

On the Complexity of Deciding Avoidability of Sets of Partial Words*

Brandon Blakeley¹ F. Blanchet-Sadri² Josh Gunter¹
Narad Rampersad³

May 31, 2010

Abstract

Blanchet-Sadri et al. have shown that AVOIDABILITY, or the problem of deciding the avoidability of a finite set of partial words over an alphabet of size $k \geq 2$, is \mathcal{NP} -hard [Theoret. Comput. Sci. **410** (2009) 968–972]. Building on their work, we analyze in this paper the complexity of natural variations on the problem. While some of them are \mathcal{NP} -hard, others are shown to be efficiently decidable. Using some combinatorial properties of de Bruijn graphs, we establish a correspondence between lengths of cycles in such graphs and periods of avoiding words, resulting in a tight bound for periods of avoiding words. We also prove that AVOIDABILITY can be solved in polynomial space, and reduces in polynomial time to the problem of deciding the avoidability of a finite set of partial words of equal length over the binary alphabet. We give a polynomial bound on the period of an infinite avoiding word, in the case of sets of full words, in terms of two parameters: the length and the number of words in the set. We give a polynomial space algorithm to decide if a finite set of partial words is avoided by a non-ultimately periodic infinite word. The same algorithm also decides if the number of words of length n avoiding a given finite set of partial words grows polynomially or exponentially with n .

Keywords: Automata and formal languages; Computational complexity; Combinatorics on words; Partial words; Unavoidable sets; \mathcal{NP} -hard problems; De Bruijn graphs.

*This material is based upon work supported by the National Science Foundation under Grant No. DMS-0754154. The Department of Defense is also gratefully acknowledged. Part of this paper was presented at DLT'09 [2]. We thank the referees of a preliminary version of this paper for their very valuable comments and suggestions.

¹Department of Computer Sciences, The University of Texas at Austin, 1 University Station C0500 Taylor Hall 2.124, Austin, TX 78712-0233, USA

²Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402-6170, USA, blanchet@uncg.edu

³Department of Mathematics and Statistics, University of Winnipeg, 515 Portage Avenue, Winnipeg, MB R3B 2E9, CANADA

1 Introduction

A set of (full) words X over a finite alphabet A is called *unavoidable* if every two-sided infinite word over A has a factor in X (when a word w has no factor in X , we say that w avoids X); otherwise X is called *avoidable*. Consequently, a subset X of A^* is unavoidable if and only if $A^* \setminus A^*XA^*$ is finite, and any unavoidable set contains a finite one. An alternate characterization of a finite unavoidable set is that every periodic two-sided infinite word has a factor in X [15]. Among other topics, the cardinality of such sets has been investigated [15]. If we take, for example, one element from each of the conjugacy classes $\{aa\}$, $\{bb\}$ and $\{ab, ba\}$ of the set of length two words over the alphabet $\{a, b\}$, then we build an unavoidable set. Note that there is at least one element from each class in an unavoidable set of words of length two since we can construct an infinite word whose factors of length two all belong to the same class. This observation can be generalized, so that any unavoidable set of words of length m over a k -letter alphabet contains at least as many words as there are conjugacy classes. In [16], it was proved that this bound is sharp (see [6] for a simpler proof).

A set of partial words X over a finite alphabet A is called *unavoidable* if every two-sided infinite full word over A has a factor *compatible* with a member of X . Partial words are sequences that may contain some “holes,” denoted by “ \diamond ’s,” that match any letter of the alphabet (we also say that \diamond is compatible with any letter of the alphabet). For instance, $a\diamond bca\diamond b$ is a partial word with two holes over the alphabet $\{a, b, c\}$. Unavoidable sets of partial words were introduced in [4] where the number theoretic problem of classifying such sets of size $n \geq 2$ over a k -letter alphabet with $k \leq n$ was initiated. (Note that if X is unavoidable, then every infinite unary word has a factor compatible with a member of X , in particular, X cannot have fewer elements than the alphabet.)

Efficient algorithms to determine if a finite set of full words is unavoidable are well known [15]. For example, we can check whether there is a loop in the finite automaton of Aho and Corasick [1] recognizing $A^* \setminus A^*XA^*$. 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} , the set of all full words compatible with an element of X . Indeed, by the definition of \hat{X} , a two-sided infinite word w has a factor in \hat{X} if and only if that factor is compatible with a member of X . Thus the infinite words which avoid $X \subset A_\diamond^*$ are exactly those which avoid $\hat{X} \subset A^*$, and $X \subset A_\diamond^*$ is unavoidable if and only if $\hat{X} \subset A^*$ is unavoidable. However this incurs a dramatic loss in efficiency, as each partial word u in X can contribute as many as $|A|^{|H(u)|}$ elements to \hat{X} ($H(u)$ denotes the set of holes of u).

In [5], it was proved that testing the unavoidability of a finite set of partial words is much harder to handle than the similar problem for full words. Indeed, it turns out that the problem of deciding whether a finite set

of partial words over a k -letter alphabet where $k \geq 2$ is unavoidable is \mathcal{NP} -hard (the complexity class of those decision problems that are at least as hard as any problem that can be solved by a non-deterministic Turing machine in polynomial time), which is in contrast with the well known feasibility results for unavoidability of a set of full words [7, Chapter 7.4] (note that the case $k = 1$ is trivial).

The enumeration problem for words of length n avoiding a finite set of full words has been studied by several authors. For example, Kobayashi [14] presented a matrix-theoretic approach to this problem; Goulden and Jackson [11] describe another method. A set of words L is of *polynomial growth* if there exists a polynomial $p(n)$ such that the number of words in L of length n is at most $p(n)$ for all $n \geq 0$. The set L is of *exponential growth* if there exists a real number $r > 1$ such that for infinitely many $n \geq 0$, the number of words in L of length n is at least r^n . Over any fixed alphabet, the set of finite words avoiding any given finite set of partial words can be of either polynomial growth or exponential growth; no intermediate growth is possible. This is a consequence of previous work on the avoidability of sets of full words (see [14] for example).

The contents of our paper is as follows: In Section 2, we review basic concepts on partial words and discuss previous work on avoidability of sets of such words. In Section 3, we analyze the complexity of natural variations on AVOIDABILITY, or the problem of deciding the avoidability of a finite set of partial words over an alphabet of size $k \geq 2$. While some of them are shown to be \mathcal{NP} -hard, others are shown to be efficiently decidable. We establish a correspondence between lengths of cycles in de Bruijn graphs and periods of avoiding words, resulting in a bound for periods of avoiding words. We also show that AVOIDABILITY can be solved in polynomial space, and reduces in polynomial time to the problem of deciding the avoidability of a finite set of partial words of equal length over the binary alphabet. In Section 4, we give a polynomial bound on the period of an avoiding word, in the case of sets of full words, in terms of two parameters: the length and the number of words in the set. In Section 5, we give a polynomial space algorithm to decide if a finite set of partial words is avoided by a non-ultimately periodic infinite word over a fixed alphabet. Our algorithm also decides if the number of words of length n avoiding a given finite set of partial words grows polynomially or exponentially with n . We also apply the probabilistic method to show that if a set X of partial words is not too large, the number of words of length n avoiding X grows exponentially with n . Finally in Section 6, we conclude with some remarks.

2 Preliminaries

Throughout this paper A is a fixed non-empty finite set called an *alphabet* whose elements we call *letters*. A *word* of length n over A is a finite sequence of elements of A . We denote by A^* (respectively, A^n) the set of finite words (respectively, the set of words of length n) over A . 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 .

A *two-sided infinite word* w is a function $w : \mathbb{Z} \rightarrow A$. A finite word u is a factor of w if there exists some $i \in \mathbb{Z}$ such that $u = w(i) \cdots w(i + |u| - 1)$. For a positive integer p , w has *period* p , or w is *p-periodic*, if $w(i) = w(i + p)$ for all $i \in \mathbb{Z}$. If w has period p for some p , then we call w *periodic*. If v is a non-empty finite word, then we denote by $v^{\mathbb{Z}}$ the unique two-sided infinite word w with period $|v|$ such that $v = w(0) \cdots w(|v| - 1)$. A *one-sided infinite word* w is a function $w : \mathbb{N} \rightarrow A$. It is *ultimately periodic* if it can be written as $w = uvvvv \cdots$ for some finite words u and v , where v is non-empty.

A *partial word* u of length n over an alphabet A can be defined as a function $u : [0..n - 1] \rightarrow A_{\diamond}$, where $A_{\diamond} = A \cup \{\diamond\}$, and will be written as $u(0)u(1) \cdots u(n - 1)$. For $0 \leq i < n$, if $u(i) \in A$, then i belongs to the *domain* of u , denoted $D(u)$; otherwise, i belongs to the *set of holes* of u , denoted $H(u)$. Whenever $H(u)$ is empty, we say that u is a *full word*. We refer to an occurrence of the symbol \diamond as a *hole*. We denote by A_{\diamond}^* the set of all partial words over A with an arbitrary number of holes. For $u \in A_{\diamond}^*$, we also write $|u|$ for the length of u . A partial word v is a *factor* of the partial word u if there exist x, y such that $u = xvy$. We denote the set of all factors of u by $F(u)$. If $x = \varepsilon$, then v is a *prefix* of u ; if $y = \varepsilon$, v is a *suffix* of u .

Two partial words u and v of equal length are *compatible*, denoted $u \uparrow v$, if $u(i) = v(i)$ whenever $i \in D(u) \cap D(v)$. If X is a set of partial words, we denote by \hat{X} the set of all full words compatible with an element of X . The partial word u is *contained* in v , denoted $u \subset v$, if $|u| = |v|$ and $u(i) = v(i)$ for all $i \in D(u)$. Two partial words u and v are *conjugate* if there exist partial words x, y such that $u \subset xy$ and $v \subset yx$. It is well-known that conjugacy on full words is an equivalence relation, but it is not such a relation on partial words [3]. If a partial word u can be written as $u = u_1 \diamond u_2 \diamond \cdots u_{n-1} \diamond u_n$, then the set $\{u_1 a_1 u_2 a_2 \cdots u_{n-1} a_{n-1} u_n \mid a_1, a_2, \dots, a_{n-1} \in A\}$ is called a *partial expansion* on u (note that u_1, u_2, \dots, u_n are partial words that may contain holes, and also note that $u \subset v$ for every member v of a partial expansion on u).

A two-sided infinite full word w over A *avoids* $X \subset A_{\diamond}^*$ if no factor of w is an element of \hat{X} . We say that X is *unavoidable* if no two-sided infinite word over A avoids X . Previous work shows that AVOIDABILITY is \mathcal{NP} -hard. In [5], it is proved that determining if a finite set of partial words over an alphabet of size $k \geq 2$ is avoidable or not is much harder to handle than the similar problem for full words. This is done by using a reduction from

the 3SAT problem, known to be \mathcal{NP} -complete [10]. In [4], an algorithm, that will be used in some of the proofs, for deciding AVOIDABILITY is given based on the following reductions from a set X to a set Y that maintain avoidability:

1. Factoring – if $x, y \in X$ and $y' \in F(x)$ where $x \neq y \subset y'$, then $Y = X \setminus \{x\}$.
2. Prefix-Suffix – if there exists a partial word $x = ya \in X$ with $a \in A$ such that for every $b \in A$ there exists a suffix z of y and a partial word $v \in X$ with $v \subset zb$, then $Y = (X \setminus \{x\}) \cup \{y\}$.
3. Hole Truncation – if $x \diamond^n \in X$ for some positive integer n , then $Y = (X \setminus \{x \diamond^n\}) \cup \{x\}$.
4. Expansion – $Y = (X \setminus \{x\}) \cup W$ where W is a partial expansion on $x \in X$.

A set $X \subset A_\diamond^*$ is unavoidable if and only if X can be reduced to $\{\varepsilon\}$ by these reductions. To improve the efficiency of the algorithm, the Expansion operation is only considered valid if there exists $x = ya \in X$ with $a \in A$ such that for every $b \in A$ there exist a suffix z of y and a partial word $v \in X$ where $zb \uparrow v$, and the positions chosen to expand are such that a Prefix-Suffix operation might become valid. Note that if the algorithm reduces X down to Y and no more reductions are valid, then there exist no valid Prefix-Suffix operations on the set \hat{Y} .

We end this section with a few remarks. Any set of partial words containing the empty word or \diamond^n for some non-negative integer n will be called a *trivial* unavoidable set. If a set of partial words is unavoidable, then it must have an element compatible with a factor of each two-sided infinite unary word. In particular no non-trivial unavoidable set can have fewer elements than the alphabet. Every word avoids the empty set, so there are no unavoidable sets of size 0. It is also clear that unless the alphabet is unary the only unavoidable sets of size 1 are the trivial ones. If the alphabet is unary, then every non-empty set is unavoidable and in that case there is only one two-sided infinite word. We will not consider the unary alphabet further in this paper.

3 Complexity of Avoidability Problems

In this section, we discuss natural variations on AVOIDABILITY. While some of them are \mathcal{NP} -hard, others are shown to be efficiently decidable. We establish a correspondence between lengths of cycles in de Bruijn graphs and periods of avoiding words. We also show that the problem of deciding the avoidability of a finite set of partial words over a k -letter alphabet can be

solved in polynomial space, and reduces in polynomial time to the problem of deciding the avoidability of such a set over the binary alphabet.

Testing if a word avoids a finite set can be done using Lemma 1 which we will implicitly use when proving membership in certain complexity classes in the following results concerning restricted AVOIDABILITY.

Lemma 1. *Given a finite word v , the problem of deciding if the infinite periodic word $v^{\mathbb{Z}}$ avoids a finite set of partial words can be solved in polynomial time.*

Proof. Let x be an element in a finite set of partial words X . We claim that $v^{\mathbb{Z}}$ avoids x if and only if v^p avoids x , where $p \geq \lceil \frac{|x|}{|v|} \rceil + 1$. This follows because every factor of length $|x|$ of $v^{\mathbb{Z}}$ is also a factor of v^p . Our algorithm first identifies the longest string $x' \in X$, computes $p = \lceil \frac{|x'|}{|v|} \rceil + 1$, and constructs $w = v^p$. Then, for each partial word x in X , we decide if x is compatible with any factor of w using any efficient string matching algorithm with wildcards (such as [8], for example). \square

Theorem 1. *The problem of deciding the avoidability of a finite set of partial words, such that each element has at most two defined positions, is \mathcal{NP} -hard.*

Proof. Our proof proceeds by reduction from the Directed Hamiltonian Circuit problem, one of Karp's original twenty-one \mathcal{NP} -complete problems [13]. In the Directed Hamiltonian Circuit problem, we decide whether a given digraph has a Hamiltonian circuit. Given a digraph $G = (V, E)$, we construct a set X of partial words, where each element has at most two defined positions, such that X is avoidable if and only if G has a Hamiltonian circuit. Our alphabet is $V = \{v_1, v_2, \dots, v_n\}$ and the set X is composed of the following three parts: (1) $\{v_i v_j \mid (v_i, v_j) \notin E\}$, (2) $\{v_i \diamond^{n-1} v_j \mid v_i \neq v_j\}$, and (3) $\{v_i \diamond^j v_i \mid 0 \leq j < n - 1\}$.

For the forward implication, suppose there exists a Hamiltonian circuit in G , say $(u_1, u_2, \dots, u_n, u_1)$. We claim that $w = (u_1 u_2 \dots u_n)^{\mathbb{Z}}$ avoids X . Indeed, w avoids Part (1) of X because each $(u_i, u_{i+1}) \in E$ for $0 < i < n$ and $(u_n, u_1) \in E$. Part (2) is avoided because w is n -periodic. Part (3) is avoided because consecutive occurrences of the same letter are separated by $n - 1$ other letters. For the reverse implication, suppose there exists a two-sided infinite word w which avoids X . To avoid Part (3), consecutive occurrences of the same letter must be separated by at least $n - 1$ other letters. To avoid Part (2), $w(i) = w(i + n)$ for all $i \in \mathbb{Z}$, so w must be n -periodic. From our previous observations, this period must be of the form $u_1 u_2 \dots u_n$, where each u_i is distinct. Finally, to avoid Part (1), $(u_i, u_{i+1}) \in E$ where $0 < i < n$ and $(u_n, u_1) \in E$. Therefore, $(u_1, u_2, \dots, u_n, u_1)$ is a Hamiltonian circuit in G . \square

The following proposition shows membership in \mathcal{NP} of the problem defined in Theorem 1 over a binary alphabet.

Proposition 1. *The problem of deciding the avoidability of a finite set of partial words over the binary alphabet, such that each element has at most two defined positions, can be solved in non-deterministic polynomial time.*

Proof. Let a, b be the two distinct letters in the alphabet. Note that, for such a set X to not be avoided by $a^{\mathbb{Z}}$ or $b^{\mathbb{Z}}$, for some m and n , X must have elements $a \diamond^m a$ and $b \diamond^n b$. Furthermore, if X is avoidable, then it must be avoided by an infinite word with period $m + n + 2$. Therefore, we can decide these sets by non-deterministically selecting a word v of length $m + n + 2$ and verifying that $v^{\mathbb{Z}}$ avoids X . \square

The next theorem shows that another natural variation (see, for example, [6]), constant length sets, on the problem of deciding avoidability is \mathcal{NP} -hard.

Remark 1. *When we consider constant length sets of partial words, we implicitly require that neither the first or last position in any of the words be a hole.*

Theorem 2. *The problem of deciding the avoidability of a finite set of partial words of equal length over an alphabet of size $k \geq 2$ is \mathcal{NP} -hard.*

Proof. We present a reduction from the \mathcal{NP} -hard unrestricted AVOIDABILITY problem. Given a finite set X of partial words over a k -size alphabet A , we construct a set X' of partial words of equal length as follows. Let l denote the maximum length of the words in X . Then X' is formed by the following two parts: $\{u \diamond^{l-|u|-1} a \mid u \in X, |u| < l, a \in A\}$ and $\{u \mid u \in X, |u| = l\}$. We show that X' is avoided by the same words as X . Consider for any $u \in X$ where $|u| < l$ the set $X'_u = \{u \diamond^{l-|u|-1} a \mid a \in A\}$ which has the same avoidability as $Y'_u = \{u \diamond^{l-|u|}\}$ because an Expansion operation on Y'_u results in X'_u . Furthermore, a Hole Truncation operation on Y'_u results in the set $\{u\}$. Therefore, X'_u is avoided by the same words as $\{u\}$. By our construction of X' , clearly X' is avoided by the same words as X . Therefore, X is avoidable if and only if X' is avoidable. Finally, we note that the length of the description of X , that is $\|X\| = \sum_{x \in X} |x|$, satisfies $\|X'\| < \|X\|lk$, and so this reduction runs in polynomial time. \square

A tractable variation is provided in the next theorem. As a direct corollary, combining the two restrictions presented in the previous two theorems results in a problem which can be efficiently decided.

Theorem 3. *The problem of deciding the avoidability of a finite set X of partial words, where for some positive integer p every element $x \in X$ is defined at position $0 \leq i < |x|$ if and only if p divides i , can be solved in polynomial time.*

Proof. Our algorithm decides avoidability of X as follows: Construct the set X' by removing all holes from words in X , and then run on X' an algorithm which decides avoidability of sets of full words. We show that X is unavoidable if and only if X' is unavoidable. Suppose X' is avoidable, and $(v')^{\mathbb{Z}}$ avoids X' . It is easy to see that the word $v^{\mathbb{Z}}$, where $v = v'(0)^p v'(1)^p \cdots v'(|v'| - 1)^p$, avoids X . Now suppose X' is unavoidable. Then there exists some sequence of the operations defined in Section 2 which reduces the set X' to $\{\varepsilon\}$. We claim that the same sequence but with Hole Truncations following every Prefix-Suffix operation reduces the set X to $\{\varepsilon\}$. Observe that if $x' \in X'$ is a factor of $y' \in X'$, then the word $x \in X$ formed by placing $p - 1$ holes between each letter in x' is necessarily a factor of $y \in X$ formed by the same process. Furthermore, if there exists a word $x' = y'a \in X'$ with $a \in A$ such that for every $b \in A$ there exists a suffix z' of y' where $z'b \in X'$, then the word $x \in X$ formed by placing $p - 1$ holes between every letter in x' necessarily satisfies the same conditions for a Prefix-Suffix application. Finally, note that the property defining the elements in these sets remains invariant across any rule application. \square

Corollary 1. *The problem of deciding the avoidability of a given finite set of partial words of equal length n , where each element has at most two defined positions (by Remark 1, each element has the form $a \diamond^{n-2} b$), can be solved in polynomial time.*

Another natural variation of AVOIDABILITY is presented in the next theorem. A combinatorial problem with input containing numerical parameters is said to be *strongly \mathcal{NP} -complete* if it is \mathcal{NP} -complete even when all numerical parameters are bounded by some polynomial in the size of the input.

Theorem 4. *The problem of deciding whether a finite set of partial words is avoided by a word of length l is strongly \mathcal{NP} -complete. In other words, the problem of deciding whether a finite set of partial words is avoided by a word of length l is \mathcal{NP} -complete even when l is bounded by some polynomial in the size of the input.*

Proof. Our result is a generalization of that presented in [9] for the analogous problem over full words, and our proof follows the same structure. First, we show that the problem is contained in \mathcal{NP} by presenting a non-deterministic algorithm which decides the problem in polynomial time. In short, our algorithm non-deterministically selects a word w of length l and verifies that no element in the set is compatible with a factor of w .

Next, we show that the problem is \mathcal{NP} -hard by reducing the Long Path [10] problem to it. In the Long Path problem, we decide whether a graph has a simple path of a certain length. Given an instance $G = (V, E), l$ of the Long Path problem, we construct a set X of partial words such that G has a simple path of length l if and only if X is avoided by a word of

length l . Our alphabet is $V = \{v_1, v_2, \dots, v_n\}$ and our set $X = \{v_i v_j \mid (v_i, v_j) \notin E\} \cup \{v_i \diamond^j v_i \mid 0 \leq j \leq l - 2\}$. Suppose there exists a simple path u_1, u_2, \dots, u_l in G . Then the word $u_1 u_2 \dots u_l$ avoids X , as each of $(u_i, u_{i+1}) \in E$ and $u_i \neq u_j$ for every $1 \leq i, j \leq l, i \neq j$. Now, suppose a word w of length l avoids X . Then w must have only distinct symbols, and each pair of adjacent letters in w must be elements of E . So let $w = u_1 u_2 \dots u_l$. Then the sequence u_1, u_2, \dots, u_l is a simple path in G . \square

Using de Bruijn graphs, well known to be Hamiltonian and Eulerian, Theorem 5 will give a bound on periods of avoiding words by establishing a correspondence between them and lengths of cycles in such graphs. Recall that the *line digraph* (V', E') of a digraph $G = (V, E)$, denoted $L(G)$, is such that $V' = E$ and $E' = \{((u, v), (v, w)) \mid u, v, w \in V\}$. The *de Bruijn graph* of order m over a k -size alphabet A , denoted $G(m, k)$, is the digraph (V, E) defined as follows: if $m = 1$, then $V = A$ and $E = \{(a, b) \mid a, b \in A\}$, and if $m > 1$, then $G(m, k) = L(G(m - 1, k))$. Equivalently, for $m \geq 1$, $V = A^m$ and $E = \{(z, z') \mid z' = \sigma(z, a)\}$, where we denote by $\sigma(z, a)$ the word z' of length m such that for some $b \in A$, $za = bz'$ (here $b = z(0)$, $z'(0) = z(1)$, $z'(1) = z(2)$, \dots , $z'(m - 2) = z(m - 1)$, $z'(m - 1) = a$).

Theorem 5. *If a finite set of partial words of length m over an alphabet A is avoidable, then it is avoided by a word of period at most $|A|^m$.*

Proof. Let X be a finite avoidable set of partial words of length m . Consider the subgraph $G = (V, E)$ of $G(m, k)$ induced by the set $\{u \mid u \not\supset x \text{ for all } x \in X\}$. Essentially, incidence in G corresponds to transitions in the automaton of Aho and Corasick [1]. We claim that there exists a cycle in G of length p if and only if there exists an infinite word with period p which avoids X . Consider any cycle C in G of length p . Construct the word v_C formed by concatenating the first letters of each vertex along the cycle. By our construction of G , no subword of $(v_C)^\mathbb{Z}$ of length m is compatible with any word in X . Therefore, the infinite word $(v_C)^\mathbb{Z}$ of period p avoids X . Now suppose there exists a cycle C in G of length greater than $|V|$. Then, by the pigeonhole principle, C is not simple, and so we can find a simple cycle C' of length at most $|V|$. Therefore, since $|V| \leq |A|^m$, there exists an infinite word with period at most $|A|^m$ that avoids X . \square

Theorem 6. *The problem of deciding the avoidability of a finite set of partial words of equal length can be solved in polynomial space.*

Proof. We apply the bound found in Theorem 5 to obtain a polynomial space algorithm which decides the avoidability of a finite set X of partial words of length m over alphabet A . Algorithm 1 searches for a cycle in the graph defined in Theorem 5 without constructing the graph. The correctness of this algorithm can be proved with the loop invariant that, at iteration i ,

Algorithm 1 Deciding Avoidability of a Set X of Partial Words of Length m over alphabet A in Polynomial Space

- 1: Non-deterministically select a word w of length m
 - 2: Set $z = w$, $i = 0$
 - 3: **while** $i < |A|^m$ **do**
 - 4: Increment i
 - 5: Non-deterministically select a letter $a \in A$
 - 6: Set $z = \sigma(z, a)$
 - 7: If $\exists x \in X$ such that $z \uparrow x$, reject
 - 8: If $z = w$, accept
 - 9: Reject
-

there is a path of length i from w to z . So if there is a cycle in G , then there is a cycle with at most $|A|^m$ vertices and our algorithm will accept. If there is no cycle, then there is no path from w to w of length at least 1 and our algorithm will reject. Because our algorithm stores only two words of length m and a counter of length $m \log |A|$, it uses $\mathcal{O}(m)$ non-deterministic space, and so, by Savitch's theorem [17], only $\mathcal{O}(m^2)$ deterministic space, which is polynomial in the input's length. \square

Generalizing to arbitrary sets, we get the following corollary.

Corollary 2. AVOIDABILITY *is in* PSPACE.

Proof. The problem of deciding the avoidability of an arbitrary finite set of partial words over a k -letter alphabet can be solved in polynomial space. Indeed, the polynomial time reduction to constant length sets in Theorem 2 followed by the polynomial space Algorithm 1 decides the problem. \square

We now consider reducing AVOIDABILITY to the binary alphabet.

Theorem 7. *The problem of deciding the avoidability of a finite set of partial words over an alphabet of size $k > 2$ reduces in polynomial time to the problem of deciding the avoidability of a finite set of partial words over the binary alphabet.*

Proof. Given a finite set X of partial words over alphabet $A = \{a_1, a_2, \dots, a_k\}$, we construct a set X' of partial words over the alphabet $B = \{0, 1\}$ such that X' is avoidable if and only if X is avoidable. At a high level, our reduction encodes each symbol in a binary representation and delimits adjacent encodings with a special binary word. Let $l = \lceil \log_2 |A| \rceil$ be the length of an encoding, $d = 101$ be the delimiting word, and define the sets $S = \{00, 11\}$ and $T = B^3 \setminus \{101\}$. Finally, define the function $b : A_\diamond \rightarrow S^l \cup \{(\diamond\diamond)^l\}$ to be such that $b(a_i)$ equals the binary representation of the natural number $i - 1$, where each bit is replaced with two copies of itself, and $b(\diamond) = (\diamond\diamond)^l$. We now describe the elements in X' :

1. First, add every word of length $2l + 3$ which does not contain 101 as a factor in order to ensure that any avoiding infinite word has 101 as a factor.
2. Second, for each $t \in T$, add $101(\diamond\diamond)^lt$. To avoid these words, every occurrence of 101 in an infinite avoiding word must be followed by another 101 after $2l$ other bits.
3. Third, for each $0 \leq i < l$, add the words $101(\diamond\diamond)^i01$ and $101(\diamond\diamond)^i10$. This forces avoiding words to have only valid binary representations (that is, words from S^l) between consecutive pairs of 101.
4. Fourth, for each word $u_0u_1 \cdots u_{m-1} \in X$, where each $u_i \in A_\diamond$, add to X' the word $101b(u_0)101b(u_1) \cdots 101b(u_{m-1})$ which enforces a bijection between words which avoid X and words which avoid X' .
5. Finally, for each $|A| < i \leq 2^l$, add the corresponding binary representation of $i - 1$ where each bit is replaced with two copies of itself. This ensures that every factor of length $2l$ delimited by 101 in an infinite avoiding word corresponds to a symbol in A .

Suppose some infinite word $w = (w_0w_1 \cdots w_{n-1})^{\mathbb{Z}}$ avoids X , where each $w_i \in A$. Then clearly the word $w' = (101b(w_0)101b(w_1) \cdots 101b(w_{n-1}))^{\mathbb{Z}}$ avoids X' . Next, suppose that some infinite word w' avoids X' . Then, to avoid the first part, w' must have 101 as a factor. Additionally, to avoid the second part, following every occurrence of 101 in w' there must be another occurrence of 101 in w' after $2l$ other bits. Furthermore, to avoid the third part, these bits must come in pairs. Moreover, to avoid the last part, these $2l$ bits must form a binary representation of some symbol in A . So one period of our word w' must be of the form $101b(u_0)101b(u_1) \cdots 101b(u_{n-1})$ for some $u_i \in A$. Finally, to avoid the fourth part, $(u_0u_1 \cdots u_{n-1})^{\mathbb{Z}}$ must avoid X .

Note that all but the fourth part of X' are functions of only the size of the alphabet. Because the alphabet is constant, these sets are constant with respect to the input. Therefore, because the fourth part grows linearly with respect to X , this reduction can be performed in polynomial time. \square

Theorem 7 shows that problems of deciding avoidability of sets over alphabets of sizes at least two are equivalent with respect to polynomial time reductions; that is, they are all in the same complexity class. A more rigorous analysis of the space complexity of our reduction, in conjunction with Theorem 1, provides an alternate proof of the \mathcal{NP} -hardness of the general problem.

Corollary 3. *AVOIDABILITY is \mathcal{NP} -hard.*

Proof. The problem of deciding the avoidability of a finite set of partial words over an alphabet of size $k \geq 2$ is \mathcal{NP} -hard. Indeed, we prove that the reduction from the Directed Hamiltonian Circuit problem followed by the reduction to the binary alphabet is polynomial time, and therefore suffices to show the avoidability problem \mathcal{NP} -hard. The reduction in Theorem 1 uses $\mathcal{O}(|V|^3)$ space, while the reduction in Theorem 7 uses $\mathcal{O}(|A|^2 + |X| \log |A|)$ space. Because $A = V$, the composition of these reductions uses $\mathcal{O}(|V|^2 + |V|^3 \log |V|) = \mathcal{O}(|V|^3 \log |V|)$ space. As both are polynomial time, so is their composition. This concludes the proof when $k = 2$. As in [5], for $k > 2$, we simply forbid the other letters, a_3, \dots, a_k , of the alphabet by including them in the set. \square

Note that in some avoidability problems the alphabet is not fixed, but in AVOIDABILITY it is, so that the problem in Theorem 1 is not a special case of AVOIDABILITY and does not directly imply \mathcal{NP} -hardness (Corollary 3).

Additionally, Theorem 7 shows that if every finite avoidable set of partial words over some alphabet of size k is avoided by an infinite word with a period bounded by a polynomial in the size of the set, then so is every finite avoidable set over an alphabet of any size. Moreover, by applying Theorem 2, we can reduce all these avoidability problems to the problem of deciding the avoidability of a finite set of partial words of equal length over the binary alphabet. In the next section, we exploit properties of these reduced sets to present some partial results towards a polynomial bound on the period of an avoiding word.

4 Polynomial Bound on Periods of Avoiding Words

In the previous section, we reduced the problem of deciding the avoidability of a finite set of partial words over an alphabet of size $k \geq 2$ to the problem of deciding the avoidability of a finite set of partial words of equal length over the binary alphabet. This reduction simplifies our problem significantly, most notably by allowing us to consider only two parameters when establishing a bound on the shortest period of an avoiding word: the length of the words in the set and the number of elements in the set.

The following theorem establishes a bijection from simple cycles to subset-minimal cycles in de Bruijn graphs (a cycle C in a graph G is *subset-minimal* if there does not exist a shorter cycle D such that every vertex in D is also in C).

Theorem 8. *Let $G(m, k)$ be the de Bruijn graph of order m over an alphabet of size k . There exists a bijection from simple cycles in $G(m, k)$ to subset-minimal cycles in $G(m + 1, k)$ which preserves cycle length.*

Proof. We claim that the following function from simple cycles in $G(m, k)$ to subset-minimal cycles in $G(m + 1, k)$ is a bijection:

$$f([v_1, v_2, \dots, v_n]) = [(v_1, v_2), (v_2, v_3), \dots, (v_n, v_1)]$$

where the edge (v_i, v_{i+1}) in $G(m, k)$ is a vertex in $G(m + 1, k)$. Consider any cycle $C = [v_1, v_2, \dots, v_n]$ in $G(m, k)$. Note that $|C| = |f(C)|$, so f preserves cycle length. First, we show that if C is simple, then $f(C)$ is subset-minimal. Suppose for contradiction that there exists a proper subset of the set of vertices in $f(C)$ which forms a cycle; then some v_i would appear as a starting point in $f(C)$ more than once for some $i \in [1..n]$. This corresponds to v_i appearing in C more than once, meaning that C is not simple, a contradiction. Now, suppose C is not simple. So some v_i appears more than once in C . Therefore v_i appears as a starting point and ending point in $f(C)$ more than once. Consider the subpath $D = [(v_i, v_j), \dots, (v_{j'}, v_i)]$ where $(v_{j'}, v_i)$ is the first occurrence of v_i as an ending point after (v_i, v_j) . Then D forms a cycle, and since it has v_i as a starting and ending point exactly once, D is proper; therefore $f(C)$ is not subset-minimal. \square

The first corollary gives the lengths of the longest subset-minimal cycles in de Bruijn graphs, the second, which strengthens Theorem 5, provides a tight bound for periods of avoiding words, and the third is a negative result on polynomially bounded periods.

Corollary 4. *Let $G(m, k)$ be the de Bruijn graph of order m over an alphabet of size k . The length of the longest subset-minimal cycle in $G(m, k)$ is k^{m-1} .*

Proof. If $m = 1$, then the longest subset-minimal cycle for $G(1, k)$ has length $k^{1-1} = 1$ since every vertex contains a self-loop. If $m > 1$, then by Theorem 8, the longest subset-minimal cycle in $G(m, k)$ corresponds to the longest simple cycle in $G(m - 1, k)$. Since de Bruijn graphs are Hamiltonian, the longest simple cycle in $G(m - 1, k)$ is Hamiltonian. Since $G(m - 1, k)$ has k^{m-1} vertices, k^{m-1} is the length of the longest subset-minimal cycle in $G(m, k)$. \square

Corollary 5. *If a finite set of partial words of length m over a k -letter alphabet is avoidable, then it is avoided by a word with period at most k^{m-1} . Furthermore, for every m , a finite set of partial words of length m over a k -letter alphabet exists such that the smallest period of an infinite word which avoids the set is k^{m-1} .*

Proof. Because words which avoid a finite set of partial words of length m over a k -letter alphabet correspond to cycles in the subgraph of the de Bruijn graph $G(m, k)$ induced by removing vertices compatible with elements in X ,

the shortest cycle corresponds to the shortest period of an avoiding word. Since the longest subset-minimal cycle in $G(m, k)$ has length k^{m-1} , the shortest period of an avoiding word is at most k^{m-1} (refer to the proof of Theorem 5). Furthermore, if $G(m, k)$ contains only this cycle, the shortest period of an avoiding word is exactly k^{m-1} , and hence the bound is optimal. \square

Corollary 6. *No polynomial function p of m exists such that all avoidable sets of partial words of length m over a k -letter alphabet are avoided by an infinite word with period at most $p(m)$.*

Proof. By Corollary 5, for every m there exists a finite avoidable set of partial words of length m over a k -letter alphabet such that the smallest period of an infinite word which avoids the set is k^{m-1} . Suppose there were such a polynomial function $p(m)$. Hence, there exists some m' such that $p(m') < k^{m'-1}$, a contradiction. \square

In short, if there exists a polynomial function p from a set of n partial words of length m over a k -letter alphabet to an upper bound on the shortest period of an infinite word which avoids the set, then p is a function of both n and m . We now state our conjecture regarding the existence of a polynomial bound in terms of n and m .

Conjecture 1. *If a set of n partial words of length m over a k -letter alphabet is avoidable, it is avoided by an infinite periodic word with period at most mn .*

The following propositions present some positive results towards verifying Conjecture 1. Recall that we assume without loss in generality that every element in the set is defined at both its first and last position.

Proposition 2. *Conjecture 1 is true when $n \leq 2$.*

Proof. Let a, b be any two distinct letters in the alphabet. When $n = 1$, the set is avoided either by $a^{\mathbb{Z}}$ or $b^{\mathbb{Z}}$. When $n = 2$, if both words in the set contain a common letter (without loss of generality suppose it is a), then the set is avoided by $b^{\mathbb{Z}}$. Otherwise, if both words in the set have no letters in common, then without loss of generality suppose that one of the words begins with a and the other ends with b . In this case the set is avoided by $(a^{m-1}b^{m-1})^{\mathbb{Z}}$. \square

We note that proving Conjecture 1 is much more difficult for sets of partial words than for sets of full words, as we must consider how many fewer elements are needed in a set of partial words. Consequently, the following result is restricted to sets of full words. For positive integers m and k , let $c(m, k)$ be the number of conjugacy classes of words of length m over a k -letter alphabet. It was shown in [6] that for every m and k , there

exists an unavoidable set of full words of length m over an alphabet of size k having $c(m, k)$ elements.

Proposition 3. *Conjecture 1 is true for sets of full words.*

Proof. We show that every avoidable set X of n full words of length m over a k -letter alphabet is avoided by an infinite word of period at most mn . We first claim that $c(m, k) \geq \frac{k^m}{m}$. This follows because there are k^m words of length m and each conjugacy class has at most m elements. We now consider two possible cases: First, suppose $n < \frac{k^m}{m}$. Then because $n < c(m, k)$, some conjugacy class is not represented in X ; that is, there exists some word v of length m such that for all $u \in [v]$ and $x \in X$, $u \neq x$. Therefore, the word $v^{\mathbb{Z}}$ of period m avoids X . Now, suppose $n \geq \frac{k^m}{m}$. By Corollary 5, X is avoided by an infinite word with period at most k^{m-1} . The result follows since $mn \geq k^m > k^{m-1}$. \square

There is no difficulty in proving the conjecture for sets containing $n \geq c(m, k)$ partial words since in this case $n \geq c(m, k) \geq \frac{k^m}{m}$. Hence, by Corollary 5, the set is avoided by an infinite word with period at most $k^{m-1} < k^m \leq mn$. On the other hand, the conjecture is not clear if $n < c(m, k)$, since the set of full words compatible with the original set of n partial words can contain elements from each conjugacy class.

Referring to Theorem 5, Conjecture 1 is true if and only if the shortest cycle in the subgraph of $G(m, k)$ induced by the set of words not compatible with elements in an avoidable set of n partial words of length m over a k -letter alphabet has at most mn vertices.

5 Sets Avoidable by Aperiodic Infinite Words

We now consider the problem of determining whether or not there is a non-ultimately periodic infinite word avoiding a given set of partial words.

Theorem 9. *There is a polynomial space algorithm to decide if a finite set of partial words over a k -letter alphabet is avoided by a non-ultimately periodic infinite word. Equivalently, there is a polynomial space algorithm to decide if the number of words of length n that avoid a finite set of partial words over a k -letter alphabet grows polynomially or exponentially with n .*

Proof. Suppose we are given a finite set X of partial words over a k -letter alphabet. Let us first perform the transformation of X to X' as described in the proof of Theorem 2. The set X' consists of partial words of the same length m , and the words avoiding X' are exactly the words avoiding X .

Let $G(m, k)$ be the de Bruijn graph of order m and let G be the subgraph of $G(m, k)$ induced by the set $\{u \mid u \not\sim x \text{ for all } x \in X'\}$. It is clear that there is a non-ultimately periodic infinite word avoiding X' if and only if G

contains two distinct directed cycles C_1 and C_2 such that there is a directed path P_1 from C_1 to C_2 and a directed path P_2 from C_2 to C_1 . Similarly, the number of words of length n that avoid X' grows exponentially with n if and only if there exist C_1 , C_2 , P_1 , and P_2 as described above.

To determine the existence of C_1 , C_2 , P_1 , and P_2 , we apply a variation of Algorithm 1. We only describe the changes required to Algorithm 1. Instead of non-deterministically choosing a single word w , we instead choose two distinct words w and v of length m . We then non-deterministically search for cycles in G from w to w and from v to v of lengths at most k^m , just as in Algorithm 1. Using the same technique, we non-deterministically search for paths P_1 from w to v and P_2 from v to w in G , where P_1 and P_2 have length at most k^m . This non-deterministic algorithm uses only $\mathcal{O}(m)$ space, so there is an equivalent deterministic algorithm that runs in $\mathcal{O}(m^2)$ space by Savitch's theorem [17]. \square

Our next theorem uses the probabilistic method and is therefore non-constructive. Let A_1, \dots, A_n be events in a probability space. A graph $G = (V, E)$ is a *dependency graph* if $V = \{1, \dots, n\}$ and for all i , A_i is mutually independent of all the A_j 's for which there is no edge $\{i, j\} \in E$.

Lemma 2 ([12], Lemma 19.1). *Let $G = (V, E)$ be a dependency graph for events A_1, \dots, A_n in a probability space. Suppose that the maximum degree of G is d and that there is a real number p for which $\Pr[A_i] \leq p$ for all $i = 1, \dots, n$. If $4pd \leq 1$, then $\Pr[\cap \overline{A_i}] \geq (1 - 2p)^n > 0$.*

We use the above result, known as Lovász Local Lemma (symmetric version), to prove that if a set X of partial words is not too large, the number of words of length n avoiding X grows exponentially with n .

Theorem 10. *Let X be a set of partial words of length $m \geq 2$ with at most $h < m$ holes over an alphabet A of size $k \geq 2$. If $|X| \leq \frac{k^{m-h}}{4(2m-1)}$, then for $n \geq 1$, there are at least $\left[k \left(1 - \frac{1}{4m-2} \right) \right]^n$ words of length n over A that avoid X . Furthermore, there is a non-ultimately periodic infinite word over A that avoids X .*

Proof. Let n be an arbitrary positive integer and let w be a random word of length n over A . For $i = 1, \dots, n$, let A_i denote the event that w contains a factor compatible with a partial word in X at position $i - 1$. Let $p = \frac{|X|}{k^{m-h}}$, so that for all i , $\Pr[A_i] \leq p$. To apply the local lemma we may take $d = 2m - 1$, since there can be at most $2m - 1$ overlapping pairs of occurrences of factors of length m in w . Observe that for $|X| \leq \frac{k^{m-h}}{4(2m-1)}$, we have $p = \frac{|X|}{k^{m-h}} \leq \frac{1}{4(2m-1)}$, so that $4pd \leq 1$. By the local lemma, with probability at least $(1 - 2p)^n \geq (1 - \frac{1}{4m-2})^n$, w contains no factor compatible with a partial word in X . There are therefore at least $\left[k \left(1 - \frac{1}{4m-2} \right) \right]^n$ words of

length n over A that avoid X . Since $k, m \geq 2$, we have $k(1 - \frac{1}{4m-2}) > 1$, so the number of words of length n over A that avoid X grows exponentially with n . We conclude by observing that by our discussion in the proof of Theorem 9, there are exponentially many words of length n avoiding X if and only if there is a non-ultimately periodic infinite word avoiding X . \square

6 Conclusion and Open Problems

In this paper, we have established the membership of AVOIDABILITY in PSPACE, have reduced AVOIDABILITY to constant length sets over the binary alphabet, have formulated a conjecture about polynomially bounding periods of infinite avoiding words and have proven it for the special case of sets of full words, have given a polynomial space algorithm that determines if a given finite set of partial words is avoided by a non-ultimately periodic infinite word and that also determines if the number of words of length n avoiding the given set grows polynomially or exponentially with n , and have also applied the probabilistic method to show that if a set of partial words is not too large, the number of words of length n avoiding it grows exponentially with n . However, membership of AVOIDABILITY in \mathcal{NP} remains open. A World Wide Web server interface has been established at www.uncg.edu/cmp/research/unavoidablesets3 for automated use of a program that when given as input a finite set of partial words over a given alphabet will output the shortest period of an infinite avoiding word in case the set is avoidable.

References

- [1] Aho, A., Corasick, M.: Efficient string machines, an aid to bibliographic research. *Comm. of the ACM* **18** (1975) 333–340
- [2] Blakeley, B., Blanchet-Sadri, F., Gunter, J., Rampersad, N.: On the complexity of deciding avoidability of sets of partial words. In Diekert, V., Nowotka, D. (eds.): *DLT 2009, LNCS 5583* (Springer-Verlag, Berlin, Heidelberg, 2009) 113–124
- [3] Blanchet-Sadri, F.: *Algorithmic Combinatorics on Partial Words*. Chapman & Hall/CRC Press (2008)
- [4] Blanchet-Sadri, F., Brownstein, N.C., Kalcic, A., Palumbo, J., Weyand, T.: Unavoidable sets of partial words. *Theor. Comput. Sys.* **45** (2009) 381–406
- [5] Blanchet-Sadri, F., Jungers, R., Palumbo, J.: Testing avoidability on sets of partial words is hard. *Theoret. Comput. Sci.* **410** (2009) 968–972

- [6] Champarnaud, J.M., Hansel, G., Perrin, D.: Unavoidable sets of constant length. *Internat. J. Algebra Comput.* **14** (2004) 241–251
- [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] Clifford, P., Clifford, R.: Simple deterministic wildcard matching. *Inform. Process. Lett.* **101** (2007) 53–54
- [9] Evdokimov, A., Kitaev, S.: Crucial words and the complexity of some extremal problems for sets of prohibited words. *J. Combin. Theory Ser. A* **105** (2004) 273–289
- [10] Garey, M.R., Johnson, D.S.: *Computers and Intractability - A Guide to the Theory of NP-Completeness*. Freeman (1979)
- [11] Goulden, I., Jackson, D.: *Combinatorial Enumeration*. Dover (2004)
- [12] Jukna, S.: *Extremal Combinatorics*. Springer (2001)
- [13] Karp, R.M.: Reducibility among combinatorial problems. In Miller, R.E., Thatcher, J.W. (eds.): *Complexity of Computer Computations*. Plenum, New York (1972) 85–103
- [14] Kobayashi, Y.: Repetition-free words, *Theoret. Comput. Sci* **44** (1986) 175–197
- [15] Lothaire, M.: *Algebraic Combinatorics on Words*. Cambridge University Press (2002)
- [16] Mykkeltveit, J.: A proof of Golomb’s conjecture for the de Bruijn graph. *J. Combin. Theory Ser. B* **13** (1972) 40–45
- [17] Savitch, W.J.: Relationship between nondeterministic and deterministic tape classes. *J. Comput. Syst. Sci.* **4** (1970) 177–192