Journal of Integer Sequences, Vol. 23 (2020),
Article 20.10.8

Two Infinite Words with Cubic Subword Complexity

Luke Schaeffer
Institute for Quantum Computing
University of Waterloo
Waterloo, ON N2L 3G1
Canada
lrschaeffer@gmail.com

Kaiyu Wu
School of Computer Science
University of Waterloo
Waterloo, ON N2L 3G1
Canada
k29wu@uwaterloo.ca

Abstract

We consider two natural infinite words whose subword complexity is cubic, and determine their exact subword complexity. As a consequence, it follows that neither word is morphic.

1 Introduction

In this paper we are concerned with words over a finite alphabet Σ. We say a word y is a *subword* of a word w if there exist (possibly empty) words x, z such that $w = xyz$. For example, `bank` is a subword of `embankment`. The *subword complexity* (also called *factor complexity* or just *complexity*) of an infinite word $\mathbf{w}$ is the function $\rho_{\mathbf{w}}(n)$ that maps n to

the number of distinct subwords of length n in $\mathbf{w}$. Subword complexity is a natural measure of the complexity of a word, and has been extensively studied (see, e.g., [3, 7, 4, 5, 6, 2]).

We say that a morphism $h : \Sigma^* \to \Sigma^*$ is *prolongable* on a letter a if $h(a) = ax$ for some word x such that $h^i(x) \neq \epsilon$ for all $i \geq 0$. In this case, it is meaningful to define $h^\omega(a) := a\,x\,h(x)\,h^2(x)\cdots$, which is an infinite word that is a fixed point of the extension of h to infinite sequences. A word of the form $h^\omega(a)$ is called *pure morphic*.

A *coding* is a particular type of morphism that maps every letter to a word of length 1. An infinite word $\mathbf{w}$ is said to be *morphic* if it can be expressed as the image, under a coding, of a morphic word. The class of morphic words has been widely studied (see, e.g., [1, Chap. 7]).

Pansiot [8, 9] classified the subword complexity of morphic words. He showed that pure morphic words have subword complexity $O(n^2)$, and thus all morphic words have subword complexity $O(n^2)$. A random infinite word will almost surely have exponential subword complexity, and hence is not morphic. However, there are not that many explicit examples of non-morphic words that are easy to write down.

Recently Tim Smith [10] introduced a class of infinite words which he called *zigzag words*. In this note we determine the exact subword complexities of two natural zigzag words and show they are cubic. These, then, provide additional natural examples of non-morphic words.

2 Definitions of words

Let

$$\mathbf{w}_1 = \prod_{i\geq 1}\prod_{j=1}^{i} a^{i-j+1}b^j = (ab)(aab\cdot abb)(aaab\cdot aabb\cdot abbb)\cdots$$

and

$$\mathbf{w}_2 = \prod_{i\geq 1}\prod_{j=1}^{i} a^{j}b^{i-j+1} = (ab)(abb\cdot aab)(abbb\cdot aabb\cdot aaab)\cdots .$$

We will show that the exact subword complexities of $\mathbf{w}_1$ and $\mathbf{w}_2$ are as follows:

$$\rho_{\mathbf{w}_1}(n) = \frac{n^3}{6} + \frac{n^2}{2} - \frac{5n}{3} + 3 \quad \text{for } n \geq 4;$$

$$\rho_{\mathbf{w}_2}(n) = \frac{n^3}{6} - \frac{2n}{3} + \frac{19 + (-1)^n}{4} \quad \text{for } n \geq 4.$$

Call each factor $\prod_{j=1}^{i} a^{i-j+1}b^j$ of $\mathbf{w}_1$ and $\prod_{j=1}^{i} a^j b^{i-j+1}$ of $\mathbf{w}_2$ as a *minute* (or more specifically the $(i+1)$-st minute), and call each natural division within a minute a *second*. Note that seconds are unique; every word of the form a^ib^j occurs infinitely many times in $\mathbf{w}_1$ and $\mathbf{w}_2$, but only the one in the $(i+j)$-th minute will be considered a second. Equivalently, a^ib^j is a second if it is preceded by b and followed by a. Finally, every occurrence of the subword ba marks the boundaries between two seconds, since no second contains ba.

3 Subword complexity of $\mathbf{w}_1$

Using the above observation, whenever we can guarantee that a subword s of $\mathbf{w}_1$ contains a second, we uniquely determine what s must be. This is equivalent to s containing two occurrences of ba, one at the start of the second, and one at the end. Thus if s contains a factor of the form $a^ib^ja^kb^la$ with $j, k, l \geq 1$, we may identify a^kb^l as a second occurring in the $(k+l)$th minute. Since we uniquely determine what s must be, this gives us conditions on the values of i and j. In the case that we cannot guarantee that s contains a second (s may appear in $\mathbf{w}_1$ many times; some of these occurrences will contain a second, but others will not), s must be of the form $a^*b^*a^*b^*$.

Theorem 1. *The subword complexity of* $\mathbf{w}_1$ *is*

$$\rho_{\mathbf{w}_1}(n) = \begin{cases} n^3/6 + n^2/2 - 5n/3 + 3, & \text{for } n \geq 4; \\ 2^n, & \text{otherwise.} \end{cases}$$

Proof. First consider the case when we cannot guarantee that a subword s of w contains a second. Then s must be of the form $a^*b^*a^*b^*$. Note that for all $p, q \geq 1$, the word $a^{p+1}b^qa^pb^{q+1}$ is a subword in the $(p+q+1)$st minute. Thus by taking p and q sufficiently large, we can get any subword of the form $a^*b^*a^*$ or $b^*a^*b^*$. Hence all words in

$$A := a^*b^*a^* \cup b^*a^*b^*$$

are subwords of w. The remaining words in $a^*b^*a^*b^* \setminus A = a^+b^+a^+b^+$ are of the form $a^ib^ja^kb^l$ with $i, j, k, l \geq 1$. Since ba marks the boundary between two seconds and no second has a factor ba, a^ib^j must be the suffix of one second and a^kb^l the prefix of the next. There are two possibilities:

- If the seconds are in the same minute, then $a^ib^ja^kb^l$ is a subword of $a^{p+1}b^qa^pb^{q+1}$ for some $p, q \geq 1$. Thus $i \leq k+1$ and $l \leq j+1$.
- If the seconds are in different minutes, then $a^ib^ja^kb^l$ is a subword of $ab^pa^{p+1}b$. Thus, $i = l = 1, j+1 = k$. But note that in this case, we have $i \leq k+1, l \leq j+1$, so we have already counted these subwords in the first case.

Thus there is a subword $a^ib^ja^kb^l$ in $\mathbf{w}_1$ if and only if the tuple (i, j, k, l) is in

$$B = \{(i, j, k, l) \,:\, i, j, k, l \geq 1, i \leq k+1, l \leq j+1\}.$$

Now consider the case when we can guarantee that s contains a second. By looking at the first second s contains, we conclude that s must be prefixed by a word of the form $a^ib^ja^kb^la$ with $j, k, l \geq 1$ and $i \geq 0$. The second contained is a^kb^l and a^ib^j is the suffix of the second preceding it. There are two possibilities:

- The second $a^k b^l$ is the first second in the minute, so $l = 1$. The preceding second is ab^{k-1}, so either $i = 1$ and $j = k - 1$ or $i = 0$ and $j \le k - 1$. Thus we obtain a unique subword for each tuple in

$$C := \{(0, j, k, 1) : k \ge 2, 1 \le j \le k - 1\} \cup \{(1, k - 1, k, 1) : k \ge 2\}.$$

- The second $a^k b^l$ is not the first second in the minute, so $l > 2$. The preceding second is $a^{k+1} b^{l-1}$, so either $i = 0$ and $j \le l - 1$ or $1 \le 1 \le k + 1$ and $j = l - 1$. Thus we obtain a unique subword for each tuple in

$$\begin{aligned} D :=& \{(i, l - 1, k, l) : i, k, l - 1 \ge 1, i \le k + 1\} \\ & \cup \{(0, j, k, l) : j, k, l - 1 \ge 1, j \le l - 1\}. \end{aligned}$$

We will find generating functions a, b, c, d such that $[x^n]a(x)$ is the number of subwords of length n corresponding to A, etc.

- $A = \epsilon \cup a^+ \cup b^+ \cup a^+b^+ \cup b^+a^+ \cup a^+b^+a^+ \cup b^+a^+b^+$ is uniquely generated. This can be translated to

$$\begin{aligned} a(x) &= 1 + \frac{2x}{1-x} + \frac{2x^2}{(1-x)^2} + \frac{2x^3}{(1-x)^3} \\ &= \frac{1 - x + x^2 + x^3}{(1-x)^3}. \end{aligned}$$

- There is a length $i + j + k + l$ subword for each $(i, j, k, l) \in B$. Thus,

$$\begin{aligned} b(x) &= \sum_{(i,j,k,l) \in B} x^{i+j+k+l} = \sum_{k \ge 1, 1 \le i \le k+1} \sum_{j \ge 1, 1 \le l \le j+1} x^{i+j+k+l} \\ &= \left(\sum_{k \ge 1, 1 \le i \le k+1} x^{i+k} \right)^2 \\ &= \left(\sum_{k \ge 1} \sum_{i=1}^{k+1} x^{i+k} \right)^2 \\ &= \left(\frac{x^2 + x^3 - x^4}{(1-x)^2(1+x)} \right)^2 = \frac{x^4 + 2x^5 - x^6 - 2x^7 + x^8}{(1-x)^4(1+x)^2}. \end{aligned}$$

- The tuples in C, D corresponds to prefixes. For each tuple $(i, j, k, l) \in C, D$, there is

exactly one subword with prefix $a^i b^j c^k d^l a$. Thus we have

$$\begin{aligned} c(x) &= \sum_{(i,j,k,l)\in C} \frac{x^{i+j+k+l+1}}{1-x} = \sum_{k\geq 2}\sum_{j=1}^{k-1} \frac{x^{j+k+2}}{1-x} + \sum_{k\geq 2} \frac{x^{2k+2}}{1-x} \\ &= \frac{x^5}{(1-x)^3(1+x)} + \frac{x^6}{(1-x)^2(1+x)} \\ &= \frac{x^5+x^6-x^7}{(1-x)^3(1+x)}. \end{aligned}$$

$$\begin{aligned} d(x) &= \sum_{l\geq 2,k\geq 1,1\leq u\leq k+1} \frac{x^{i+k+2l}}{1-x} + \sum_{l\geq 2,k\geq 1,1\leq j\leq l-1} \frac{x^{j+k+l+1}}{1-x} \\ &= \frac{x^6+x^7-x^8}{(1-x)^4(1+x)^2} + \frac{x^5+x^6}{(1-x)^4(1+x)^2} = \frac{x^5+2x^6+x^7-x^8}{(1-x)^4(1+x^2)}. \end{aligned}$$

The sum of the generating functions $F(x) = a(x) + b(x) + c(x) + d(x)$ encodes the subword complexity as follows

$$F(x) = \sum_{n\geq 0} \rho_{\mathbf{w}_1}(n) x^n.$$

Thus we obtain

$$\begin{aligned} F(x) &= \frac{1-x+x^2+x^3}{(1-x)^3} + \frac{x^4+2x^5-x^6-2x^7+x^8}{(1-x)^4(1+x)^2} \\ &\quad + \frac{x^5+x^6-x^7}{(1-x)^3(1+x)} + \frac{x^5+2x^6+x^7-x^8}{(1-x)^4(1+x^2)} \\ &= \frac{1-2x+x^2+2x^5-3x^6+x^7}{(1-x)^4} \\ &= \frac{1}{(1-x)^4} - \frac{1}{(1-x)^3} - \frac{2}{(1-x)^2} + \frac{5}{1-x} - 2 + x^2 + x^3. \end{aligned}$$

Thus

$$\rho_{\mathbf{w}_1}(n) = \binom{n+3}{3} - \binom{n+2}{2} - 2\binom{n+1}{1} + 5 = \frac{n^3}{6} + \frac{n^2}{2} - \frac{5n}{3} + 3$$

for all $n \geq 0$, with the exception of $n = 0, 2, 3$ due to the $-2 + x^2 + x^3$ terms. □

4 Subword complexity of $\mathbf{w}_2$

The word $\mathbf{w}_2$ is very similar to $\mathbf{w}_1$, in that the seconds in each minute are concatenated in the reverse order. Thus, the technique used in the previous section will work here as well. We will give the classification of subwords t of $\mathbf{w}_2$, but omit the details of the computations.

Theorem 2. *The subword complexity of* $\mathbf{w}_2$ *is*

$$\rho_{\mathbf{w}_2}(n) = \begin{cases} n^3/6 - 2n/3 + (19 + (-1)^n)/4, & \text{for } n \geq 4; \\ 2^n, & \text{otherwise.} \end{cases}$$

Proof. As before, if a subword t contains $a^i b^j a^k b^l a$, then it contains a second, and that second must be within the $(k+l)$th minute. Thus this uniquely determines t. We may choose this second to be the first second contained in t, so that $a^i b^j a^k b^l a$ is the prefix of t.

- If $a^k b^l$ is the first second of a minute, so $k = 1$, then $a^i b^j$ is the suffix of $a^{l-1}b$. Thus $0 \leq i \leq l-1$ and $j = 1$. Therefore, we obtain a prefix for each tuple in
$$C := \{(i, 1, 1, l) \,:\, 0 \leq i \leq l-1 \text{ and } l \geq 2\}.$$
- If $a^k b^l$ is not the first second of a minute, so $k \geq 2$, then $a^i b^j$ is the suffix of $a^{k-1}b^{l+1}$. Thus either $i = 0$ and $1 \leq j \leq l+1$ or $1 \leq i \leq k-1$ and $j = l+1$. Thus we obtain a prefix for each tuple in
$$\begin{aligned} D :=& \{(i, l+1, k, l) \,:\, 1 \leq i \leq k-1, k \geq 2, l \geq 1\} \\ &\cup \{(0, j, k, l) \,:\, 1 \leq j \leq l+1, k \geq 2, l \geq 1\}. \end{aligned}$$

Next, consider t of the form $a^*b^*a^*b^*$. It must lie within two consecutive seconds, either $a^p b^{q+1} a^{p+1} b^q$ for $p, q \geq 1$ if they are within the same minute, or $a^p b a b^q$ if they are not. Clearly, every word in $a^*b^* \cup b^*a^*$ is a valid subword, by taking p, q large enough. We also get every word in $a^+bb^+a^+ \cup b^+aa^+b^+$. But every word appearing in a^+ba^+ must be a subword of $a^p b a b^q$. Therefore, the second block of as must have length 1, and similarly for words in b^+ab^+. Therefore, every word in

$$A := a^*b^*a^* \cup b^*a^*b^* \setminus \left(a^+baa^+ \cup bb^+ab^+\right)$$

appears as a subword of $\mathbf{w}_2$.

The remaining words of the form $a^*b^*a^*b^*$ are $a^i b^j a^k b^l$ with $i, j, k, l \geq 1$. For this to be a subword of $a^p b^{q+1} a^{p+1} b^q$, we must have $j, k \geq 2, i \leq k-1, l \leq j-1$. For it to be a subword of $a^p b a b^q$, we must have $j, k = 1$. Thus we obtain a subword for each tuple in

$$B := \{(i, j, k, l) \,:\, j, k \geq 2, 1 \leq i \leq k-1, 1 \leq l \leq j-1\} \cup \{(i, 1, 1, l) \,:\, i, l \geq 1\}.$$

As before, we construct generating functions for each of the sets, and add them. This gives a generating function $F(x)$ which encodes the subword complexity

$$F(x) = a(x) + b(x) + c(x) + d(x) = \sum_{n \geq 0} \rho_{\mathbf{w}_2}(n) x^n.$$

After the calculations, we obtain

$$F(x) = \frac{5 - 11x + 3x^2 + 10x^3 - 5x^4}{(1-x)^4(1+x)} - 4 - 2x - x^2 + x^3,$$

so that $\rho_{\mathbf{w}_2}(n) = \frac{n^3}{6} - \frac{2n}{3} + \frac{19+(-1)^n}{4}$ for $n \geq 0$ with exceptions at $n = 0, 1, 2, 3$ due to the $-4 - 2x - x^2 + x^3$ terms. □

5 Acknowledgments

We would like to thank Tim Smith and Jeffrey Shallit for suggesting the problem.

References

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

[2] Jean-Paul Allouche, Sur la complexité des suites infinies, *Bull. Belg. Math. Soc. Simon Stevin* **1** (1994), 133–143.

[3] J. Cassaigne, Special factors of sequences with linear subword complexity. In J. Dassow, G. Rozenberg, and A. Salomaa, editors, *Developments in Language Theory II*, pp. 25–34. World Scientific, 1996.

[4] A. Ehrenfeucht and G. Rozenberg, On the subword complexity of square-free D0L languages, *Theoret. Comput. Sci.* **16** (1981), 25–32.

[5] Sébastien Ferenczi, Complexity of sequences and dynamical systems, *Discrete Math.* **206** (1999), 145–154.

[6] Sébastien Ferenczi and Zoltán Kása, Complexity for finite factors of infinite sequences, *Theoret. Comput. Sci.* **218** (1999), 177–195.

[7] F. Mignosi, Infinite words with linear subword complexity, *Theoret. Comput. Sci.* **65** (1989), 221–242.

[8] J.-J. Pansiot, Complexité des facteurs des mots infinis engendrés par morphismes itérés. In J. Paredaens, editor, *Proc. 11th Int'l Conf. on Automata, Languages, and Programming (ICALP)*, Vol. 172 of *Lecture Notes in Computer Science*, pp. 380–389. Springer-Verlag, 1984.

[9] J.-J. Pansiot, Bornes inférieures sur la complexité des facteurs des mots infinis engendrés par morphismes itérés. In M. Fontet and K. Mehlhorn, editors, *STACS 84, Proc. 1st Symp. Theoretical Aspects of Comp. Sci.*, Vol. 166 of *Lecture Notes in Computer Science*, pp. 230–240. Springer-Verlag, 1984.

[10] T. Smith, A characterization of morphic words with polynomial growth, *Discrete Mathematics & Theoretical Computer Science* **22** (2020), Paper #3.

2010 *Mathematics Subject Classification*: Primary 68R15; Secondary 05A15.
Keywords: subword complexity, morphic word.

(Concerned with sequence A338760, A338761.)

Received September 10 2020; revised version received October 27 2020. Published in *Journal of Integer Sequences*, November 7 2020.

Return to Journal of Integer Sequences home page.