Extremal infinite overlap-free binary words

Jean-Paul Allouche
CNRS, LRI
Bâtiment 490
F-91405 Orsay Cedex (France)
allouche@lri.fr

James Currie
Department of Mathematics
University of Winnipeg
Winnipeg, Manitoba R3B 2E9 (Canada)
currie@io.uwinnipeg.ca

Jeffrey Shallit
Department of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1 (Canada)
shallit@graceland.uwaterloo.ca

Submitted: August 29, 1997; Accepted: May 3, 1998.

Abstract

Let $\overline{\mathbf{t}}$ be the infinite fixed point, starting with 1, of the morphism $\mu : 0 \to 01$, $1 \to 10$. An infinite word over $\{0,1\}$ is said to be *overlap-free* if it contains no factor of the form $axaxa$, where $a \in \{0,1\}$ and $x \in \{0,1\}^*$. We prove that the lexicographically least infinite overlap-free binary word beginning with any specified prefix, if it exists, has a suffix which is a suffix of $\overline{\mathbf{t}}$. In particular, the lexicographically least infinite overlap-free binary word is $001001\overline{\mathbf{t}}$.

Keywords: Homomorphism, fixed point, overlap-free word.
1991 Mathematics Subject Classification: Primary 68R15.

1 Introduction

Since the pioneering work of Thue [14, 15] (see also [5]) the overlap-free words on a finite alphabet, i.e., those words that do not contain a factor $axaxa$, where x is a word and a a letter, have been studied extensively. The question of extremality (for the lexicographic order) of overlap-free binary infinite words seems to have been addressed only once: Berstel proved [4], (see also [5]), that the lexicographically *greatest* infinite overlap-free word on the binary alphabet $\{0,1\}$ that begins with 0 is the Thue-Morse sequence $\mathbf{t} = 01101001\cdots$, which shows once more the ubiquity of this sequence. (Recall that $\mathbf{t}$ is one of the fixed points of the morphism $0 \longrightarrow 01$, $1 \longrightarrow 10$. We let $\overline{\mathbf{t}} = 10010110\cdots$ denote the other fixed point.) The following natural question then arises: what is the lexicographically *least* overlap-free word on $\{0,1\}$ that begins with 0? Computing the first few terms suggests that this word is $001001100101100110100 1\cdots = 001001\overline{\mathbf{t}}$.

The main result of this paper is the following: *the lexicographically least overlap-free sequence with any specified prefix, if it exists, has a suffix equal to a suffix of* $\overline{\mathbf{t}}$. Furthermore, we give an algorithm to construct this sequence. Of course, replacing "least" by "greatest" and interchanging 0's and 1's gives a dual result.

2 Notation

In what follows we consider words or infinite words (sequences) on the binary alphabet $\{0,1\}$. Words will be denoted by lower-case letters (usually r, s, $\cdots$). The set of all finite words on $\{0,1\}$ is denoted by $\{0,1\}^*$. Elements of $\{0,1\}$ will be denoted by a, b, c, d, $\cdots$. Infinite words will be denoted by boldface small letters $\mathbf{x}$, $\mathbf{y}$, $\cdots$. The length (number of letters) of a (finite) word w is denoted by $|w|$. If w is a word, $\overline{w}$ is the word obtained from w by replacing the 0's by 1's and the 1's by 0's.

We define the morphism μ on $\{0,1\}^*$ by $\mu(0) = 01$, $\mu(1) = 10$, and we extend it to infinite words by continuity. The infinite fixed point of μ beginning with 0 is denoted by $\mathbf{t}$, hence

$$\mathbf{t} = 0110100110010110\cdots.$$

The infinite fixed point of μ beginning with 1 is $\overline{\mathbf{t}}$. These sequences are called the Thue-Morse sequences.

The lexicographic order for infinite words is defined by $\mathbf{x} < \mathbf{y}$ if and only if there exists $k \geq 0$ such that the prefixes of length k of $\mathbf{x}$ and $\mathbf{y}$ are equal, and the $(k+1)$-th letter of $\mathbf{x}$ is 0 while the $(k+1)$-th letter of $\mathbf{y}$ is 1. We note that the morphism μ is order-preserving on the set of infinite words; more precisely, $\mathbf{x} < \mathbf{y} \Leftrightarrow \mu(\mathbf{x}) < \mu(\mathbf{y})$.

3 Tools

In this section we provide some basic tools that will be useful in the proof of our main result. The three lemmas we give demonstrate the close relationship between the morphism μ and overlap-free words. Parts of the lemmas below can be found in the literature (essentially in

[14, 15]; see also [6]). Nevertheless, it seems that in the factorization lemma we give below (Lemma 3), the question of uniqueness of the decomposition was first addressed in [11].

We start with a technical lemma.

Lemma 1

1. If $y, y' \in \{0,1\}^*$, and if there exist $c, d \in \{0,1\}$ such that $u = c\mu(y) = \mu(y')d$, then $u = c\,(\overline{c}\,c)^{|y|}$.

2. If $y, z \in \{0,1\}^*$ satisfy $zz = \mu(y)$, then there exists $x \in \{0,1\}^*$ such that $z = \mu(x)$.

Proof.

1. We first note that y and y' must have the same length. Let $y = a_1\,a_2\,\cdots\,a_s$ and $y' = b_1\,b_2\,\cdots\,b_s$. Then we have:

$$c\,a_1\,\overline{a_1}\,a_2\,\overline{a_2}\,\cdots\,a_{s-1}\,\overline{a_{s-1}}\,a_s\,\overline{a_s} = b_1\,\overline{b_1}\,b_2\,\overline{b_2}\,\cdots\,b_{s-1}\,\overline{b_{s-1}}\,b_s\,\overline{b_s}\,d.$$

Hence $b_1 = c$, $a_1 = \overline{b_1} = \overline{c}$, $b_2 = \overline{a_1} = c$, $a_2 = \overline{b_2} = \overline{c}$, $\cdots$, $b_s = \overline{a_{s-1}} = c$, $a_s = \overline{b_s} = \overline{c}$, and $\overline{a_s} = d$, i.e., $c = d$, and hence finally $u = c(\overline{c}\,c)^s$.

2. Suppose that $zz = \mu(y)$. If $|z|$ is even, the result is clear, since then $|\mu(y)| \equiv 0 \pmod 4$, and hence $|y|$ is even. Hence z is the image by μ of the prefix of y of length $|y|/2$. Let us show that $|z|$ cannot be odd. If it were, let $z = au = vb$, where $a, b \in \{0,1\}$ and u and v are binary words of even length. Then $\mu(y) = zz = vbau$; hence u and v must be the images by μ of binary words, say $u = \mu(r)$ and $v = \mu(s)$. Hence $zz = \mu(s)ba\mu(r)$, and $b = \overline{a}$. But we have $z = a\mu(r) = \mu(s)b$, hence from assertion 1 above, $z = a\,(\overline{a}\,a)^{|r|}$. But the last letter of z cannot be simultaneously equal to a and to $\overline{a}$. ■

Next, we prove that the morphism μ behaves nicely when applied to overlap-free words.

Lemma 2

1. Let $w \in \{0,1\}^*$. Then w is overlap-free if and only if $\mu(w)$ is overlap-free.

2. Let $\mathbf{x}$ be an infinite word on $\{0,1\}$. Then

(a) $\mu(\mathbf{x})$ is overlap-free if and only if $\mathbf{x}$ is overlap-free.

(b) $0\mu(\mathbf{x})$ is overlap-free if and only if $1\mathbf{x}$ is overlap-free.

(c) $1\mu(\mathbf{x})$ is overlap-free if and only if $0\mathbf{x}$ is overlap-free.

(d) $00\mu(\mathbf{x})$ is overlap-free if and only if $1\mathbf{x}$ is overlap-free and $\mathbf{x}$ begins with 101.

(e) $11\mu(\mathbf{x})$ is overlap-free if and only if $0\mathbf{x}$ is overlap-free and $\mathbf{x}$ begins with 010.

3. The Thue-Morse sequences $\mathbf{t}$ and $\overline{\mathbf{t}}$ are overlap-free. The sequences $0\mathbf{t}$, $1\mathbf{t}$, $0\,\overline{\mathbf{t}}$, and $1\,\overline{\mathbf{t}}$ are overlap-free.

4. The sequences $01\,\overline{\mathbf{t}}$, $10\,\overline{\mathbf{t}}$, $110\,\overline{\mathbf{t}}$, and $001001\,\overline{\mathbf{t}}$ are overlap-free.

Proof.
1. This is proved in [15].

2. (a) As in [15], assertion 1 above immediately extends to infinite words, showing that $\mu(\mathbf{x})$ is overlap-free if and only if $\mathbf{x}$ is overlap-free.

(b) Now if the infinite word $0\mu(\mathbf{x})$ contains an overlap, then the word $\mathbf{s} = 10\mu(\mathbf{x})$ *a fortiori* contains an overlap. Since $\mathbf{s} = \mu(1\mathbf{x})$, this implies that $1\mathbf{x}$ contains an overlap. If conversely the word $1\mathbf{x}$ contains an overlap, it can occur in two ways: either the overlap is a prefix of $1\mathbf{x}$, hence $\mathbf{x}$ begins with $z1z1$ for some finite word z, or the overlap occurs “inside” $1\mathbf{x}$, which means that $\mathbf{x}$ itself contains an overlap. In the latter case, $\mu(\mathbf{x})$ contains an overlap, hence $0\mu(\mathbf{x})$ contains an overlap. In the former case, $0\mu(\mathbf{x})$ begins with $0\mu(z)10\mu(z)10$ and hence contains an overlap. This proves (b). The proof of (c) is similar.

(d) Suppose now that $00\mu(\mathbf{x})$ is overlap-free. Then $0\mu(\mathbf{x})$ is also overlap-free; hence, from the preceding argument, $1\mathbf{x}$ is overlap-free. Furthermore, if $\mathbf{x}$ begins with abc where $a, b, c \in \{0, 1\}$, then $0\,0\,a\,\overline{a}\,b\,\overline{b}\,c\,\overline{c}$ and $1abc$ must be overlap-free; hence, easily, $a = 1$, $b = 0$, and $c = 1$. Conversely, suppose that $1\mathbf{x}$ is overlap-free and begins with 101. From the preceding argument this implies that $0\mu(\mathbf{x})$ is overlap-free. Hence any overlap of the word $00\mu(\mathbf{x})$ must in fact be a prefix. Hence, in order to show that $00\mu(\mathbf{x})$ is overlap-free, we will suppose that $00\mu(\mathbf{x})$ begins with $azaza$, with $a \in \{0, 1\}$ and $z \in \{0, 1\}^*$ and obtain a contradiction. By the hypothesis, the word $\mathbf{x}$ begins with 101, say $\mathbf{x} = 101\mathbf{y}$ for some infinite word $\mathbf{y}$, hence the word $00\mu(\mathbf{x})$ is equal to $00100110\mu(\mathbf{y})$, and begins with $azaza$. By inspection, we see that $|az| \geq 6$, hence the word 001001 occurs as a factor of $10\mu(\mathbf{y})$. Since this last word begins with 1, this means that there exists a nonempty word w such that $10\mu(\mathbf{y}) = w001001\cdots$. But $10\mu(\mathbf{y})$ is overlap-free, as it is a factor of the overlap-free word $0\mu(\mathbf{x})$. This implies that 001001 can be extended to the left (with the last letter of w) to an overlap-free word, although 001001 is clearly not extendable to the left to an overlap-free word. This proves (d). The proof of (e) is similar.

3. The Thue-Morse sequence $\overline{\mathbf{t}}$ is the limit, as $n \to \infty$, of the binary words $\mu^n(1)$ which are all overlap-free, since 1 is overlap-free: this is Thue's proof [15].

Let $a \in \{0, 1\}$, and suppose that $a\,\overline{\mathbf{t}}$ contains an overlap. Then the overlap must be a prefix of $a\,\overline{\mathbf{t}}$. Hence there exists a finite word z such that $\overline{\mathbf{t}}$ begins with $zaza$. We will show that this is impossible. More precisely we show that the sequence $\overline{\mathbf{t}}$ cannot begin with $z0z0$ nor $z1z1$ for any finite word z. Let us suppose that $\overline{\mathbf{t}}$ begins with $zaza$ with z a finite word and $a \in \{0, 1\}$, and take z a word of minimal length for which there exists $a \in \{0, 1\}$ such that $\overline{\mathbf{t}}$ begins with $zaza$. Since $zaza$ has even length and $\overline{\mathbf{t}} = \mu(\overline{\mathbf{t}})$, the word $zaza$ is the image by μ of a binary word. Hence, applying Lemma 1, the word za itself is the image of a binary word by μ, say $za = \mu(y)$. Hence y must end with $\overline{a}$, say $y = x\,\overline{a}$. Then $\overline{\mathbf{t}}$ begins with $zaza = \mu(x\,\overline{a}\,x\,\overline{a})$. But $\overline{\mathbf{t}} = \mu(\overline{\mathbf{t}})$. Hence $\overline{\mathbf{t}}$ begins with $x\,\overline{a}\,x\,\overline{a}$, which contradicts the minimality of the length of z.

4. Since $01\,\overline{\mathbf{t}} = \mu(0\,\overline{\mathbf{t}})$, and $10\,\overline{\mathbf{t}} = \mu(1\,\overline{\mathbf{t}})$, we deduce from 2.(a) and 3 above that these two words are overlap-free. The word $110\,\overline{\mathbf{t}}$ is overlap-free, since it is a suffix of the word

$0110\,\overline{\mathbf{t}} = \mu^2(0\,\overline{\mathbf{t}})$ that is overlap-free from 2.(a) and 3 above. Finally to prove that the word $001001\,\overline{\mathbf{t}}$ is overlap-free, we write $001001\,\overline{\mathbf{t}} = 00\mu(10\,\overline{\mathbf{t}})$. This last word is overlap-free from (d) above, since we have just proved that $110\,\overline{\mathbf{t}}$ is overlap-free. ■

We now give a factorization lemma for both finite and infinite overlap-free binary words.

Lemma 3

1. If $x \in \{0,1\}^*$ is overlap-free, then there exist u, v, y with $u, v \in \{\varepsilon, 0, 1, 00, 11\}$ and $y \in \{0,1\}^*$ an overlap-free word, such that $x = u\mu(y)v$. Furthermore this decomposition is unique if $|x| \geq 7$, and u (resp. v) is completely determined by the prefix (resp. suffix) of length 7 of x. The bound 7 is sharp as shown by the example $x = 001011 = 00\mu(1)11 = 0\mu(00)1$.

2. If $\mathbf{x}$ is an infinite overlap-free word on $\{0,1\}$, then there exist $u \in \{\varepsilon, 0, 1, 00, 11\}$ and an infinite overlap-free word $\mathbf{y}$ on $\{0,1\}$ such that $\mathbf{x} = u\mu(\mathbf{y})$. The prefix u is completely determined by the prefix of $\mathbf{x}$ of length 4, except if $\mathbf{x}$ begins with 0010 or 1101, in which case the word u is completely determined by the prefix of $\mathbf{x}$ of length 5.

Proof. First note that the word y is necessarily overlap-free by assertion 1 of Lemma 2.

The existence of the decomposition for finite words is proved in [10], in the proof of Theorem 6.4. The uniqueness is proved in [11], Lemma 2.2.

Finally in this factorization of x, the word u (resp. v) depends only on the prefix (resp. suffix) of x of length 7. This is recalled in the following table of all possible prefixes (resp. suffixes) of length ≥ 7: by mere inspection, we see that the word u (resp. v) is uniquely determined, knowing the decomposition does exist. (Note however that some of the words below, e.g., 0011011, might be non-extendable to longer overlap-free words.)

x	u
$0010011\cdots$	00
$0010110\cdots$	0
$0011001\cdots$	0
$0011010\cdots$	0
$0011011\cdots$	0
$0100101\cdots$	0

x	u
$0100110\cdots$	0
$0101100\cdots$	ε
$0101101\cdots$	ε
$0110010\cdots$	ε
$0110011\cdots$	ε
$0110100\cdots$	ε

x	u
$1001011\cdots$	ε
$1001100\cdots$	ε
$1001101\cdots$	ε
$1010010\cdots$	ε
$1010011\cdots$	ε
$1011001\cdots$	1

x	u
$1011010\cdots$	1
$1100100\cdots$	1
$1100101\cdots$	1
$1100110\cdots$	1
$1101001\cdots$	1
$1101100\cdots$	11

x	v
$\cdots 0010011$	1
$\cdots 0010110$	ε
$\cdots 0011001$	ε
$\cdots 0011010$	ε
$\cdots 0011011$	11
$\cdots 0100101$	ε

x	v
$\cdots 0100110$	ε
$\cdots 0101100$	0
$\cdots 0101101$	1
$\cdots 0110010$	0
$\cdots 0110011$	1
$\cdots 0110100$	0

x	v
$\cdots 1001011$	1
$\cdots 1001100$	0
$\cdots 1001101$	1
$\cdots 1010010$	0
$\cdots 1010011$	1
$\cdots 1011001$	ε

x	v
$\cdots 1011010$	ε
$\cdots 1100100$	00
$\cdots 1100101$	ε
$\cdots 1100110$	ε
$\cdots 1101001$	ε
$\cdots 1101100$	0

2. For any prefix x_n of $\mathbf{x}$ such that, say $|x_n| = n$, assertion 1 of Lemma 3 above gives the existence of u_n and v_n in $\{\varepsilon, 0, 1, 00, 11\}$ and of an overlap-free word y_n on $\{0,1\}$, such that $x_n = u_n \mu(y_n) v_n$. Furthermore, u_n does not depend on n for $n \geq 7$. Say $u_n = u$. Hence $\mathbf{x} = \lim_{n\to\infty} x_n = \lim_{n\to\infty} u\mu(y_n)v_n$. Since $\mu(y_n)$ goes to infinity, this implies $\mathbf{x} = \lim_{n\to\infty} u\mu(y_n)$. Hence $\lim_{n\to\infty} \mu(y_n)$ exists, which gives the existence of $\mathbf{y} = \lim_{n\to\infty} y_n$. This sequence is overlap-free as it is a limit of overlap-free words, and we have $\mathbf{x} = u\mu(\mathbf{y})$. Note that in the decomposition of x_n, v_n cannot be equal to 00 or 11, since these words are not images of a word by μ. Hence v_n is either empty or is equal to 0 or 1. Finally, inspecting the table above shows that the prefix of length 4 of a 7-letter word determines the word u in all cases but the two cases where the word begins with 0010 or 1101 for which we need to look at the prefix of length 5. ■

4 The main result

Before proving our main theorem, we need to state two results on overlap-free sequences and to study eight particular cases. This is the purpose of the following proposition.

Proposition 1 *Let w be a binary word, and let $\mathbf{x}(w)$ be the lexicographically least overlap-free sequence (if it exists) beginning with w. Let $\overline{\mathbf{t}}$ be the Thue-Morse sequence beginning with 1. Then*

(a) $\mathbf{x}(1) = \overline{\mathbf{t}}$.

(b) If $\mathbf{x}(w)$ exists, and if z is a finite word such that $z\,\mathbf{x}(w)$ is overlap-free, then $\mathbf{x}(zw)$ exists and $\mathbf{x}(zw) = z\,\mathbf{x}(w)$. In particular $\mathbf{x}(01) = 0\overline{\mathbf{t}}$ and $\mathbf{x}(11) = 1\overline{\mathbf{t}}$.

If $\mathbf{x}(w)$ exists, and if ww_1 is a prefix of $\mathbf{x}(w)$, then $\mathbf{x}(ww_1)$ exists and $\mathbf{x}(ww_1) = \mathbf{x}(w)$. In particular $\mathbf{x}(10) = \overline{\mathbf{t}}$.

(c) We have $\mathbf{x}(001) = 001001\overline{\mathbf{t}}$. Furthermore, $\mathbf{x}(0) = \mathbf{x}(00) = 001001\overline{\mathbf{t}}$.

(d) $\mathbf{x}(010) = 0\overline{\mathbf{t}}$.

(e) $\mathbf{x}(011) = 01\overline{\mathbf{t}}$.

(f) $\mathbf{x}(100) = \overline{\mathbf{t}}$.

(g) $\mathbf{x}(101) = 10\overline{\mathbf{t}}$.

(h) $\mathbf{x}(110) = 1\overline{\mathbf{t}}$.

(i) $\mathbf{x}(0010) = 001001\overline{\mathbf{t}}$.

(j) $\mathbf{x}(1101) = 110\overline{\mathbf{t}}$.

Proof.

(a) We note that $\overline{\mathbf{t}}$ is an overlap-free sequence beginning with 1. On the other hand, the set of overlap-free sequences beginning with a given word is a closed set (a limit of overlap-free sequences is also overlap-free). Let $\mathbf{y}$ be the least overlap-free sequence beginning with 1. Then $\mathbf{y} \leq \overline{\mathbf{t}}$. Hence $\mathbf{y}$ must start with 1001. Now, from Lemma 3 there exists an infinite sequence $\mathbf{x}$ such that $\mathbf{y} = \mu(\mathbf{x})$. The sequence $\mathbf{x}$ is overlap-free from assertion 2 (a) in Lemma 2. As $\mathbf{x}$ begins with 1, we have $\mathbf{y} \leq \mathbf{x}$. Hence $\mu(\mathbf{y}) \leq \mu(\mathbf{x}) = \mathbf{y}$, since μ is order-preserving. The sequence $\mu(\mathbf{y})$ is overlap-free as it is the image under μ of an overlap-free sequence. Hence we must have $\mu(\mathbf{y}) = \mathbf{y}$ from the minimality of $\mathbf{y}$. Hence $\mathbf{y} = \overline{\mathbf{t}}$. [We remark that this result was already proved by Berstel [4] in the following "dual" form: the lexicographically *greatest* infinite overlap-free word on the binary alphabet $\{0, 1\}$ that begins with 0 is the Thue-Morse sequence $\mathbf{t} = 01101001 \cdots$.]

(b) Suppose that $\mathbf{x}(w)$ exists, and that $z\,\mathbf{x}(w)$ is overlap-free. Hence $z\,\mathbf{x}(w)$ is an overlap-free sequence beginning with zw. Let $z\,\mathbf{y}$ be an infinite overlap-free sequence beginning with zw, that satisfies $z\,\mathbf{y} \leq z\,\mathbf{x}(w)$. Then $\mathbf{y}$ is an overlap-free sequence beginning with w, and that satisfies $\mathbf{y} \leq \mathbf{x}(w)$. Hence $\mathbf{y} = \mathbf{x}(w)$, and $z\,\mathbf{y} = z\,\mathbf{x}(w)$.

Suppose now that $\mathbf{x}(w)$ exists, and that ww_1 is a prefix of $\mathbf{x}(w)$. Suppose that $\mathbf{y}$ is an overlap-free sequence that begins with ww_1, and that $\mathbf{y} \leq \mathbf{x}(w)$. Since $\mathbf{y}$ also begins with w, we must have $\mathbf{y} = \mathbf{x}(w)$.

(c) We note that an overlap-free sequence beginning with 001 cannot be smaller than $00100 \cdots$, hence (no overlap) than $0010011 \cdots$. Now $001001\,\overline{\mathbf{t}}$ is overlap-free from assertion 4 of Lemma 2. It then suffices to note that, from the first assertion in (b) above, $001001\,\overline{\mathbf{t}}$ is the lexicographically least sequence beginning with 0010011. Hence $\mathbf{x}(001) = \mathbf{x}(0010011) = 001001\,\overline{\mathbf{t}}$. Similarly an overlap-free sequence beginning with 0 or 00 cannot be smaller than $00100 \cdots$, and the same reasoning applies: hence $\mathbf{x}(0) = \mathbf{x}(00) = 001001\overline{\mathbf{t}}$.

(d) to (j) Using properties 3 and 4 of Lemma 2, and assertions (a), (b) and (c) above, we have:
$\mathbf{x}(010) = 0\,\overline{\mathbf{t}}$, since $\mathbf{x}(01) = 0\,\overline{\mathbf{t}}$;
$\mathbf{x}(011) = 01\,\overline{\mathbf{t}}$;
$\mathbf{x}(100) = \overline{\mathbf{t}}$, since $\mathbf{x}(1) = \overline{\mathbf{t}}$;
$\mathbf{x}(101) = 10\,\overline{\mathbf{t}}$;
$\mathbf{x}(110) = 1\,\overline{\mathbf{t}}$, since $\mathbf{x}(10) = \overline{\mathbf{t}}$;
$\mathbf{x}(0010) = 001001\,\overline{\mathbf{t}}$, since $\mathbf{x}(001) = 001001\,\overline{\mathbf{t}}$;
$\mathbf{x}(1101) = 110\,\overline{\mathbf{t}}$, since $\mathbf{x}(110) = 1\,\overline{\mathbf{t}}$. ■

We are now ready to state and prove our main theorem.

Theorem 1 *Let w be a finite binary word such that there exists an infinite overlap-free binary sequence which begins with w. Let $\mathbf{x}$ be the lexicographically least infinite overlap-free binary sequence which begins with w. Then there exists a suffix of $\mathbf{x}$ that is equal to a suffix of the Thue-Morse sequence $\overline{\mathbf{t}} = 10010110 \cdots$. The construction of this minimal sequence is given by an explicit algorithm.*

Proof. We will prove the result by induction on the length of w. The induction will show the construction to be algorithmic.

Using Proposition 1 above, we see that the result has already been obtained if the word w has length at most 3 or is equal to 0010 or 1101. The result is also true if $w = 00100$ and $w = 001001$, using Proposition 1 (b), and the equality $\mathbf{x}(0010) = 001001\overline{\mathbf{t}}$: this gives $\mathbf{x}(00100) = \mathbf{x}(001001) = 001001\overline{\mathbf{t}}$. Furthermore the result holds also true if $w = 11011$ or $w = 110110$, since we obtain, using Lemma 2 (e) and Proposition 1 (i) and (b), that $\mathbf{x}(11011) = \mathbf{x}(110110) = 110110010110\overline{\mathbf{t}}$.

If w has length ≥ 4, and is not one of the six words above, we will prove that $\mathbf{x}(w)$ ends with $\mu(\mathbf{x}(w'))$ for some w' such that $|w'| < |w|$. Since $\mu(\overline{\mathbf{t}}) = \overline{\mathbf{t}}$, the image by μ of an infinite word having a suffix in common with $\overline{\mathbf{t}}$ has the same property: hence the induction hypothesis for w' will imply the required result for w.

Using Lemma 3, write $w = u\mu(y)v$. Note that, from the proof of the second part of Lemma 3, the word v is either empty or has length 1. Furthermore $\mathbf{x}(w)$ has to begin with $u\mu(y)v\overline{v} = w\overline{v}$, hence $\mathbf{x}(w) = \mathbf{x}(w\overline{v})$.

- If u is empty, then $\mathbf{x}(w) = \mathbf{x}(w\overline{v}) = \mathbf{x}(\mu(yv)) = \mu(\mathbf{x}(yv))$. We have $|yv| < |\mu(y)v| = |w|$, since y cannot be empty ($|w| \geq 4$).

- If $u = 0$, using Lemma 2 and Proposition 1 we have $1\mathbf{x}(w) = 1\mathbf{x}(w\overline{v}) = 1\mathbf{x}(0\mu(yv)) = \mathbf{x}(10\mu(yv)) = \mathbf{x}(\mu(1yv)) = \mu(\mathbf{x}(1yv))$. We have $|1yv| < |0\mu(y)v| = |w|$, since y cannot be empty. The same reasoning holds for $u = 1$.

- If $u = 00$, we have $\mathbf{x}(w) = \mathbf{x}(w\overline{v}) = \mathbf{x}(00\mu(yv))$. Let $\mathbf{x}(w) = 00\mu(yv)\mu(\mathbf{z})$. Using Lemma 2, we know that $1yv(\mathbf{z})$ has to be overlap-free, and that $yv(\mathbf{z})$ has to begin with 101.

- If $4 \leq |w| \leq 6$, the equality $w = 00\mu(y)v$ easily implies that w is one of the words 0010, 00100, 00101, or 001001. The cases $w \in \{0010, 00100, 001001\}$ have been excluded. The case $w = 00101$ is impossible: namely $\mathbf{x}(00101) = \mathbf{x}(001011)$, since 001010 contains an overlap. But the word 001011 is not of the form $00\mu(y)v$ with $|v| \leq 1$.

- If $|w| \geq 7$, then the equality $w = 00\mu(y)v$ shows that $|yv| \geq 3$, hence yv has to begin with 101. Let $yv = 101y'$. We thus have $\mathbf{x}(00\mu(yv)) = \mathbf{x}(00\mu(101y')) = 00\mu(101y'\mathbf{z})$. Then, by Lemma 2, $1(101y'\mathbf{z}) = \mathbf{x}(1101y')$. Finally $|1101y'| = |1yv| < |00\mu(y)v| = |w|$.

The same reasoning holds if $u = 11$. ■

Remark 1.
The algorithm we give resembles an algorithm given in [11] to decide (in linear time) whether a *finite* binary word contains an overlap.

Remark 2.
- Let us show an example of how the algorithm works. This example also shows that it is possible for the lexicographically least sequence beginning with a word w to have no suffix equal to the Thue-Morse sequence $\overline{\mathbf{t}}$. Let $w = 0010110$. If there exists an infinite overlap-free

sequence $\mathbf{s}$ beginning with w, say $\mathbf{s} = w\mathbf{x}$, then its factorization must be $\mathbf{s} = 0\mu(001\mathbf{y})$. Now $\mathbf{s}$ is the least overlap-free sequence beginning with w if and only if $1001\mathbf{y}$ is the least overlap-free sequence beginning with 1001 (assertion 2 (b) in Lemma 2). Now the factorization of $1001\mathbf{y}$ must be $\mu(10\mathbf{z})$, with $\mathbf{y} = \mu(\mathbf{z})$. And $1001\mathbf{y}$ is the least overlap-free sequence beginning with 1001 if and only if $10\mathbf{z}$ is the least overlap-free sequence beginning with 10. From Proposition 1, we see that the least overlap-free sequence beginning with 100 is $\overline{\mathbf{t}}$, and that the least overlap-free sequence beginning with 101 is $10\overline{\mathbf{t}}$. Hence the least overlap-free sequence beginning with 10 is $\overline{\mathbf{t}}$. Hence $10\mathbf{z} = \overline{\mathbf{t}}$. We thus have successively $1001\mathbf{y} = 1001\mu(\mathbf{z}) = \mu(10\mathbf{z}) = \mu(\overline{\mathbf{t}}) = \overline{\mathbf{t}}$. Then $1\mathbf{s} = 10\mu(001\mathbf{y}) = \mu(1001\mathbf{y}) = \mu(\overline{\mathbf{t}}) = \overline{\mathbf{t}}$. Hence $\mathbf{s}$ is the sequence obtained by deleting the first letter of $\overline{\mathbf{t}}$.

Finally, we give as a corollary of the above result a new proof of a result obtained in [1, 3] in relation with iterations of continuous functions and kneading sequences.

Corollary 1 *For a binary sequence* $\mathbf{s}$ *let* $S(\mathbf{s})$ *be the sequence obtained by deleting the first letter of* $\mathbf{s}$ *(S is also called the shift). Let* S^k *be the k-th iterate of the map* S. *Let* $\mathbf{a} = S(\mathbf{t}) = 11010011001\cdots$. *Then, for each* $k \geq 1$, *we have* $\overline{\mathbf{a}} < S^k(\mathbf{a}) < \mathbf{a}$.

Proof. Since the sequence $\mathbf{t}$ is overlap-free, so is the sequence $\mathbf{a}$. Hence $\mathbf{a}$ cannot be periodic. This implies that $S^k(\mathbf{a})$ cannot be equal to either $\mathbf{a}$ or $\overline{\mathbf{a}}$ for $k \geq 1$. We thus have only to prove that, for all $k \geq 1$ (actually this is true for all $k \geq 0$), $\overline{\mathbf{a}} \leq S^k(\mathbf{a}) \leq \mathbf{a}$. We first prove that, for each $k \geq 0$, we have $\overline{\mathbf{a}} \leq S^k(\mathbf{a})$. Since $\overline{\mathbf{a}} = 0010110\cdots$ we are done if $S^k(\mathbf{a})$ begins with 1 or with 01 or with 0011. If $S^k(\mathbf{a})$ begins with 0010, it cannot begin with 00100. For if this were the case, $S^k(\mathbf{a})$ would begin with 001001 since it is overlap-free. But the word 001001 is not extendable to the left into an overlap-free word, hence cannot occur in the sequence $\mathbf{a}$, since this sequence is extendable to the left into $\mathbf{t}$ which is overlap-free.

Hence $S^k(\mathbf{a})$ begins with 00101, hence with 001011, hence with 0010110. But, as shown by the remark above, the least overlap-free sequence beginning with 0010110 is $S(\overline{\mathbf{t}}) = \overline{\mathbf{a}}$. Hence $S^k(\mathbf{a}) \geq \overline{\mathbf{a}}$.

We now prove that, for each $k \geq 0$, we have $S^k(\mathbf{a}) \leq \mathbf{a}$. Since $\mathbf{a} = 11010011\cdots$, we are done if $S^k(\mathbf{a})$ begins with 0, 10, or 1100. Hence we can suppose that $S^k(\mathbf{a})$ begins with 1101.

It is easy to see that $S^k(\mathbf{a})$ cannot begin with 11011 because the only substrings of $\mathbf{t}$ starting at even positions are 10 and 01. Hence $S^k(\mathbf{a})$ begins with 11010, hence with 110100, hence with 1101001. But the "dual" of the above remark shows that the lexicographically greatest overlap-free sequence beginning with 110100110 is $S(\mathbf{t}) = \mathbf{a}$ and the result is proved. ■

Remark 3.

An even simpler argument shows that the sequence $\mathbf{b} = 001001\overline{\mathbf{t}}$ satisfies, for all $k \geq 0$, $\mathbf{b} \leq S^k(\mathbf{b}) \leq \overline{\mathbf{b}}$.

We give another simple corollary.

Corollary 2 *The lexicographically least overlap-free sequence that is extendable to the left to a doubly-infinite overlap-free sequence is* $S(\overline{\mathbf{t}})$.

Proof. We have seen that the lexicographically least sequence on $\{0,1\}$ is the sequence $001001\overline{\mathbf{t}}$. This sequence cannot be extended to the left to an overlap-free sequence, since concatenating either 0 or 1 to the front creates an overlap. Now let us consider the sequence $S(\overline{\mathbf{t}})$. It begins with 0010110 and is the least overlap-free sequence having this prefix from Remark 2 above. It is not possible to find a smaller sequence that is extendable to the left to an overlap-free doubly-infinite sequence, since the prefix 00100 is forbidden from the claim above. Hence such a sequence must begin with 00101, hence with 0010110.

On the other hand, it is well-known that $\overline{\mathbf{t}}$ is extendable to the left to a doubly-infinite overlap-free sequence ([15, Satz 7], [12], [5, p. 30]). ■

5 The lexicographically least square-free sequence

As already proved by Thue, it is possible to construct, on a three-letter alphabet, a sequence without squares, i.e., without any factor of the form ww, where w is a nonempty word. (It is readily checked that there is no square-free word of length ≥ 4, and hence no square-free infinite word, on a two-letter alphabet.) Now take the alphabet to be $\{0,1,2\}$. What is the lexicographically least square-free sequence over $\{0,1,2\}$? We do not know a simple description of this sequence. Using the results of Shelton [13, 8] it can be proved that the first twenty terms are

$$01020120210120102012.$$

In particular, is it true that this sequence can be generated by a finite automaton (in the sense of [7]; see also [2, 9])? As proved above the lexicographically least overlap-free sequence over $\{0,1\}$ is $001001\overline{\mathbf{t}}$, and hence is 2-automatic.

6 Acknowledgments

We thank the referee for many useful remarks, particularly those concerning our proof of the main theorem. We also thank Larry Cummings, who read a draft of this paper and made many valuable suggestions. The first two authors acknowledge with thanks the hospitality of the University of Waterloo, where this paper was largely written.

References

[1] J.-P. Allouche, Théorie des nombres et automates, Thèse d'État, Université Bordeaux I, 1983.

[2] J.-P. Allouche, Automates finis en théorie des nombres, *Exposition. Math.* **5** (1987), 239–266.

[3] J.-P. Allouche and M. Cosnard, Itérations de fonctions unimodales et suites engendrées par automates, *C. R. Acad. Sci. Paris, Sér. A* **296** (1983), 159–162.

[4] J. Berstel, A rewriting of Fife's theorem about overlap-free words, in J. Karhumäki, H. Maurer, G. Rozenberg, eds., *Results and Trends in Theoretical Computer Science*, Lecture Notes in Computer Science **812**, Springer-Verlag, 1994, pp. 19–29.

[5] J. Berstel, Axel Thue's papers on repetitions in words: a translation, Publications du Laboratoire de Combinatoire et d'Informatique Mathématique, Université du Québec à Montréal **20**, 1995.

[6] J. Berstel, P. Séébold, A characterization of overlap-free morphisms, *Disc. Appl. Math.* **46** (1993), 275–281.

[7] A. Cobham, Uniform tag sequences, *Math. Systems Theory* **6** (1972), 164–192.

[8] J. D. Currie, On the structure and extendibility of k-power free words, *Eur. J. Comb.* (1995) **16**, 111–124.

[9] F. M. Dekking, M. Mendès France, and A. van der Poorten, Folds!, *Math. Intelligencer* **4** (1982), 130–138, 173–181, 190–195.

[10] Y. Kobayashi, Repetition-free words, *Theoret. Comput. Sci.* **44** (1986), 175–197.

[11] A. J. Kfoury, A linear-time algorithm to decide whether a binary word contains an overlap, *RAIRO Inform. Théor. App.* **22** (1988), 135–145.

[12] M. Morse and G. A. Hedlund, Unending chess, symbolic dynamics and a problem in semigroups, *Duke Math. J.* **11** (1944), 1-7.

[13] R. O. Shelton (I, II, III) and R. P. Soni (II, III), Aperiodic words on three symbols I, II, III, *J. Reine Angew. Math.* **321; 327; 330** (1981), 195–209; 1–11; 44–52.

[14] A. Thue, Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in T. Nagell, ed., *Selected Mathematical Papers of Axel Thue*, Universitetsforlaget, Oslo, 1977, pp. 139–158.

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