Theoretical Computer Science 506 (2013) 17–28

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Strict bounds for pattern avoidance ☆

F. Blanchet-Sadri [a,*], Brent Woodhouse [b]

[a] Department of Computer Science, University of North Carolina, P.O. Box 26170, Greensboro, NC 27402-6170, USA
[b] Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907-2067, USA

ARTICLE INFO

Article history:
Received 15 March 2013
Received in revised form 24 July 2013
Accepted 19 August 2013
Communicated by A. Apostolico

Keywords:
Formal languages
Combinatorics on words
Pattern avoidance
Power series
Partial words

ABSTRACT

Cassaigne conjectured in 1994 that any pattern with m distinct variables of length at least $3(2^{m-1})$ is avoidable over a binary alphabet, and any pattern with m distinct variables of length at least 2^m is avoidable over a ternary alphabet. Building upon the work of Rampersad and the power series techniques of Bell and Goh, we obtain both of these suggested strict bounds. Similar bounds are also obtained for pattern avoidance in partial words, sequences where some characters are unknown.

1. Introduction

Let Σ be an alphabet of letters, denoted by $a, b, c, \dots$, and Δ be an alphabet of variables, denoted by $A, B, C, \dots$. A *pattern* p is a word over Δ. A word w over Σ is an *instance* of p if there exists a non-erasing morphism $\varphi : \Delta^* \to \Sigma^*$ such that $\varphi(p) = w$. A word w is said to *avoid* p if no factor of w is an instance of p. For example, $\underline{aa}\ \underline{b}\ \underline{aa}\ c$ contains an instance of ABA while *abaca* avoids AA.

A pattern p is *avoidable* if there exist infinitely many words w over a finite alphabet such that w avoids p, or equivalently, if there exists an infinite word that avoids p. Otherwise p is *unavoidable*. If p is avoided by infinitely many words over a k-letter alphabet, p is said to be k-*avoidable*. Otherwise, p is k-*unavoidable*. If p is avoidable, the minimum k such that p is k-avoidable is called the *avoidability index* of p. If p is unavoidable, the avoidability index is defined as ∞. For example, ABA is unavoidable while AA has avoidability index 3.

If a pattern p occurs in a pattern q, we say p *divides* q. For example, $p = ABA$ divides $q = \underline{ABC}\ \underline{BB}\ \underline{ABC}\ A$, since we can map A to ABC and B to BB and this maps p to a factor of q. If p divides q and p is k-avoidable, there exists an infinite word w over a k-letter alphabet that avoids p; w must also avoid q, thus q is necessarily k-avoidable. It follows that the avoidability index of q is less than or equal to the avoidability index of p. Chapter 3 of Lothaire [6] is a nice summary of background results in pattern avoidance.

It is not known if it is generally decidable, given a pattern p and integer k, whether p is k-avoidable. Thus various authors compute avoidability indices and try to find bounds on them. Cassaigne [5] listed avoidability indices for unary, binary, and most ternary patterns (Ochem [8] determined the remaining few avoidability indices for ternary patterns). Based on this data, Cassaigne conjectured in his 1994 Ph.D. thesis [5, Conjecture 4.1] that any pattern with m distinct variables of length

☆ This material is based upon work supported by the National Science Foundation under Grant No. DMS-1060775. We thank the referees of preliminary versions of this paper for their very valuable comments and suggestions.

* Corresponding author.
E-mail addresses: blanchet@uncg.edu (F. Blanchet-Sadri), bwoodhou@purdue.edu (B. Woodhouse).

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2013.08.010

at least $3(2^{m-1})$ is avoidable over a binary alphabet, and any pattern with m distinct variables of length at least 2^m is avoidable over a ternary alphabet. This is also [6, Problem 3.3.2].

The contents of our paper are as follows. In Section 2, we establish that both bounds suggested by Cassaigne are strict by exhibiting well-known sequences of patterns that meet the bounds. Note that the results of Section 2 were proved by Cassaigne in his Ph.D. thesis with the same patterns (see [5, Proposition 4.3]). We recall them here for sake of completeness. In Section 3, we provide foundational results for the power series approach to this problem taken by Bell and Goh [1] and Rampersad [10], then proceed to prove the strict bounds in Section 4. In Section 5, we apply the power series approach to obtain similar bounds for avoidability in partial words, sequences that may contain some do-not-know characters, or holes, which are *compatible* or match any letter in the alphabet. The modifications include that now we must avoid all partial words compatible with instances of the pattern. Lots of additional work with inequalities is necessary. Finally in Section 6, we conclude with various remarks and conjectures.

2. Two sequences of unavoidable patterns

The following proposition allows the construction of sequences of unavoidable patterns.

Proposition 1. *(See [6, Proposition 3.1.3].) Let p be a k-unavoidable pattern over Δ and $A \in \Delta$ be a variable that does not occur in p. Then the pattern pAp is k-unavoidable.*

Let $A_1, A_2, \ldots$ be distinct variables in Δ. Define $Z_0 = \varepsilon$, the empty word, and for all integers $m \geqslant 0$, $Z_{m+1} = Z_m A_{m+1} Z_m$. The patterns Z_m are called Zimin words. Since ε is k-unavoidable for every positive integer k, Proposition 1 implies Z_m is k-unavoidable for all $m \in \mathbb{N}$ by induction on m. Thus all the Zimin words are unavoidable. Note that Z_m is over m variables and $|Z_m| = 2^m - 1$. Thus there exists a 3-unavoidable pattern over m variables with length $2^m - 1$ for all $m \in \mathbb{N}$.

Likewise, define $R_1 = A_1 A_1$ and for all integers $m \geqslant 1$, $R_{m+1} = R_m A_{m+1} R_m$. Since $A_1 A_1$ is 2-unavoidable, Proposition 1 implies R_m is 2-unavoidable for all $m \in \mathbb{N}$ by induction on m. Note that R_m is over m variables; induction also yields $|R_m| = 3(2^{m-1}) - 1$. Thus there exists a 2-unavoidable pattern over m variables with length $3(2^{m-1}) - 1$ for all $m \in \mathbb{N}$.

3. The power series approach

The following theorem was originally presented by Golod (see [12, Lemma 6.2.7]) and rewritten and proven with combinatorial terminology by Rampersad.

Theorem 1. *(See [10, Theorem 2].) Let S be a set of words over a k-letter alphabet with each word of length at least two. Suppose that for each $i \geqslant 2$, the set S contains at most c_i words of length i. If the power series expansion of*

$$B(x) := \left(1 - kx + \sum_{i \geqslant 2} c_i x^i\right)^{-1}$$

has non-negative coefficients, then there are at least $[x^n]B(x)$ words of length n over a k-letter alphabet that have no factors in S.

To count the number of words of length n avoiding a pattern p, we let S consist of all instances of p. To use Theorem 1, we require an upper bound c_i on the number of words of length i in S. The following lemma due to Bell and Goh provides a useful upper bound.

Lemma 1. *(See [1, Lemma 7].) Let $m \geqslant 1$ be an integer and p be a pattern over an alphabet $\Delta = \{A_1, \ldots, A_m\}$. Suppose that for $1 \leqslant i \leqslant m$, the variable A_i occurs $d_i \geqslant 1$ times in p. Let $k \geqslant 2$ be an integer and let Σ be a k-letter alphabet. Then for $n \geqslant 1$, the number of words of length n over Σ that are instances of the pattern p is no more than $[x^n]C(x)$, where*

$$C(x) := \sum_{i_1 \geqslant 1} \cdots \sum_{i_m \geqslant 1} k^{i_1 + \cdots + i_m} x^{d_1 i_1 + \cdots + d_m i_m}.$$

Note that this approach for counting instances of a pattern is based on the frequencies of each variable in the pattern, so it will not distinguish $AABB$ and $ABAB$, for example.

4. Derivation of the strict bounds

First we prove a technical inequality.