

Two Element Unavoidable Sets of Partial Words* (Extended Abstract)

F. Blanchet-Sadri¹, N.C. Brownstein², and Justin Palumbo³

¹ Department of Computer Science, University of North Carolina,
P.O. Box 26170, Greensboro, NC 27402–6170, USA

² Department of Mathematics, University of Central Florida,
P.O. Box 161364, Orlando, FL 32816–1364, USA

³ Department of Mathematics,
Rutgers The State University of New Jersey,
110 Frelinghuysen Road, Piscataway, NJ 08854–8019, USA

Abstract. The notion of an unavoidable set of words appears frequently in the fields of mathematics and theoretical computer science, in particular with its connection to the study of combinatorics on words. The theory of unavoidable sets has seen extensive study over the past twenty years. In this paper we extend the definition of unavoidable sets of words to unavoidable sets of partial words. Partial words, or finite sequences that may contain a number of “do not know” symbols or holes, appear in natural ways in several areas of current interest such as molecular biology, data communication, DNA computing, etc. We demonstrate the utility of the notion of unavoidability on partial words by making use of it to identify several new classes of unavoidable sets of full words. Along the way we begin work on classifying the unavoidable sets of partial words of small cardinality. We pose a conjecture, and show that affirmative proof of this conjecture gives a sufficient condition for classifying all the unavoidable sets of partial words of size two. Lastly we give a result which makes the conjecture easy to verify for a significant number of cases.

1 Introduction

An *unavoidable* set of words X over an alphabet A is a set for which any sufficiently long word over A will have a factor in X . It is clear from the definition that from each unavoidable set we can extract a finite unavoidable subset, so the study can be reduced to finite unavoidable sets. This concept was explicitly introduced in 1983 in connection with an attempt to characterize the rational languages among the context-free ones [8]. Since then it has been consistently

* This material is based upon work supported by the National Science Foundation under Grant No. DMS–0452020. A World Wide Web server interface at www.uncg.edu/mat/research/unavoidablesets has been established for automated use of the program. We thank the referees of preliminary versions of this paper for their very valuable comments and suggestions.

studied by researchers in both mathematics and theoretical computer science. Testing the unavoidability of X can be done in different ways [7]: Check whether there is a loop in the finite automaton of Aho and Corasick [1] recognizing $A^* \setminus A^*XA^*$, or simplify X as much as possible. There is a large literature on unavoidable sets of words and we refer the reader to [6,14] for more information.

Another concept relevant to this paper is that of a *partial word*, or a finite sequence of symbols from a finite alphabet that may contain a number of “do not know” symbols or “holes”. Partial words appear in natural ways in several areas of current interest such as molecular biology, data communication, and DNA computing. In this paper, we introduce unavoidable sets of partial words. In terms of unavoidability, sets of partial words serve as efficient representations of sets of full words. This is strongly analogous to the study of unavoidable patterns, in which sets of patterns are used to represent infinite sets of full words [13]. The main goal here is to demonstrate that the study of unavoidable sets of partial words leads to new insights both on the theory of unavoidable sets and on the combinatorial structure of the set of words A^* as a whole. In accomplishing this we mainly focus on the problem of classifying the unavoidable sets of size two.

The contents of our paper are summarized as follows. In Section 2, we review some basic definitions related to words and partial words. In Section 3, we recall the definition of unavoidable sets of words and some useful elementary properties. There, we present our definition of unavoidable sets of partial words and we introduce the problem of classifying such sets of small cardinality and in particular those with two elements, x_1, x_2 , with respect to the regular constraints: x_1 matches the pattern $(a\diamond^*)^*a$ and x_2 the pattern $(b\diamond^*)^*b$ where \diamond denotes the “do not know” symbol and a, b denote distinct letters of the alphabet. In Section 4, we give an elegant characterization of the particular case of this problem when x_1 matches $a\diamond^*a$ and x_2 matches $b\diamond^*b$, propose a conjecture characterizing the case where x_1 matches $a\diamond^*a$ and x_2 matches $b\diamond^*b\diamond^*b$, and prove that verifying this conjecture is sufficient for solving the problem in general. There, we also prove one direction of our conjecture. In Section 5, we give partial results towards the other direction of our conjecture and in particular prove that it is easy to verify in a large number of cases. Finally in Section 6, we pose several natural and interesting questions related to unavoidable sets of partial words.

2 Preliminaries

We begin this section with the following basic terms and definitions.

Throughout this paper A is a fixed finite set called the *alphabet* whose elements we call *letters*. We use A^* to denote the set of words over A , that is the set of finite sequences of letters in the alphabet. For $u \in A^*$ we write $|u|$ for the length of u . Under the concatenation operation of words, A^* forms a free monoid whose identity is the empty word which we denote by ε . If there exist $x, y \in A^*$ such that $u = xvy$ then we say that v is a *factor* of u .

A *two-sided infinite word* w is a function $w : \mathbb{Z} \rightarrow A$. A finite word u is a factor of the two-sided infinite word w if u is a finite subsequence of w , that is if

there exists some $i \in \mathbb{Z}$ so that $u = w(i+1) \dots w(i+|u|)$. For a positive integer p , we say that w has *period* p if $w(i) = w(j)$ for all $i, j \in \mathbb{Z}$ satisfying $i \equiv j \pmod{p}$. If w has period p for some p then we call w *periodic*. We can now define infinite powers of a word: if v is a nonempty finite word, then we denote by $v^{\mathbb{Z}}$ the unique two-sided infinite word w such that w has period $|v|$ and $w(0) \dots w(|v|-1) = v$.

A word of finite length n over an alphabet A can be defined as a total function $w : \{0, \dots, n-1\} \rightarrow A$. Analogously a *partial word* (or, *pword*) of length n over A is a partial function $u : \{0, \dots, n-1\} \rightarrow A$. For $0 \leq i < n$, if $u(i)$ is defined, then we say that i belongs to the domain of u (denoted by $i \in D(u)$). Otherwise we say that i belongs to the *set of holes* of u (denoted by $i \in H(u)$). In cases where $H(u)$ is empty we say that u is a *full word*.

Let $A_{\diamond} = A \cup \{\diamond\}$. If u is a partial word of length n over A , then the *companion* of u is the total function $u_{\diamond} : \{0, \dots, n-1\} \rightarrow A_{\diamond}$ defined by

$$u_{\diamond} = \begin{cases} u(i) & \text{if } i \in D(u) \\ \diamond & \text{otherwise} \end{cases}$$

Throughout this paper we identify a partial word with its companion. We reserve the term *letter* for members of A . We will refer to an occurrence of the symbol \diamond in a partial word as a *hole*.

Two partial words u and v of equal length are said to be *compatible*, denoted by $u \uparrow v$, if $u(i) = v(i)$ for every $i \in D(u) \cap D(v)$. If X is a set of partial words, we use \hat{X} to denote the set of all full words compatible with a member of X .

3 Unavoidable sets

We first recall the definition of an unavoidable set of full words and some relevant properties. Let $X \subseteq A^*$. A two-sided infinite word w *avoids* X if no factor of w is a member of X . We say that X is *unavoidable* if no two-sided infinite word avoids X . In other words X is unavoidable if every two-sided infinite word has a factor in X .

Following are two useful facts giving alternative characterizations of unavoidable sets: (1) The set $X \subseteq A^*$ is unavoidable if and only if there are only finitely many words in A^* with no member of X as a factor; and (2) If the set $X \subseteq A^*$ is finite, then X is unavoidable if and only if no periodic two-sided infinite word avoids it. Proofs can be found in [13].

We now give our extension of the definition of unavoidable sets of words to unavoidable sets of partial words.

Definition 1. Let $X \subseteq A_{\diamond}^*$. A two-sided infinite word w *avoids* X if no factor of w is a member of \hat{X} . We say that X is *unavoidable* if no two-sided infinite word avoids X . In other words X is unavoidable if every two-sided infinite word has a factor compatible with a member of X .

There is a simple connection between sets of partial words and sets of full words that is worth noting. By the definition of \hat{X} , w has a factor in \hat{X} if and only

if that same factor is compatible with a member of X . Thus the two-sided infinite words which avoid X are exactly those which avoid \hat{X} , and X is unavoidable if and only if \hat{X} is unavoidable.

With regards to unavoidability X is then essentially a representation of a set of full words. This representation makes possible new approaches to unavoidable sets of full words. It is easier to consider the two-sided infinite words avoiding $X = \{aa, b\diamond^3b\}$ as those without an occurrence of aa and no two occurrences of b separated by three letters rather as the words avoiding

$$\hat{X} = \{aa, baaab, baabb, babab, babbb, bbaab, bbabb, bbbab, bbbbb\}.$$

It is most natural to look first for the unavoidable sets of partial words that have small cardinality. Insight into the structure of A^* can be gained by identifying an unavoidable set, especially if that set contains few elements.

Any set of partial words containing the empty word or \diamond^n for some $n \in \mathbb{N}$ will be called a trivial unavoidable set. To find nontrivial unavoidable sets of size 2 we may assume that $A = \{a, b\}$. Classifying the unavoidable sets of size 2 is a daunting task and is the focus of this paper.

Say $X = \{x_1, x_2\}$ is unavoidable. As mentioned before if X is nontrivial it must be that one member of X is compatible with a power of a and the other is compatible with a power of b , as that is the only way to guarantee that both $a^{\mathbb{Z}}$ and $b^{\mathbb{Z}}$ will not avoid X . So in order to classify the unavoidable sets of size 2, it is sufficient to determine for which m_1, m_2, \dots, m_k and n_1, n_2, \dots, n_l the set

$$X_{m_1, \dots, m_k | n_1, \dots, n_l} = \{a\diamond^{m_1}a \dots a\diamond^{m_k}a, b\diamond^{n_1}b \dots b\diamond^{n_l}b\}$$

is unavoidable. We can in fact simplify the situation a little further. The following lemma tells us that it is enough to solve the problem for cases where $m_1 + 1, m_2 + 1, \dots, m_k + 1$ and $n_1 + 1, n_2 + 1, \dots, n_l + 1$ are relatively prime.

Lemma 1. *Let $p \in \mathbb{N}$. The set $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is unavoidable if and only if the set*

$$Y = \{a\diamond^{p(m_1+1)-1}a \dots a\diamond^{p(m_k+1)-1}a, b\diamond^{p(n_1+1)-1}b \dots b\diamond^{p(n_l+1)-1}b\}$$

is unavoidable.

Proof. In terms of notation it will be helpful to define

$$M_j = \sum_{i=1}^j m_i + 1$$

Now suppose the two-sided infinite word w avoids $X_{m_1, \dots, m_k | n_1, \dots, n_l}$, and let

$$v = \dots w(-1)^p w(0)^p w(1)^p \dots$$

We claim that v avoids Y . Suppose otherwise. Then v has a factor compatible with some $x \in Y$. Without loss of generality say that

$$x = a \diamond^{p(m_1+1)-1} a \dots \diamond^{p(m_k+1)-1} a$$

Then to say that v has a factor compatible with x is equivalent to saying that there exists $i \in \mathbb{Z}$ for which

$$v(i) = v(i + pM_1) = \dots = v(i + pM_k) = a$$

But if we set $h = \lfloor \frac{i}{p} \rfloor$ then this implies that

$$w(h) = w(h + M_1) = \dots = w(h + M_k) = a$$

contradicting the fact that w avoids $X_{m_1, \dots, m_k | n_1, \dots, n_l}$.

We prove the other direction analogously. Suppose now that the two-sided infinite word w avoids Y , and set $v = \dots w(-p)w(0)w(p) \dots$. We claim that v avoids $X_{m_1, \dots, m_k | n_1, \dots, n_l}$. Otherwise v has a factor compatible with some $x \in p$ which we may suppose without loss of generality is $a \diamond^{m_1} a \dots \diamond^{m_k} a$. Then there exists $i \in \mathbb{Z}$ for which

$$v(i) = v(i + M_1) = \dots = v(i + M_k)$$

but this implies that

$$w(pi) = w(pi + pM_1) = \dots = w(pi + pM_k)$$

which contradicts the fact that w avoids Y .

In order to solve the problem of identifying when $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is unavoidable we start with small values of k and l . The set $\{a, b \diamond^{n_1} b \dots b \diamond^{n_l} b\}$ is unavoidable for if w is a two-sided infinite word which lacks a factor compatible with a it must be $b^{\mathbb{Z}}$. This handles the case where $k = 0$ (and symmetrically $l = 0$).

4 Special cases

We first consider the case where $k = 1$ and $l = 1$, that is, we consider the set $X_{m|n} = \{a \diamond^m a, b \diamond^n b\}$. In this case, we can give a nice characterization of which integers m, n make this set avoidable.

Theorem 1. *Write $m + 1 = 2^s r_0$, $n + 1 = 2^t r_1$ where r_0, r_1 are odd. Then $X_{m|n} = \{a \diamond^m a, b \diamond^n b\}$ is avoidable if and only if $s = t$.*

Proof. Let w be a two-sided infinite word avoiding $X_{m|n}$. Then w also avoids $b \diamond^m b$. Otherwise for some $i \in \mathbb{Z}$, $w(i) = b$ and $w(i + m + 1) = b$. Since w avoids $b \diamond^n b$ we must have that $w(i + n + 1) = a$ and $w(i + m + 1 + n + 1) = a$, which contradicts the fact that w avoids $a \diamond^m a$. A symmetrical argument shows that w avoids $a \diamond^n a$.

For ease of notation, write $\bar{a} = b$ and $\bar{b} = a$. Let $p \in \mathbb{N}$. We will say that a two-sided infinite word is p -alternating if for all $i \in \mathbb{Z}$, $w(i) = w(i + p)$. By

our previous observation it is easy to see that w avoids $X_{m|n}$ if and only if w is $m + 1$ -alternating and $n + 1$ -alternating. Thus to prove the theorem it is sufficient to show that a two-sided infinite word exists which is p -alternating and q -alternating if and only if $s = t$ where $p = 2^s r_0$ and $q = 2^t r_1$ with r_0 and r_1 odd.

Notice that if w is p -alternating then it has period $2p$: for $i \in \mathbb{Z}$,

$$w(i) = \overline{w(i+p)} = \overline{\overline{w(i+2p)}} = w(i+2p)$$

Now suppose $s \neq t$. Without loss of generality say $s < t$. Then $s+1 \leq t$. Let l be the least common multiple of p and q . The prime factorization of l must have no greater power of 2 than the prime factorization of q . Thus there exists an odd number k such that $kq \equiv 0 \pmod{2p}$. If there were a two-sided infinite word w which was p -alternating and q -alternating we would have $w(0) = w(2p) = w(kq)$ since w has period $2p$. But since k is odd and w is q -alternating we also have $w(0) = \overline{w(kq)}$. This is a contradiction. We have half of the necessary implication.

Now suppose $s = t$. Then $p = 2^s r_0$, $q = 2^s r_1$. We only need to prove that there exists some w which is p -alternating and q -alternating and we do this by induction on s . If $s = 0$, then p and q are odd. Then the word $\dots ababab \dots$ is p -alternating and q -alternating. This handles our base case. Now say w is $2^s r_0$ and $2^s r_1$ -alternating. Then $v = \dots w(-1)w(-1)w(0)w(0)w(1)w(1) \dots$ is $2^{s+1} r_0$ and $2^{s+1} r_1$ -alternating. This finishes the induction and our proof.

We next consider the case where $k = 1$ and $l = 2$, that is, sets of the form $X_{m|n_1, n_2} = \{a \diamond^m a, b \diamond^{n_1} b \diamond^{n_2} b\}$. We believe, based on extensive experimental evidence, that we have identified the cases for which $X_{m|n_1, n_2}$ is unavoidable which we state in this section (Conjecture 1). As a result of this conjecture, $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is avoidable for all larger k, l . Here we prove one direction of our conjecture, and in Section 5, we give partial results towards the other direction which turns out to be easy for even values of m .

There is a delicate tension in the change of difficulty of the problem as we increase k and l . On the one hand, we have identified a large number of avoidable sets of the form $\{a \diamond^m a, b \diamond^n b\}$. For $X_{m|n_1, n_2}$ to be avoidable it is sufficient that $\{a \diamond^m a, b \diamond^{n_1} b\}$, $\{a \diamond^m a, b \diamond^{n_2} b\}$ or $\{a \diamond^m a, b \diamond^{n_1+n_2+1} b\}$ be avoidable. Thus by first identifying the avoidable sets for smaller values of k and l our job has gotten a little easier. On the other hand the structure of words avoiding $\{a \diamond^m a, b \diamond^{n_1} b \diamond^{n_2} b\}$ is not nearly as nice as those avoiding $\{a \diamond^m a, b \diamond^n b\}$. There is no simple characterization akin to p -alternation.

In proving that a set of the form $X_{m|n_1, n_2}$ is unavoidable our strategy is to derive a contradiction using structural properties that any potential two-sided infinite word w avoiding X would have. These properties take the form of certain rules involving the occurrences of letters in w . For example, whenever $w(i) = w(i+n_1+1) = b$ in w , we must have that $w(i+n_1+n_2+2) = a$. The presence of an a also has implications: if $w(i) = a$ then $w(i-m-1) = b$ and $w(i+m+1) = b$. Often particular values of m, n_1 and n_2 have a relationship that cause these patterns to reoccur and perpetuate themselves, making a contradiction easy to

find. In order for this to happen we also need a starting point for the perpetuation the ground. For this Theorem 1 is a very handy tool.

We give an example of this in action. The set $\{a\diamond^7a, b\circ b\circ^3b\}$ is unavoidable. Suppose instead that there exists a two-sided infinite word w which avoids it. We know from Theorem 1 that $\{a\circ^7a, b\circ b\}$ is unavoidable, thus w must have a factor compatible with $b\circ b$. Say without loss of generality that $w(0) = w(2) = b$. This implies that $w(6) = a$, which in turn implies that $w(-2) = b$. Then we have that $w(-2) = w(0) = b$, forcing $w(4) = a$. This propagation continues: $w(-4) = w(-2) = b$ and so $w(2) = a$, which makes $w(-6) = b$ giving $w(0) = a$, a contradiction. This example is part of a more general phenomenon. Notice how in this example as the patterns reoccur, we have a sequence of a 's traveling to the left toward the b at $w(0)$. There is a symmetric situation in which the b 's travel to the right towards the a at $w(n_1 + 1)$. Both scenarios are covered by the following proposition.

Proposition 1. *Suppose either $m = 2n_1 + n_2 + 2$ or $m = n_2 - n_1 - 1$, and $n_1 + 1$ divides $n_2 + 1$. Then $X_{m|n_1, n_2}$ is unavoidable if and only if $\{a\circ^m a, b\circ^{n_1} b\}$ is unavoidable.*

One notable consequence of Proposition 1 is that if m is odd, then both $\{a\circ^m a, bb\circ^{m+1}b\}$ and $\{a\circ^m a, bb\circ^{m-2}b\}$ are unavoidable.

The next theorem takes advantage of the perpetuating pattern phenomenon in a more complicated context. Proposition 1 held because each a forced a b into the next position of an occurrence of $w(i) = w(i+n_1+1) = b$, which in turn forced a new a in w . This created a single traveling sequence of a 's and b 's, causing an a to overlap with the b at $w(0)$, yielding a contradiction. In the next argument, we take notice of the fact that each a occurring in w may contribute to two occurrences of $w(i) = w(i+n_1+1) = b$ simultaneously so that a contradiction will occur after many traveling sequences of letters appear and overlap.

Theorem 2. *Say that $m = n_2 - n_1 - 1$ or $m = 2n_1 + n_2 + 2$, and that the highest power of 2 dividing $n_1 + 1$ is less than the highest power of 2 dividing $m + 1$. Then $X_{m|n_1, n_2}$ is unavoidable.*

Proof. Since the highest power of 2 dividing $n_1 + 1$ is different than the highest power of 2 dividing $m + 1$, we have that the set $Y = \{a\circ^m a, b\circ^{n_1} b\}$ is unavoidable. Consider first the case where $m = n_2 - n_1 - 1$ and suppose for contradiction that there exists a two-sided infinite word w that avoids X . Then w has no factor compatible with $\{a\circ^m a\}$, and so since Y is unavoidable it must have a factor compatible with $\{b\circ^{n_1} b\}$. Assume without loss of generality that $w(0) = b$ and $w(n_1 + 1) = b$.

We now generate an infinite table of facts about w . Two horizontally adjacent entries in the table will represent positions in w which are $n_1 + 1$ letters apart. Two vertically adjacent entries in the table will represent positions in w which are $m + 1 = n_2 - n_1$ letters apart. The two upper left entries of our table are $w(0) = b$ and $w(n_1 + 1) = b$, two facts we have already assumed. Since w avoids $X_{m|n_1, n_2}$ we have more information relevant to the table: two horizontally

adjacent b entries force an a entry diagonally down and to the right from them, and an a entry forces a b entry in the vertically adjacent positions. From these rules we can build the following table, labeling the columns C_0, C_1, \dots :

$$\begin{array}{ccccccc}
C_0 & C_1 & C_2 & C_3 & \dots & & \\
w(0) = b & w(n_1 + 1) = b & w(2n_1 + 2) = b & w(3n_1 + 3) = b & & & \\
& & w(n_1 + n_2 + 2) = a & w(2n_1 + n_2 + 3) = a & & & \\
& & w(2n_2 + 2) = b & w(n_1 + 2n_2 + 3) = b & & &
\end{array}$$

For $i \in \mathbb{N}$, we shall define v_i to be the factor of w represented by C_i . If i is odd then C_i has i entries, and if i is even then C_i has $i + 1$ entries. Thus we define

$$v_i = \begin{cases} w(in_1 + i)w(in_1 + i + 1) \dots w(in_2 + i) & \text{if } i \text{ even} \\ w((i-1)n_1 + i)w((i-1)n_1 + i + 1) \dots w(n_1 + (i-1)n_2 + i) & \text{if } i \text{ odd} \end{cases}$$

Two adjacent entries in C_i represent a distance of $m + 1$ positions between letters in v_i . Thus for i even we have that $|v_i| = im + 1$ and for i odd we have that $|v_i| = (i - 1)m + 1$. We can also use the table to get some partial information about the positions of a 's and b 's in v_i . For $j \in \mathbb{N}$, $v_i(j) = b$ if $j \equiv 0 \pmod{2m + 2}$, and $v_i(j) = a$ if $j \equiv m + 1 \pmod{2m + 2}$.

Because the highest power of 2 dividing $n_1 + 1$ is no greater than the highest power of 2 dividing $m_1 + 1$, there exists some k for which $k(n_1 + 1) \equiv m + 1 \pmod{2m + 2}$. Take i sufficiently large so that $|v_i| > kn_1 + k$. Because of how k was chosen, we have that $v_i(kn_1 + k) = a$. However examining the table we see that

$$w((i + k)n_1 + i + k) = v_i(kn_1 + k) = v_{i+k}(0) = b$$

a contradiction. This handles the situation where $m = n_2 - n_1 - 1$. The proof for the case where $m = 2n_1 + n_2 + 2$ is similar, the only difference is that the table will represent increasingly negative positions of w , rather than increasingly positive ones.

As an application of Theorem 2, take $m = 1$. Let us see for which $n_1 \in \mathbb{N}$ the hypotheses of the theorem hold to make $X_{m|n_1, n_2}$ unavoidable. The highest power of 2 dividing $n_1 + 1$ should be less than the highest power of 2 dividing $m + 1 = 2$. Thus $n_1 + 1$ must be odd, n_1 is even. Since $m = 1$ we cannot have $m = 2n_1 + n_2 + 2$. Say we have $m = n_2 - n_1 - 1$. Then $n_2 = n_1 + 2$. So we have that for any even n_1 , the set $\{a \triangleright a, b \triangleright^n b \triangleright^{n+2} b\}$ is unavoidable. We will prove in Section 5 that this is a complete characterization of unavoidability of $X_{m|n_1, n_2}$ for $m = 1$.

The next proposition identifies another large class of unavoidable sets using a modification of the strategies discussed so far.

Proposition 2. *Suppose $n_1 < n_2$, $2m = n_1 + n_2$ and $|m - n_1|$ divides $m + 1$. Then $X_{m|n_1, n_2}$ is unavoidable.*

We believe that together Lemma 1, Proposition 1, Proposition 2, and Theorem 2 nearly give a complete characterization of when $X_{m|n_1, n_2}$ is unavoidable. Following is what we believe to be the only exception.

Proposition 3. *The set $X_{6|1,3} = \{a\diamond^6 a, b\circ b\circ^3 b\}$ is unavoidable.*

We now state our conjecture.

Conjecture 1. The set $X_{m|n_1, n_2}$ is unavoidable precisely when the hypotheses of at least one of Lemma 1, Proposition 1, Proposition 2, Proposition 3 or Theorem 2 hold. Restated, $X_{m|n_1, n_2}$ is unavoidable for relatively prime $m+1$, n_1+1 and n_2+1 with $n_1 \leq n_2$ if and only if one of the following conditions (or their symmetric equivalents) hold:

- Proposition 1: The case where the set $\{a\circ^m a, b\circ^{n_1} b\}$ is unavoidable, $m = 2n_1 + n_2 + 2$ or $m = n_2 - n_1 - 1$, and $n_1 + 1$ divides $n_2 + 1$.
- Theorem 2: The case where $m = n_2 - n_1 - 1$ or $m = 2n_1 + n_2 + 2$, and the highest power of 2 dividing $n_1 + 1$ is less than the highest power of 2 dividing $m + 1$.
- Proposition 2: The case where $n_1 < n_2$, $2m = n_1 + n_2$ and $|m - n_1|$ divides $m + 1$.
- Proposition 3: The case where $m = 6$, $n_1 = 1$ and $n_2 = 3$.

The reader may verify that for any fixed m the only one of the above conditions that contributes infinitely many unavoidable sets to $X_{m|n_1, n_2}$ is Theorem 2, and that this theorem never applies to even m . Thus the conjecture states that there are only finitely many values of m, n_1, n_2 with m fixed and even and $X_{m|n_1, n_2}$ unavoidable. We will prove in Section 5 that this is indeed the case.

Using Lemma 1 we may assume without loss of generality that $m+1, n_1+1, n_2+1$ are relatively prime. An important consequence of the conjecture is that in order for $X_{m|n_1, n_2}$ to be unavoidable it is necessary that either $m = 6$ and $n_1, n_2 = 1, 3$, or that one of the following equations hold:

$$m = 2n_1 + n_2 + 2 \tag{1}$$

$$m = 2n_2 + n_1 + 2 \tag{2}$$

$$m = n_1 - n_2 - 1 \tag{3}$$

$$m = n_2 - n_1 - 1 \tag{4}$$

$$2m = n_1 + n_2 \tag{5}$$

Using this fact we can show that an affirmative proof of the conjecture has a powerful consequence.

We end this section with the following proposition which implies that if Conjecture 1 is true then we have completely classified the unavoidable sets of size two.

Proposition 4. *If Conjecture 1 holds, then $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is avoidable for all $k \geq 2$ and $l \geq 3$.*

Proof. Assuming Conjecture 1 holds, it is enough to prove that both $X_{m_1, m_2 | n_1, n_2}$ and $X_{m | n_1, n_2, n_3}$ are avoidable for all m_1, m_2, n_1, n_2 . We handle the case of $X_{m_1, m_2 | n_1, n_2}$. Assume without loss of generality that m_1, m_2, n_1, n_2 are relatively prime. In order for this set to be unavoidable it is necessary that the sets $\{a \diamond^{m_1} a, b \diamond^{n_1} b \diamond^{n_2} b\}$, $\{a \diamond^{m_2} a, b \diamond^{n_2} b \diamond^{n_1} b\}$, $\{a \diamond^{m_1} a \diamond^{m_2} a, b \diamond^{n_1} b\}$ and the set $\{a \diamond^{m_1} a \diamond^{m_2} a, b \diamond^{n_2} b\}$ are unavoidable as well. For each of these sets, Conjecture 1 gives a necessary condition: either $m = 6$ and $n_1 = 1, n_2 = 3$ (or symmetrically $n_1 = 3, n_2 = 1$) or one of Equations 1, 2, 3, 4 or 5 must hold. Consider the following tables:

$m_1 = 2n_1 + n_2 + 2$	$m_2 = 2n_1 + n_2 + 2$
$m_1 = 2n_2 + n_1 + 2$	$m_2 = 2n_2 + n_1 + 2$
$m_1 = n_1 - n_2 - 1$	$m_2 = n_1 - n_2 - 1$
$m_1 = n_2 - n_1 - 1$	$m_2 = n_2 - n_1 - 1$
$m_1 = 6, n_1 = 1, n_2 = 3$	$m_2 = 6, n_1 = 1, n_2 = 3$
$m_1 = 6, n_2 = 1, n_1 = 3$	$m_2 = 6, n_2 = 1, n_1 = 3$
$2m_1 = n_1 + n_2$	$2m_2 = n_1 + n_2$

$n_1 = 2m_1 + m_2 + 2$	$n_2 = 2m_1 + m_2 + 2$
$n_1 = 2m_2 + m_1 + 2$	$n_2 = 2m_2 + m_1 + 2$
$n_1 = m_1 - m_2 - 1$	$n_2 = m_1 - m_2 - 1$
$n_1 = m_2 - m_1 - 1$	$n_2 = m_2 - m_1 - 1$
$n_1 = 6, m_1 = 1, m_2 = 3$	$n_2 = 6, m_1 = 1, m_2 = 3$
$n_1 = 6, m_2 = 1, m_1 = 3$	$n_2 = 6, m_2 = 1, m_1 = 3$
$2n_1 = m_1 + m_2$	$2n_2 = m_1 + m_2$

In order for $X_{m_1, m_2 | n_1, n_2}$ to be unavoidable it is necessary that at least one equation from each column be satisfied. It is easy to verify using a computer algebra system that this is impossible except in the case where the last equation in each column is satisfied. However in this case $m_1 = m_2 = n_1 = n_2$ and so by Theorem 1 the set is avoidable.

5 Avoidability results for $k = 1$ and $l = 2$

In order to prove the conjecture, only one direction remains. We must show that if none of the hypotheses of Lemma 1, Proposition 1, Proposition 2, Proposition 3 or Theorem 2 hold then $X_{m | n_1, n_2}$ is avoidable. In this section we give partial results towards this goal.

We have found that in general identifying sets of the form $X_{m | n_1, n_2}$ as avoidable tends to be a more difficult task than identifying them as unavoidable. In the case of unavoidability we needed only consider a single word then derive a contradiction from its necessary structural properties. To find a class of avoidable sets we must invent some general procedure for producing a two-sided infinite word which avoids each such set. This is precisely what we move towards in the following propositions in which we verify that the conjecture holds for certain values of m and n_1 .

It is easy to see that none of Equations 1, 2, 3, 4 or 5 are satisfied when $n_1, n_2 < m \leq n_1 + n_2 + 2$. Thus the conjecture for such values is that $X_{m|n_1, n_2}$ is avoidable. The following fact verifies that this is indeed the case.

Proposition 5. *If $n_1, n_2 < m < n_1 + n_2 + 2$ then $X_{m|n_1, n_2}$ is avoidable.*

The next proposition makes the conjecture easy to verify for even values of m .

Proposition 6. *Assume m is even and that $2m \leq n_1, n_2$. Then $X_{m|n_1, n_2}$ is avoidable.*

For any fixed even m there are then only finitely many values of n_1, n_2 which might be unavoidable. The reader may verify that this is consistent with the conjecture. The reader may also verify that the conjecture for $m = 0$ is that $X_{0|n_1, n_2}$ is always avoidable, and indeed this is given by Proposition 6. Similarly the conjecture for $m = 2$ is that $X_{2|n_1, n_2}$ is avoidable except for $n_1 = 1, n_2 = 3$ or $n_2 = 3, n_1 = 1$. It is easy to find avoiding two-sided infinite words for other values of n_1 and n_2 less than 5 when $m = 2$. By Proposition 6 this is all that is necessary to confirm the conjecture for $m = 2$. In this way we have been able to verify the conjecture for all even m up to very large values. The odd values of m seem to be much more difficult and will most likely require more sophisticated techniques. The following proposition gives our confirmation of the conjecture for $m = 1$.

Proposition 7. *The conjecture holds for $m = 1$; that is $X_{1|n_1, n_2}$ is unavoidable if and only if n_1 and n_2 are even numbers with $|n_1 - n_2| = 2$.*

The following and final proposition says that if m and n_1 are close enough in value then $X_{m|n_1, n_2}$ is avoidable for large enough n_2 .

Proposition 8. *Let $s \in \mathbb{N}$ with $s < m - 2$. Then for $n > 2(m + 1)^2 + m - 1$, $X_{m|m+s, n} = \{a \diamond^m a, b \diamond^{m+s} b \diamond^n b\}$ is avoidable.*

6 Open questions

Conjecture 1, although tested in numerous cases via computer, and verified for $m = 1$ and a large number of even values of m , still remains to be proven. As was shown in Section 4, an affirmative answer to this question would imply that $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is avoidable for all $k, l \geq 3$. Given that avoidable sets of the form $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ for small k and l translate directly to avoidable sets for larger k and l , it might seem intuitive that for some sufficiently large fixed k and l there exists an easy proof that $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is always avoidable, and thus all larger values are. This is a deceptively difficult question. There is an interesting tension occurring between the increase in avoidability of $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ and the structural complication of $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ as k and l increase.

We pose two open questions that propose direction for further research.

Open question 1 *Can one find some sufficiently large values of k and l for which it is easy to prove that $X_{m_1, \dots, m_k | n_1, \dots, n_l}$ is always avoidable?*

Efficient algorithms to determine if a finite set of full words is unavoidable are well known, see for example [6]. These same algorithms can be used to decide if a finite set of partial words X is unavoidable by determining the unavoidability of \hat{X} . However this incurs a dramatic loss in efficiency, as each pword u in X can contribute as many as $\|A\|^{\|H(u)\|}$ elements to \hat{X} . There are algorithms for finding repetitions with gaps that could be useful for answering Open question 2, for instance [9,10,11,12,15].

Open question 2 *Is there an efficient procedure to determine if a finite set of partial words is unavoidable?*

References

1. Aho, A.V., Corasick, M.J.: Efficient string machines, an aid to bibliographic research. *Comm. ACM* **18** (1975) 333–340
2. Berstel, J., Boasson, L.: Partial words and a theorem of Fine and Wilf. *Theoret. Comput. Sci.* **218** (1999) 135–141
3. Blanchet-Sadri, F.: Codes, orderings, and partial words. *Theoret. Comput. Sci.* **329** (2004) 177–202
4. Blanchet-Sadri, F.: Primitive partial words. *Discrete Appl. Math.* **148** (2005) 195–213
5. Blanchet-Sadri, F., Duncan, S.: Partial Words and the Critical Factorization Theorem. *J. Combin. Theory Ser. A* **109** (2005) 221–245 <http://www.uncg.edu/mat/cft/>
6. Choffrut, C., Culik II, K.: On extendibility of unavoidable sets. *Discrete Appl. Math.* **9** (1984) 125–137
7. Choffrut, C., Karhumäki, J.: Combinatorics of Words. In Rozenberg, G., Salomaa, A. (eds.): *Handbook of Formal Languages*. Vol. 1. Springer-Verlag, Berlin (1997) 329–438
8. Ehrenfeucht, A., Haussler, D., Rozenberg, G.: On regularity of context-free languages. *Theoret. Comput. Sci.* **27** (1983) 311–322
9. Kolpakov, R., Kucherov, G.: Finding Approximate Repetitions Under Hamming Distance. *Lecture Notes in Comput. Sci.* Vol. 2161. Springer-Verlag, Berlin (2001) 170–181
10. Kolpakov, R., Kucherov, G.: Finding Approximate Repetitions Under Hamming Distance. *Theoret. Comput. Sci.* **33** (2003) 135–156
11. Landau, G., Schmidt, J.: An Algorithm for Approximate Tandem Repeats. *Lecture Notes in Comput. Sci.* Vol. 684. Springer-Verlag, Berlin (1993) 120–133
12. Landau, G.M., Schmidt, J.P., Sokol, D.: An Algorithm for Approximate Tandem Repeats. *J. Comput. Biology* **8** (2001) 1–18
13. Lothaire, M.: *Algebraic Combinatorics on Words*. Cambridge University Press, Cambridge (2002)
14. Rosaz, L.: Inventories of unavoidable languages and the word-extension conjecture. *Theoret. Comput. Sci.* **201** (1998) 151–170
15. Schmidt, J.P.: All Highest Scoring Paths in Weighted Grid Graphs and Their Application to Finding All Approximate Repeats in Strings. *SIAM J. Comput.* **27** (1998) 972–992