Words Avoiding Reversed Subwords

Narad Rampersad and Jeffrey Shallit
School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
CANADA
nrampersad@math.uwaterloo.ca
shallit@graceland.math.uwaterloo.ca

Abstract

We examine words w satisfying the following property: if x is a subword of w and $|x|$ is at least k for some fixed k, then the reversal of x is not a subword of w.

1 Introduction

Let Σ be a finite, nonempty set called an *alphabet*. We denote the set of all finite words over the alphabet Σ by Σ^*. The empty word is represented by ϵ. Let Σ_k denote the alphabet $\{0, 1, \ldots, k-1\}$.

Let $\mathbb{N}$ denote the set $\{0, 1, 2, \ldots\}$. An *infinite word* is a map from $\mathbb{N}$ to Σ. The set of all infinite words over the alphabet Σ is denoted Σ^ω.

A map $h : \Sigma^* \to \Delta^*$ is called a *morphism* if h satisfies $h(xy) = h(x)h(y)$ for all $x, y \in \Sigma^*$. A morphism may be defined by specifying its action on Σ. Morphisms may also be applied to infinite words in the natural way.

If $w \in \Sigma^*$ is written $w = w_1 w_2 \cdots w_n$, where each $w_i \in \Sigma$, then the *reversal* of w, denoted w^R, is the word $w_n w_{n-1} \cdots w_1$.

If y is a nonempty word, then the word $yyy \cdots$ is written as y^ω. If an infinite word $\mathbf{w}$ can be written in the form y^ω for some nonempty y, then $\mathbf{w}$ is said to be *periodic*. If $\mathbf{w}$ can be written in the form $y' y^\omega$ for some nonempty y, then $\mathbf{w}$ is said to be *ultimately periodic*.

A *square* is a word of the form xx, where $x \in \Sigma^*$ is nonempty. A word w' is called a *subword* (resp. a *prefix* or a *suffix*) of w if w can be written in the form $uw'v$ (resp. $w'v$ or uw') for some $u, v \in \Sigma^*$. We say a word w is *squarefree* (or *avoids squares*) if no subword of w is a square.

JCMCC **54** (2005), pp. 157-164

2 Avoiding reversed subwords

Szilard [6] asked the following question:

> Does there exist an infinite word $\mathbf{w}$ such that if x is a subword of $\mathbf{w}$, then x^R is not a subword of $\mathbf{w}$?

Clearly there must be some restriction on the length of x: if $|x| = 1$, then all nonempty words fail to have the desired property. For $|x| \geq 2$, however, we have the following result.

Theorem 1. *There exists an infinite word* $\mathbf{w}$ *over* Σ_3 *such that if* x *is a subword of* $\mathbf{w}$ *and* $|x| \geq 2$*, then* x^R *is not a subword of* $\mathbf{w}$*. Furthermore,* $\mathbf{w}$ *is unique up to permutation of the alphabet symbols.*

Proof. Note that if $|x| \geq 3$ and both x and x^R are subwords of $\mathbf{w}$, then there is a prefix x' of x such that $|x'| = 2$ and $(x')^R$ is a suffix of x^R. Hence it suffices to show the theorem for $|x| = 2$. We show that the infinite word

$$\mathbf{w} = (012)^\omega = 012012012012\cdots$$

has the desired property. To see this, consider the set $\mathcal{A}$ consisting of all subwords of $\mathbf{w}$ of length two. We have $\mathcal{A} = \{01, 12, 20\}$. Noting that if $x \in \mathcal{A}$, then $x^R \notin \mathcal{A}$, we conclude that if x is a subword of $\mathbf{w}$ and $|x| \geq 2$, then x^R is not a subword of $\mathbf{w}$.

To see that $\mathbf{w}$ is unique up to permutation of the alphabet symbols, consider another word $\mathbf{w}'$ satisfying the conditions of the theorem, and suppose that $\mathbf{w}'$ begins with 01. Then 01 must be followed by 2, 12 must be followed by 0, and 20 must be followed by 1. Hence,

$$\mathbf{w}' = (012)^\omega = 012012012012\cdots = \mathbf{w}.$$

□

Note that the solution given in the proof of Theorem 1 is periodic. In the following theorem, we give a nonperiodic solution to this problem for $|x| \geq 3$.

Theorem 2. *There exists an infinite nonperiodic word* $\mathbf{w}$ *over* Σ_3 *such that if* x *is a subword of* $\mathbf{w}$ *and* $|x| \geq 3$*, then* x^R *is not a subword of* $\mathbf{w}$*.*

Proof. By reasoning similar to that given in the proof of Theorem 1, it suffices to show the theorem for $|x| = 3$. Let $\mathbf{w}'$ be an infinite nonperiodic word over Σ_2. For example, if $\mathbf{w}' = 11010010001\cdots$, then $\mathbf{w}'$ is nonperiodic. Define the morphism $h : \Sigma_2^\omega \to \Sigma_3^\omega$ by

$$\begin{aligned} 0 &\to 0012 \\ 1 &\to 0112. \end{aligned}$$

Then $\mathbf{w} = h(\mathbf{w}')$ has the desired property. Consider the set $\mathcal{A}$ consisting of all subwords of $\mathbf{w}$ of length three. We have

$$\mathcal{A} = \{001, 011, 012, 112, 120, 200, 201\}.$$

Noting that if $x \in \mathcal{A}$, then $x^R \notin \mathcal{A}$, we conclude that if x is a subword of $\mathbf{w}$ and $|x| \geq 3$, then x^R is not a subword of $\mathbf{w}$.

To see that $\mathbf{w}$ is not periodic, suppose the contrary; *i.e.*, suppose that $\mathbf{w} = y^\omega$ for some $y \in \Sigma_3^*$. Clearly, $|y| > 4$. Suppose then that y begins with $h(0)$. Noting that the only way to obtain 00 from $h(ab)$, where $a, b \in \Sigma_2$, is as a prefix of $h(0)$, we see that $y = h(y')$ for some $y' \in \Sigma_2^*$. Hence, $\mathbf{w} = (h(y'))^\omega = h((y')^\omega)$, and so $\mathbf{w}' = (y')^\omega$ is periodic, contrary to our choice of $\mathbf{w}'$. □

Over a two-letter alphabet we have the following negative result.

Theorem 3. *Let $k \leq 4$ and let w be a word over Σ_2 such that if x is a subword of w and $|x| \geq k$, then x^R is not a subword of w. Then $|w| \leq 8$.*

Proof. As mentioned previously, if $k = 1$ the result holds trivially. If $k = 2$, note that all binary words of length at least three must contain one of the following words: 00, 11, 010, or 101. Similarly, if $k = 3$, note that all binary words of length at least five must contain one of the following words: 000, 010, 101, 111, 0110, or 1001; and if $k = 4$, note that all binary words of length at least nine must contain one of the following words: 0000, 0110, 1001, 1111, 00100, 01010, 01110, 10001, 10101, or 11011. Hence, $|w| \leq 8$, as required. □

For $|x| \geq 5$, however, we find that there *are* infinite words with the desired property.

Theorem 4. *There exists an infinite word $\mathbf{w}$ over Σ_2 such that if x is a subword of $\mathbf{w}$ and $|x| \geq 5$, then x^R is not a subword of $\mathbf{w}$.*

Proof. By reasoning similar to that given in the proof of Theorem 1, it suffices to show the theorem for $|x| = 5$. We show that the infinite word

$$\mathbf{w} = (001011)^\omega = 001011001011001011\cdots$$

has the desired property. To see this, consider the set $\mathcal{A}$ consisting of all subwords of $\mathbf{w}$ of length five. We have

$$\mathcal{A} = \{00101, 01011, 01100, 10010, 10110, 11001\}.$$

Noting that if $x \in \mathcal{A}$, then $x^R \notin \mathcal{A}$, we conclude that if x is a subword of $\mathbf{w}$ and $|x| \geq 5$, then x^R is not a subword of $\mathbf{w}$. □

Let z be the word 001011. We denote the *complement* of z by $\bar{z}$, *i.e.*, the word obtained by substituting 0 for 1 and 1 for 0 in z. Let $\mathcal{B}$ be the set defined as follows:

$$\mathcal{B} = \{x \mid x \text{ is a cyclic shift of } z \text{ or } \bar{z}\}.$$

We have the following characterization of the words satisfying the conditions of Theorem 4.

Theorem 5. *Let* $\mathbf{w}$ *be an infinite word over* Σ_2 *such that if* x *is a subword of* $\mathbf{w}$ *and* $|x| \geq 5$*, then* x^R *is not a subword of* $\mathbf{w}$*. Then* $\mathbf{w}$ *is ultimately periodic. Specifically,* $\mathbf{w}$ *is of the form* $y'y^\omega$*, where* $y' \in \{\epsilon, 0, 1, 00, 11\}$ *and* $y \in \mathcal{B}$.

Proof. By reasoning similar to that given in the proof of Theorem 1, it suffices to show the theorem for $|x| = 5$. We call a word $w \in \Sigma_2^*$ *valid* if w satisfies the property that if x is a subword of w and $|x| = 5$, then x^R is not a subword of w. We have the following two facts, which may be verified computationally.

1. All 32 valid words of length 9 are of the form $y'yy''$, where $y' \in \{\epsilon, 0, 1, 00, 11\}$, $y \in \mathcal{B}$, and $y'' \in \Sigma_2^*$.

2. Let w be one of the 20 valid words of the form yy'', where $y \in \mathcal{B}$, $y'' \in \Sigma_2^*$, and $|y''| = 9$. Then y is a prefix of y''.

We will prove by induction on n that for all $n \geq 1$, $y'y^n$ is a prefix of $\mathbf{w}$, where $y' \in \{\epsilon, 0, 1, 00, 11\}$ and $y \in \mathcal{B}$.

If $n = 1$, then by applying the first fact to the prefix of $\mathbf{w}$ of length 9, we have that $y'y$ is a prefix of $\mathbf{w}$, as required.

Assume then that $y'y^n$ is a prefix of $\mathbf{w}$. We can thus write $\mathbf{w} = y'y^{n-1}y\mathbf{w}'$, for some $\mathbf{w}' \in \Sigma_2^\omega$. By applying the second fact to the prefix of $y\mathbf{w}'$ of length 15, we have that y is a prefix of $\mathbf{w}'$. Hence $\mathbf{w} = y'y^{n-1}yy\mathbf{w}'' = y'y^{n+1}\mathbf{w}''$, for some $\mathbf{w}'' \in \Sigma_2^\omega$, as required.

We therefore conclude that if $\mathbf{w}$ satisfies the conditions of the theorem, then $\mathbf{w}$ is of the form $y'y^\omega$, where $y' \in \{\epsilon, 0, 1, 00, 11\}$ and $y \in \mathcal{B}$. □

Next we give a nonperiodic solution to this problem for $|x| \geq 6$.

Theorem 6. *There exists an infinite nonperiodic word* $\mathbf{w}$ *over* Σ_2 *such that if* x *is a subword of* $\mathbf{w}$ *and* $|x| \geq 6$*, then* x^R *is not a subword of* $\mathbf{w}$.

Proof. By reasoning similar to that given in the proof of Theorem 1, it suffices to show the theorem for $|x| = 6$. Let $\mathbf{w}'$ be an infinite nonperiodic word over Σ_2. Define the morphism $h : \Sigma_2^\omega \to \Sigma_2^\omega$ by

$$\begin{aligned} 0 &\to 0001011 \\ 1 &\to 0010111. \end{aligned}$$

We show that the infinite word $\mathbf{w} = h(\mathbf{w}')$ has the desired property. To see this, consider the set $\mathcal{A}$ consisting of all subwords of $\mathbf{w}$ of length six. We have

$$\mathcal{A} = \{000101, 001011, 010110, 010111, 011000, 011001, 011100, 100010, 100101, 101100, 101110, 110001, 110010, 111000, 111001\}.$$

Noting that if $x \in \mathcal{A}$, then $x^R \notin \mathcal{A}$, we conclude that if x is a subword of $\mathbf{w}$ and $|x| \geq 6$, then x^R is not a subword of $\mathbf{w}$.

To see that $\mathbf{w}$ is not periodic, suppose the contrary; *i.e.*, suppose that $\mathbf{w} = y^\omega$ for some $y \in \Sigma_2^*$. Clearly, $|y| > 7$. Suppose then that y begins with $h(0)$. Noting that the only way to obtain 000 from $h(ab)$, where $a, b \in \Sigma_2$, is as a prefix of $h(0)$, we see that $y = h(y')$ for some $y' \in \Sigma_2^*$. Hence, $\mathbf{w} = (h(y'))^\omega = h((y')^\omega)$, and so $\mathbf{w}' = (y')^\omega$ is periodic, contrary to our choice of $\mathbf{w}'$. □

Finally we consider words avoiding squares as well as reversed subwords. It is easy to check that no binary word of length ≥ 4 avoids squares. However, Thue [7] gave an example of a infinite squarefree ternary word. Over a four-letter alphabet we have the following negative result, which may be verified computationally.

Theorem 7. *Let w be a squarefree word over Σ_4 such that if x is a subword of w and $|x| \geq 2$, then x^R is not a subword of w. Then $|w| \leq 20$.*

In contrast with the result of Theorem 7, Alon *et al.* [1] have noted that over a four-letter alphabet there exists an infinite squarefree word that avoids palindromes x where $|x| \geq 2$. (A *palindrome* is a word x such that $x = x^R$.) However, over a five-letter alphabet there are infinite words with an even stronger avoidance property.

Theorem 8. *There exists an infinite squarefree word $\mathbf{w}$ over Σ_5 such that if x is a subword of $\mathbf{w}$ and $|x| \geq 2$, then x^R is not a subword of $\mathbf{w}$.*

Proof. By reasoning similar to that given in the proof of Theorem 1, it suffices to show the theorem for $|x| = 2$. Let $\mathbf{w}'$ be an infinite squarefree word over Σ_3. Define the morphism $h : \Sigma_3^\omega \to \Sigma_5^\omega$ by

$$\begin{aligned} 0 &\to 012 \\ 1 &\to 013 \\ 2 &\to 014. \end{aligned}$$

We show that the infinite word $\mathbf{w} = h(\mathbf{w}')$ has the desired property.

First we note that to verify that $\mathbf{w}$ is squarefree, it suffices by a theorem of Thue [8] (see also [2, 3, 4]) to verify that $h(w)$ is squarefree for all 12 squarefree words $w \in \Sigma_3^*$ such that $|w| = 3$. This is left to the reader.

To see that if x is a subword of $\mathbf{w}$ and $|x| = 2$, then x^R is not a subword of $\mathbf{w}$, consider the set $\mathcal{A}$ consisting of all subwords of $\mathbf{w}$ of length 2. We have

$$\mathcal{A} = \{01, 12, 13, 14, 20, 30, 40\}.$$

Noting that if $x \in \mathcal{A}$, then $x^R \notin \mathcal{A}$, we conclude that if x is a subword of $\mathbf{w}$ and $|x| \geq 2$, then x^R is not a subword of $\mathbf{w}$. □

Finally, we consider a slight variation of the original problem; that is, we examine words w that have the property that if x and x^R are both subwords of w, then $x = x^R$. Over a two letter alphabet, all such words w are of the form $0\cdots 0$, $1\cdots 1$, $0\cdots 01\cdots 1$, or $1\cdots 10\cdots 0$. Over a three letter alphabet, we have the following characterization.

Theorem 9. *There are $2^n - 1$ words $w \in \Sigma_3^*$ of length n that begin with 0 and have the property that if x and x^R are both subwords of w, then $x = x^R$.*

Proof. Any word w satisfying the conditions of the theorem is either of the form $0\cdots 0$, or begins with $0\cdots 01$ or $0\cdots 02$. Supose that w begins with $0\cdots 01$ (the case where w begins with $0\cdots 02$ is similar). Then $0\cdots 01$ cannot be followed by a 0, as then 01 and 10 would both be subwords of w. Extending this reasoning, we find that w must be a prefix of a word of the form

$$(0\cdots 01\cdots 12\cdots 2)(0\cdots 01\cdots 12\cdots 2)\cdots$$

(here the parentheses are not part of the word but just serve to group repeating blocks).

We see then that the language $\mathcal{L}$ of all words satisfying the conditions of the theorem can be described by the following regular expression (see [5] for more on regular expressions):

$$\mathcal{L} = (00^*11^*22^*)^*(0^* + 00^*1^*) + (00^*22^*11^*)^*(0^* + 00^*2^*).$$

The minimal (incomplete) deterministic finite automaton (again, see [5] for more on finite automata) M that accepts $\mathcal{L}$ has eight states and is given by

$$M = (\{q_1, \ldots, q_8\}, \Sigma_3, \delta, q_1, \{q_1, \ldots, q_8\}).$$

Note that all states are final. We omit the precise specification of the transition function δ and instead consider the adjacency matrix $A = (a_{ij})$, where the entries a_{ij} give the number of transitions from state q_i to state q_j. We have

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The (i, j) entry of A^n gives the number of paths of length n from state q_i to state q_j. The number of words of length n accepted by M is thus given by the sum of the values of the first row of A^n (since all states are final). An easy induction shows that

$$A^n \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2^n - 1 \\ 2^{n+1} - 1 \\ 2^n \\ 2^n \\ 2^n \\ 2^n \\ 2^n \\ 2^n \end{bmatrix} \quad \text{for } n \geq 1,$$

from which we see that $\mathcal{L}$ contains $2^n - 1$ words of length n. □

References

[1] N. Alon, J. Grytczuk, M. Haluszczak, O. Riordan, "Nonrepetitive colorings of graphs", *Random Structures and Algorithms* **21** (2002), 336–346.

[2] D. Bean, A. Ehrenfeucht, G. McNulty, "Avoidable patterns in strings of symbols", *Pacific J. Math.* **85** (1979), 261–294.

[3] J. Berstel, "Sur les mots sans carré définis par un morphisme", *Automata, languages and programming (Sixth Colloq., Graz, 1979)*, pp. 16–25, Lecture Notes in Comput. Sci. **71**, Springer, Berlin-New York, 1979.

[4] M. Crochemore, "Sharp characterizations of squarefree morphisms", *Theoret. Comput. Sci.* **18** (1982), 221–226.

[5] J. E. Hopcroft, J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, 1979.

[6] A. Szilard, personal communication, 2003.

[7] A. Thue, "Über unendliche Zeichenreihen", *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl.* 7 (1906), 1–22.

[8] A. Thue, "Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen", *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl.* 1 (1912), 1–67.