.** For any positive integer $k$, $m[0 \ldots 2^k - 1] = m[2^k + 2^{k-1} \ldots 2^{k+1} + 2^{k-1} - 1]$.

**Proof.** Let $A = m[0 \ldots 2^{k-1} - 1]$ for some positive integer $k$. Clearly $\mu^r(0)$ is a prefix of $m$ for every nonnegative integer $r$. Then $\mu^3(A) = \mu^2(A\overline{A}) = \mu(A\overline{A}A\overline{A}) = A\overline{A}A\overline{A}A\overline{A}A\overline{A}$ is a prefix of $m$. Since $|A| = 2^{k-1}$, then $A\overline{A} = m[0 \ldots 2^{k-1} - 1]$ and $A\overline{A} = m[2^k + 2^{k-1} \ldots 2^{k+1} + 2^{k-1} - 1]$. $\square$

The following upper bound for $\text{nrc}_m(n)$ follows immediately from Lemma 8.

**Proposition 9.** If $n \leq 2^k$ for $k \geq 1$, then $\text{nrc}_m(n) \leq 3(2^{k-1})$.

We make use of the next result to obtain a matching lower bound.

**Lemma 10.** \cite{[3]} Example 10.10.3] For any integer $k \geq 2$, each factor of $m$ of length $2^{k-1} + 1$ occurs in the prefix of $m$ of length $2^{k+1}$. Furthermore, each one of these factors occurs exactly once in this prefix.

**Proposition 11.** If $2^{k-1} < n$ for $k \geq 2$, then $\text{nrc}_m(n) \geq 3(2^{k-1})$. 


Proof. By Lemma 10, the first $2^{k+1} - (2^{k-1} + 1) + 1 = 3(2^{k-1})$ factors of length $2^{k-1} + 1$ appearing in $m$ are all distinct. Consequently the first $3(2^{k-1})$ length-$n$ factors appearing in $m$ must also be distinct. □

Using Propositions 9 and 11 and that $\text{nrct}_m(2) = 3$ (obtained through observation), we get Theorem 6 and thus the proof is complete. Though the theorem is not defined for $m$, we know from Lemma 16 that the Fibonacci word has the prefix $n$ please note that $nrc$.

5. Non-repetitive complexity of the Fibonacci word

Recall that the Fibonacci word is the word $f = \phi^\omega(0)$, where $\phi(0) = 01$ and $\phi(1) = 0$.

Theorem 12. If $F_{k-1} \leq n + 1 < F_k$ for $k \geq 2$, then $\text{nrct}_f(n) = F_{k-1}$.

We first need some preliminary results. Recall that $f_k = \phi^k(0)$. The next lemma is well-known.

Lemma 13. For $k \geq 2$, the words $f_k = f_{k-1}f_{k-2}$ and $f_{k-2}f_{k-1}$ differ only by their last two letters.

Lemma 14. For any positive integer $k \geq 2$, $f[0 \ldots F_k - 3] = f[F_{k-1} \ldots F_{k+1} - 3]$.

Proof. We know $f_{k+1} = f_kf_{k-1} = f_{k-1}f_{k-2}f_{k-1}$ is a prefix of $f$. By Lemma 13 $f[0 \ldots F_k - 1]$ and $f[F_{k-1} \ldots F_{k+1} - 1]$ agree up to but not including the last two positions. The result follows. □

This implies the following result.

Proposition 15. If $n + 1 < F_k$ for $k \geq 2$, then $\text{nrct}_f(n) \leq F_{k-1}$.

Furthermore, the Fibonacci word is a standard Sturmian word, so for $k \geq 1$, $f_k = u_krs$, where $rs = 01$ if $k$ is odd or $rs = 10$ if $k$ is even. The $u_k$’s are known as central words and it is known that these central words are palindromes and are bispecial (see [11]).

A *semitrivial word* [8] is a word in which the longest repeated prefix, longest repeated suffix, longest left special factor and longest right special factor are all the same word. Furthermore, this prefix/suffix/bispecial factor is a central word.

Lemma 16. [10] Proposition 16] The semitriangular prefixes of a standard Sturmian word are precisely the words of the form $u_krsu_k$ for $k \geq 1$.

The property described in the next lemma is the property of having grouped factors, which was mentioned in the introduction.

Lemma 17. [9] Corollary 1] A sequence is Sturmian if and only if, for $n \geq 0$, it has a factor of length $2n$ containing all factors of length $n$ exactly once. Furthermore, if $n \geq 1$, then there are exactly two such factors of length $2n$, namely $w01v$ and $w10v$, where $w$ is the unique right special factor of length $n - 1$ and $v$ is the unique left special factor of length $n - 1$.

Proposition 18. If $F_{k-1} \leq n + 1$ for $k \geq 2$, then $\text{nrct}_f(n) \geq F_{k-1}$.

Proof. It suffices to show that for $n = F_{k-1} - 1$, the first $F_{k-1}$ factors of $f$ of length $n$ are all distinct. We know from Lemma 16 that the Fibonacci word has the prefix $u_{k-1}rsu_{k-1}$
where \( rs = 01 \) or \( rs = 10 \). Since these prefixes are of the same construction as the factors detailed in Lemma \[17\] (\( u_{k-1} \) is the left and right special factor of length \( n - 1 \)), and since

\[
|u_{k-1}rsw_{u_{k-1}}| = 2|u_{k-1}r| = 2(F_{k-1} - 1) = 2n,
\]

it follows that this semicentral prefix contains all factors of length \( n \) exactly once. Thus for all \( n \geq F_{k-1} - 1 \), all factors of length \( n \) are distinct over the first \( 2(F_{k-1} - 1) \) positions and so the result follows.

Using Propositions \[15\] and \[18\] we get Theorem \[12\] and thus the proof is complete.

6. NON-REPETITIVE COMPLEXITY OF THE TRIBONACCI WORD

Recall that the Tribonacci word is the word \( t = \sigma^\omega(0) \), where \( \sigma(0) = 01 \), \( \sigma(1) = 02 \), and \( \sigma(2) = 0 \).

**Theorem 19.** If \( \frac{T_{k+1} + T_{k-1}}{2} - 3 < n \leq \frac{T_{k+1} + T_{k-1} - 3}{2} \) for \( k \geq 1 \), then \( \text{nrc}_t(n) = T_k \).

We first need to recall some known properties of the Tribonacci word. Recall that \( t_k = \sigma^k(0) \) and that \( D_k = t_{k-1}t_{k-2} \cdots t_2t_1t_0 \) for \( k \geq 1 \).

**Lemma 20.** \[18\] Theorem 2.5] For \( k \geq 2 \), the longest common prefix of \( t_{k-3}t_{k-1}t_{k-2} \) and \( t_k \) is \( D_{k-2} \).

**Lemma 21.** \[18\] Proposition 2.9] For \( k \geq 1 \), \( |D_k| = \frac{T_{k+1} + T_{k-1} - 3}{2} \).

**Lemma 22.** For any positive integer \( k \geq 2 \),

\[
t \left[ 0, \frac{T_{k+1} + T_{k-1} - 3}{2} - 1 \right] = t \left[ T_k \ldots T_k + \frac{T_{k+1} + T_{k-1} - 3}{2} - 1 \right].
\]

**Proof.** We know \( t_{k+2} = t_{k+1}t_k \) is a prefix of \( t \) for \( k \geq 2 \). By Lemma \[20\], we know that \( t_{k-3}t_{k-1}t_{k-2} \) agrees with \( t_k \) up to the first \( |D_{k-2}| \) symbols. It follows that \( t_{k-1}t_{k-2}t_{k-3}t_{k-1}t_{k-2} \) agrees with \( t_{k-1}t_{k-2}t_k \) up to the first \( |D_k| \) symbols. Since \( t_{k-1}t_{k-2}t_k = t[T_k \ldots T_k + T_{k+1} - 1] \), the result follows from Lemma \[21\] \( \square \)

We therefore have the following.

**Proposition 23.** If \( n \leq \frac{T_{k+1} + T_{k-1} - 3}{2} \) for \( k \geq 2 \), then \( \text{nrc}_t(n) \leq T_k \).

Before proving the lower bound for \( \text{nrc}_t(n) \), we need some additional properties of the Tribonacci word.

**Lemma 24.** \[17\] Proof of Proposition 3.3] The bispecial factors of \( t \) are precisely the palindromic prefixes of \( t \). Furthermore, the lengths of these (nonempty) prefixes are \( \frac{T_{k+2} + T_{k-3}}{2} \) for \( k \geq 0 \).

**Lemma 25.** \[18\] Lemma 2.3] If \( w \) is a palindrome, then \( \sigma(w)0 \) is a palindrome.

**Lemma 26.** If \( w \) is a palindromic prefix of \( t \) of length \( |D_k| \) for \( k \geq 1 \), then \( \sigma(w)0 \) is a palindromic prefix of \( t \) of length \( |D_{k+1}| \).

7
Proof. We know from Lemma \[24\] that all palindromic prefixes of \( t \) are of length \(|D_k|\) for \( k \geq 1 \). If \( w \) is a palindromic prefix of \( t \) of length \(|D_k|\), then clearly \( \sigma(w) \) is a prefix of \( t \). Furthermore, since \( w \) starts with a 0 and is a palindrome, it ends with a 0. So \( \sigma(w) \) ends with a 1. Since the string 11 is not in \( t \), then \( \sigma(w) \) must be followed by a 0. Thus, \( \sigma(w)0 \) is a prefix of \( t \) and we know from Lemma \[25\] that it is a palindrome. Applying the morphism \( \sigma \) to \( w \) will at most double the length. Thus

\[
|\sigma(w)0| \leq 2|w| + 1
\]

\[
= 2 \left( \frac{T_{k+1} + T_{k-1} - 3}{2} \right) + 1
\]

\[
= \frac{2T_{k+1} + 2T_{k-1} - 4}{2}
\]

\[
< \frac{T_{k+2} + T_{k+1} + T_k + T_{k+1} + T_{k-2} - 3}{2}
\]

\[
= \frac{T_{k+3} + T_{k+1} - 3}{2}
\]

\[
= |D_{k+2}|.
\]

So the only option for the length of \( \sigma(w)0 \) is \(|D_{k+1}|\).

\[\square\]

Lemma 27. If \( w \) is a (nonempty) palindromic prefix of \( t \), then the first symbols that follow each of the first two occurrences of \( w \) in \( t \) are different.

Proof. By induction on \( k \) where \(|w| = |D_k|\). Since \( t = 0102 \cdots \), the result holds for \( k = 1 \). Assume that the first symbol that follows each of the first two occurrences of \( w \) are different where \(|w| = |D_k|\).

Case 1: One \( w \) is followed by a 0 while the other \( w \) is followed by a 1.

Then \( \sigma(w0) = \sigma(w)01 \) while \( \sigma(w1) = \sigma(w)02 \).

Case 2: One \( w \) is followed by a 2 while the other \( w \) is followed by a 0.

Then \( \sigma(w2x) = \sigma(w)00y \) for some \( x = \{0, 1, 2\} \) and \( y = \{\epsilon, 1, 2\} \), while \( \sigma(w0) = \sigma(w)01 \).

Case 3: One \( w \) is followed by a 2 while the other \( w \) is followed by a 1.

Then \( \sigma(w2x) = \sigma(w)00y \) for some \( x = \{0, 1, 2\} \) and \( y = \{\epsilon, 1, 2\} \), while \( \sigma(w1) = \sigma(w)02 \).

In all three cases, both occurrences of \( \sigma(w)0 \) are followed by different symbols. Since \(|\sigma(w)0| = |D_{k+1}|\) by Lemma \[26\] this implies that the statement holds for \( k + 1 \) and thus the statement holds for all \( k \) by induction. \[\square\]

Lemma 28. \[18\] Section 1, Point 3] There is a unique left special factor and a unique right special factor of each length in \( t \).

Lemma 29. Let \( v \) denote the prefix of length \(|D_{k-1}|\) of \( t \) for \( k \geq 2 \). All the factors of length \(|D_{k-1}|\) that start between the beginning of the first occurrence of \( v \) and the beginning of the third occurrence of \( v \) are distinct (except for \( v \)).

Proof. Firstly, since \( t \) is recurrent, we know that there are three occurrences of \( v \) in \( t \). For the sake of contradiction, assume the factor \( u(\neq v) \) of length \(|D_{k-1}|\) has two occurrences in \( t \) before we reach the first symbol of the third occurrence of \( v \). For simplicity, let \( v_j \) denote the \( j \)th occurrence of \( v \) and \( u_i \) the \( i \)th occurrence of \( u \). If the starting symbol of \( u_i \) is between the starting symbol of \( v_j \) and \( v_{j+1} \), then we will denote that by \( v_j < u_i < v_{j+1} \).

Case 1: \( v_1 < u_1 < v_2 \).
If \( u_1 \) and \( u_2 \) are preceded by different symbols, then \( u \) is a left special factor. This is a contradiction since \( u \neq v \) and \( v \) is the unique left special factor of length \(|D_{k-1}|\) in \( t \). Thus, assume they are preceded by the same symbol. Then we obtain another factor (formed by the first \(|D_{k-1}| - 1 \) symbols of \( u \) and the symbol preceding \( u_1 \)), which we will call \( a \), of length \(|D_{k-1}| \) such that \( v_1 < a_1 < a_2 < v_2 \). Once again, if \( a_1 \) and \( a_2 \) are preceded by different symbols then we obtain a contradiction. By repeating this argument we eventually find that \( v_1 < v_j < u_2 \) for some \( j \), which contradicts our original assumption.

Case 2: \( v_2 < u_1 < u_2 < v_3 \).

Similar to Case 1.

Case 3: \( v_1 < u_1 < v_2 < u_2 < v_3 \).

We apply the same argument as in Case 1. We either obtain the same contradiction described in that case, or we find that the factor starting with the first symbol of \( v_1 \) and ending with the last symbol of \( u_1 \) is identical to the factor starting with the first symbol of \( v_2 \) and ending with the last symbol of \( u_2 \). This is a contradiction since the symbols following \( v_1 \) and \( v_2 \) are different by Lemma 27.

In all three cases we obtain a contradiction. Thus all the factors (except \( v \)) are distinct.

\( \square \)

**Lemma 30.** Let \( k \geq 2 \). All factors of \( t \) of length \(|D_{k-1}| + 1 \) that begin prior to the third occurrence of the prefix of \( t \) of length \(|D_{k-1}| \) are distinct.

**Proof.** It is a direct result of Lemmas 27 and 29.

\( \square \)

**Lemma 31.** [12, Section 6.3.5] If a square \( xx \) is a factor of \( t \), then \(|x| \in \{T_k, T_k + T_{k-1}\} \) for some \( k \geq 1 \).

**Lemma 32.** If a word \( v \) of length \(|D_{k-1}| \) for \( k \geq 5 \) overlaps itself in \( t \), then the shortest period of \( v \) is at least \( T_{k-2} \).

**Proof.** The largest Tribonacci number or sum of consecutive Tribonacci numbers less than \( T_{k-2} \) is \( T_{k-3} + T_{k-4} \). Let \( v = xaax \) be a factor of \( t \) of length \(|D_{k-1}| \), where \( x \) is a nonempty factor of \( t \) and \( a \) is a possibly empty factor of \( t \). Suppose that \( t \) contains the overlap \( xaxax \). Note that \(|xa|\) is a period of \( v \). Also, \(|xa| < |D_{k-1}| \) and \( 2|xa| \geq |D_{k-1}| \). However, \(|D_{k-1}| = T_{k-2} + T_{k-3} + T_{k-4} + \cdots + T_0 = 2T_{k-3} + 2T_{k-4} + 2T_{k-5} + T_{k-6} + \cdots + T_0 > 2(T_{k-3} + T_{k-4}) \) for \( k \geq 5 \). So every period of \( v \) must be larger than \( T_{k-3} + T_{k-4} \). Thus, from Lemma 31, the shortest period of \( v \) is at least \( T_{k-2} \).

\( \square \)

**Lemma 33.** If \( v \) is a prefix of \( t \) of length \(|D_{k-1}| \) for \( k \geq 2 \), then the second occurrence of \( v \) occurs at position \( T_{k-1} \) and the third occurs at position \( T_k \).

**Proof.** Since \( t = 01020201001201010201 \cdots \), it can be observed that the statement holds for \( k = 2, 3, 4 \). Thus, assume for the rest of this proof that \( k \geq 5 \). By Lemma 22, we already know that the prefix \( v \) occurs at position \( T_{k-1} \) and position \( T_k \). If there were an occurrence of \( v \) that started somewhere between the beginning of \( t \) and position \( T_{k-1} \) of \( t \), then by Lemma 32 the start of this occurrence of \( v \) must be at distance at least \( T_{k-2} \) from the beginning of \( t \) and at distance at least \( T_{k-2} \) from position \( T_{k-1} \). This implies that \( 2T_{k-2} \leq T_{k-1} \) but to the contrary we have \( 2T_{k-2} = T_{k-2} + T_{k-3} + T_{k-4} = T_{k-1} + T_{k-5} > T_{k-1} \). Furthermore, \( T_{k-1} + 2T_{k-2} > T_{k-1} + T_{k-2} + T_{k-3} = T_k \). It follows that no occurrence of \( v \) can start between the beginning of \( t \) and position \( T_{k-1} \) nor can it start anywhere between positions \( T_{k-1} \) and \( T_k \).

\( \square \)
Proposition 34. If $\frac{T_{k+T_{k-2}-3}}{2} < n$ for $k \geq 2$, then $\text{nrc}_k(n) \geq T_n$.

Proof. The result follows from Lemma 30 and Proposition 34. □

Using Propositions 23 and 34 and that $\text{nrc}_1(1) = 2$ (making the theorem hold for $k = 1$), we get Theorem 19 and thus the proof is complete.

7. Non-repetitive complexity of squarefree and overlap-free words

In this section we examine the possible behaviour of the non-repetitive complexity function for words avoiding squares, overlaps, or cubes. In particular, we attempt to construct words that avoid the desired type of repetition but have non-repetitive complexity as low as possible.

Proposition 35. There is no infinite squarefree word $x$ that has $\text{nrc}_x(n) < 2n$ for all $n$.

Proof. Since $\text{nrc}_x(n) < 2n$, $\text{nrc}_x(1) = 1$ and therefore $x = aa \cdots$, a contradiction. □

Proposition 36. There is no infinite cube-free word $x$ that has $\text{nrc}_x(n) < \frac{3}{2}n$ for all $n$.

Proof. Since $\text{nrc}_x(n) < \frac{3}{2}n$, $\text{nrc}_x(1) = 1$ and $\text{nrc}_x(2) \leq 2$. It follows that $x = aaa \cdots$ which is a contradiction. □

Proposition 37. There is no infinite overlap-free word $x$ that has $\text{nrc}_x(n) < 2n$ for all $n$.

Proof. We attempt to construct such a word over some alphabet $\Sigma$ by a standard back-tracking algorithm (by hand or by computer). We find that such a word must begin with $aabaabbabaab$ for some distinct letters $a$ and $b$ (the binary alphabet is forced). If this word is extended as $aabaabbabaab.ab$, we see that it is not possible to have $\text{nrc}_x(6) < 12$. If on the other hand we have the extension $aabaabbabaab.ba$, we obtain the overlap $abaabbaabb$ which is a contradiction. □

Consider the infinite alphabet $\Sigma = \{0, 1, 2, \ldots\}$. We define the sequence of Zimin words, $Z_0, Z_1, Z_2, \ldots$, as follows: $Z_0 = \epsilon$ and $Z_{n+1} = Z_n n Z_n$ for $n \geq 0$. Let

$$x = 0102010301020104 \cdots,$$

also known as the ruler sequence, be the limit of the $Z_n$.

Theorem 38. The infinite word $x$ is squarefree and satisfies $n < nsc_x(n) \leq 2n$ for all $n \geq 1$.

Proof. For the squarefreeness of $x$ see [13]. By the definition of $x$, if $n \leq 2^k - 1$, then $\text{nrc}_x(n) \leq 2^k$ for $k \geq 1$. It follows that if $2^k - 1 \leq n < 2^k$, then $\text{nrc}_x(n) \leq 2^k \leq 2(2^k - 1) \leq 2n$ for all $n$. Also, since $x$ is square-free, clearly $n < \text{nrc}_x(n)$ for all $n$. □

So we can obtain an infinite squarefree word over an infinite alphabet that has $\text{nrc}_x(n) \leq 2n$ for all $n$. Furthermore, for this word there are infinitely many values of $n$ such that $\text{nrc}_x(n) = 2n$.

Using an infinite alphabet may seem like “cheating”, so next we examine what can be done over a finite alphabet. We will make use of a morphism $\theta : \mathbb{N} \rightarrow \{a, b, c, d, e\}$, which maps squarefree words over an infinite alphabet to squarefree words over an alphabet of size 5. First, let

$$w = abacbabcbacabc \cdots$$

be the well-known squarefree word obtained by iterating the morphism

$$a \rightarrow abc, \quad b \rightarrow ac, \quad c \rightarrow b.$$
For $i \geq 0$, let $W_i$ be the prefix of $w$ of length $i$. We define $\theta(i) = dW_i eW_i$ for all $i \geq 0$. The map $\theta$ is squarefree [4, Corollary 1.4]; that is, if $u$ is squarefree, then $\theta(u)$ is squarefree.

**Theorem 39.** Let $x$ be the ruler sequence defined previously. Then $y = \theta(x)$ is a square-free word with $\text{nrc}_y(n) < 3n$ for all $n$ except $n = 2$.

**Proof.** It is relatively easy to see that the prefix $A$ of $x$ of length $2^k - 1$ will have $2^k - 1$ 0's, $2^{k-1}$ 1's, and so on, down to only one occurrence of $k - 1$. Furthermore, as a result of how we defined the $W_i$'s, we have $|\theta(i)| = 2^{i+1}$. Thus, we have

$$|\theta(A)| = 2k + 2(2(k - 1)) + 4(2(k - 2)) + \cdots + 2^{k-1}(2(1))$$

$$= \sum_{i=0}^{k-1} 2^{i+1}(k - i)$$

$$= 2^{k+2} - 2k - 4.$$  

Furthermore, if $B$ is the prefix of $x$ of length $2^k$, then

$$|\theta(B)| = |\theta(A)| + 2(k + 1) = 2^{k+2} - 2k - 4 + 2k + 2 = 2^{k+2} - 2,$$

since $|\theta(k)| = 2(k + 1)$. As a result of the fact that $\text{nrc}_x(n) \leq 2^k$ for $2^k - 1 < n < 2^k$, if $2^{k+1} - 2(k - 1) - 4 < n \leq 2^{k+2} - 2k - 4$, then $\text{nrc}_y(n) \leq 2^{k+2} - 2$. The expression

$$\frac{2^{k+2} - 2}{2^{k+1} - 2(k - 1) - 3}$$

is a decreasing function of $k$ and is less than 3 for $k \geq 4$. Along with the fact that $\text{nrc}_y(n) < 3n$ for $n \leq 22$ (other than $n = 2$), which can be obtained through computation, we have $\text{nrc}_y(n) < 3n$ for all $n$ except $n = 2$. \qed

It should be noted that

$$\lim_{k \to \infty} \frac{2^{k+2} - 2}{2^{k+2} - 2k - 4} = 1 \quad \text{and} \quad \lim_{k \to \infty} \frac{2^{k+2} - 2}{2^{k+1} - 2(k - 1) - 3} = 2,$$

meaning that

$$\liminf_{n \to \infty} \frac{\text{nrc}_y(n)}{n} = 1 \quad \text{and} \quad \limsup_{n \to \infty} \frac{\text{nrc}_y(n)}{n} \leq 2.$$

To obtain a result over a 3-letter alphabet we will need a morphism $\sigma$ (found by Brandenburg [7, Theorem 4]), which maps squarefree words on $\{a, b, c, d, e\}$ to squarefree words on $\{a, b, c\}$. We define it by

$$\sigma(a) = abacabcabcbabc$$

$$\sigma(b) = abacabcacabcabcbac$$

$$\sigma(c) = abacabcacbcacbc$$

$$\sigma(d) = abacabcabacbacb$$

$$\sigma(e) = abacabcacbcabc.$$

**Theorem 40.** The word $z = \sigma(y)$ is square-free and has $\text{nrc}_z(n) < 3n$ for all $n > 36.$

11
Proof. Since $|\sigma(m)| = 18$ for $m \in \{a, b, c, d, e\}$ and $\text{nrc}_y(n) \leq 2^{k+2} - 2$ if
\[2^{k+1} - 2(k - 1) - 4 < n \leq 2^{k+2} - 2k - 4,\]
then it follows that if
\[18(2^{k+1} - 2(k - 1) - 4) < n \leq 18(2^{k+2} - 2k - 4),\]
then $\text{nrc}_y(n) \leq 18(2^{k+2} - 2)$. The expression
\[
\frac{18(2^{k+2} - 2)}{18(2^{k+1} - 2(k - 1) - 4) + 1}
\]
is a decreasing function of $k$ and is less than 3 for $k \geq 4$. Along with the fact that $\text{nrc}_y(n) \leq 3n$ for $36 < n \leq (18)(22) = 396$, which can be obtained through computation, $\text{nrc}_y(n) < 3n$ for all $n > 36$.

Furthermore, similar to the word $y$, we have
\[
\lim_{k \to \infty} \frac{18(2^{k+2} - 2)}{18(2^{k+2} - 2k - 4)} = 1 \quad \text{and} \quad \lim_{k \to \infty} \frac{18(2^{k+2} - 2)}{18(2^{k+1} - 2(k - 1) - 4) + 1} = 2,
\]
meaning that
\[
\liminf_{n \to \infty} \frac{\text{nrc}_y(n)}{n} = 1 \quad \text{and} \quad \limsup_{n \to \infty} \frac{\text{nrc}_y(n)}{n} \leq 2.
\]

Since the Thue–Morse word is overlap-free, Theorem 6 shows that it is an example of an overlap-free word with $\text{nrc}_m(n) < 3n$ for all $n \geq 1$.

8. Open questions

Question 1. Is the constant $1/(1 + \varphi^2)$ in Theorem 1 best possible? Can it be replaced by 1?

Question 2. Is the word $x$ of Theorem 38 the only (up to permutation of the infinite alphabet) infinite squarefree word such that $\text{nrc}_x(n) \leq 2n$ for all $n$?

Question 3. Are the examples given in Section 7, along with the Thue–Morse word, optimal for squarefree words or overlap-free words: i.e., are there squarefree words or overlap-free words whose non-repetitive complexity functions grow even slower than the examples given here?

References

[1] J.-P. Allouche, M. Baake, J. Cassaigne, D. Damanik, “Palindrome complexity”, Theoret. Comput. Sci. 292 (2003) 9–31.
[2] J.-P. Allouche, M. Bousquet-Mélou, “On the conjectures of Rauzy and Shallit for infinite words”, Comment. Math. Univ. Carolin. 36 (1995) 705–711.
[3] J.-P. Allouche, J. Shallit, Automatic Sequences: Theory, applications, generalizations. , Cambridge, 2003.
[4] D. Bean, A. Ehrenfeucht, G. McNulty, “Avoidable patterns in strings of symbols”, Pacific J. Math. 85 (1979) 261–294.
[5] V. Becher, P. Heiber, “On extending de Bruijn sequences”, Inform. Process. Lett. 111 (2011) 930–932.
[6] V. Berthé., M. Rigo, eds., Combinatorics, Automata and Number Theory, Cambridge, 2010.
[7] F. Brandenburg, “Uniformly growing $k$-th power-free homomorphisms”, Theoret. Comput. Sci. 23 (1983) 69–82.
[8] M. Bucci, A. De Luca, G. Fici, “Enumeration and structure of trapezoidal words”, Theoret. Comput. Sci. 468 (2013) 12–22.
[9] J. Cassaigne, “Sequences with grouped factors”, in DLT 1997, pp. 211-222, 1997.
[10] A. De Luca, G. Fici, “Open and closed prefixes of Sturmian words”, in WORDS 2013, LNCS 8079, pp. 132–142, 2013.
[11] S. Fischler, “Palindromic prefixes and episturmian words”, J. Combin. Theory Ser. A 113 (2006) 1281–1304.
[12] A. Glen, On Sturmian and Episturmian Words, and Related Topics, Ph.D. thesis, University of Adelaide, 2006.
[13] M. Guay-Paquet, J. Shallit, “Avoiding squares and overlaps over the natural numbers”, Discrete Math. 309 (2009) 6245–6254.
[14] F. Mignosi, A. Restivo, S. Salemi, “Periodicity and the golden ratio”, Theoret. Comput. Sci. 204 (1998) 153–167.
[15] T. K. S. Moothathu, “Eulerian entropy and non-repetitive subword complexity”, Theoret. Comput. Sci. 420 (2012) 80–88.
[16] N. Rampersad, E. Vaslet, “On highly repetitive and power free words”, J. Integer Seq. 16 (2013), Article 13.2.7, 17 pages.
[17] G. Richomme, K. Saari, L. Q. Zamboni, “Balance and abelian complexity of the Tribonacci word”, Adv. in Appl. Math. 45 (2010) 212–231.
[18] B. Tan, Z.-Y. Wen, “Some properties of the Tribonacci sequence”, European J. Combinatorics 28 (2007) 1703–1719.

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF WINNIPEG, 515 PORTAGE AVENUE, WINNIPEG, MANITOBA R3B 2E9 (CANADA)

E-mail address: n.rampersad@uwinnipeg.ca, jnich998@hotmail.com