Further applications of a power series method for pattern avoidance

Narad Rampersad*

Department of Mathematics and Statistics
University of Winnipeg
515 Portage Avenue
Winnipeg, Manitoba R3B 2E9 (Canada)

n.rampersad@uwinnipeg.ca

Submitted: Jul 31, 2009; Accepted: Jun 10, 2011; Published: Jun 21, 2011
Mathematics Subject Classification: 68R15

Abstract

In combinatorics on words, a word w over an alphabet Σ is said to avoid a pattern p over an alphabet Δ if there is no factor x of w and no non-erasing morphism h from Δ^* to Σ^* such that $h(p) = x$. Bell and Goh have recently applied an algebraic technique due to Golod to show that for a certain wide class of patterns p there are exponentially many words of length n over a 4-letter alphabet that avoid p. We consider some further consequences of their work. In particular, we show that any pattern with k variables of length at least 4^k is avoidable on the binary alphabet. This improves an earlier bound due to Cassaigne and Roth.

1 Introduction

In combinatorics on words, the notion of an avoidable/unavoidable pattern was first introduced (independently) by Bean, Ehrenfeucht, and McNulty [1] and Zimin [22]. Let Σ and Δ be alphabets: the alphabet Δ is the *pattern alphabet* and its elements are *variables*. A *pattern* p is a non-empty word over Δ. A word w over Σ is an *instance of* p if there exists a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p) = w$. A pattern p is *avoidable* if there exists infinitely many words x over a finite alphabet such that no factor of x is an instance of p. Otherwise, p is *unavoidable*. If p is avoided by infinitely many words on an m-letter alphabet then it is said to be *m-avoidable*. The survey chapter in Lothaire [12, Chapter 3] gives a good overview of the main results concerning avoidable patterns.

*The author is supported by an NSERC Postdoctoral Fellowship.

The classical results of Thue [19, 20] established that the pattern xx is 3-avoidable and the pattern xxx is 2-avoidable. Schmidt [17] (see also [14]) proved that any binary pattern of length at least 13 is 2-avoidable; Roth [15] showed that the bound of 13 can be replaced by 6. Cassaigne [7] and Vaniček [21] (see [10]) determined exactly the set of binary patterns that are 2-avoidable.

Bean, Ehrenfeucht, and McNulty [1] and Zimin [22] characterized the avoidable patterns in general. Let us call a pattern p for which all variables occurring in p occur at least twice a *doubled pattern*. A consequence of the characterization of the avoidable patterns is that any doubled pattern is avoidable. Bell and Goh [3] proved the much stronger result that every doubled pattern is 4-avoidable. Cassaigne and Roth (see [8] or [12, Chapter 3]) proved that any pattern containing k distinct variables and having length greater than $200 \cdot 5^k$ is 2-avoidable. In this note we apply the arguments of Bell and Goh to show the following result, which improves that of Cassaigne and Roth.

Theorem 1. *Let k be a positive integer and let p be a pattern containing k distinct variables.*

(a) If p has length at least 2^k then p is 4-avoidable.

(b) If p has length at least 3^k then p is 3-avoidable.

(c) If p has length at least 4^k then p is 2-avoidable.

2 A power series approach

Rather than simply wishing to show the avoidability of a pattern p, one may wish instead to determine the number of words of length n over an m-letter alphabet that avoid p (see, for instance, Berstel's survey [4]). Brinkhuis [6] and Brandenburg [5] showed that there are exponentially many words of length n over a 3-letter alphabet that avoid the pattern xx. Similarly, Brandenburg showed that there are exponentially many words of length n over a 2-letter alphabet that avoid the pattern xxx.

As previously mentioned, Bell and Goh proved that every doubled pattern is 4-avoidable. In fact, they proved the stronger result that there are exponentially many words of length n over a 4-letter alphabet that avoid a given doubled pattern. Their main tool in obtaining this result is the following (here $[x^n]G(x)$ denotes the coefficient of x^n in the series expansion of $G(x)$).

Theorem 2 (Golod)**.** *Let S be a set of words over an m-letter alphabet, each word of length at least 2. Suppose that for each $i \geq 2$, the set S contains at most c_i words of length i. If the power series expansion of*

$$G(x) := \left(1 - mx + \sum_{i\geq 2} c_i x^i\right)^{-1} \tag{1}$$

has non-negative coefficients, then there are least $[x^n]G(x)$ words of length n over an m-letter alphabet that avoid S.

Theorem 2 is a special case of a result originally presented by Golod (see Rowen [16, Lemma 6.2.7]) in an algebraic setting. We have stated it here using combinatorial terminology. The proof given in Rowen's book also is phrased in algebraic terminology; in order to make the technique perhaps a little more accessible to combinatorialists, we present a proof of Theorem 2 using combinatorial language.

Proof of Theorem 2. For two power series $f(x) = \sum_{i\geq 0} a_i x^i$ and $g(x) = \sum_{i\geq 0} b_i x^i$, we write $f \geq g$ to mean that $a_i \geq b_i$ for all $i \geq 0$. Let $F(x) := \sum_{i\geq 0} a_i x^i$, where a_i is the number of words of length i over an m-letter alphabet that avoid S. Let $G(x) := \sum_{i\geq 0} b_i x^i$ be the power series expansion of G defined above. We wish to show $F \geq G$.

For $k \geq 1$, there are $m^k - a_k$ words w of length k over an m-letter alphabet that contain a word in S as a factor. On the other hand, for any such w either (a) $w = w'a$, where a is a single letter and w' is a word of length $k-1$ containing a word in S as a factor; or (b) $w = xy$, where x is a word of length $k-j$ that avoids S and $y \in S$ is a word of length j. There at most $(m^{k-1} - a_{k-1})m$ words w of the form (a), and there are at most $\sum_j a_{k-j} c_j$ words w of the form (b). We thus have the inequality

$$m^k - a_k \leq (m^{k-1} - a_{k-1})m + \sum_j a_{k-j} c_j.$$

Rearranging, we have

$$a_k - a_{k-1}m + \sum_j a_{k-j} c_j \geq 0, \tag{2}$$

for $k \geq 1$.

Consider the function

$$\begin{aligned} H(x) &:= F(x)\left(1 - mx + \sum_{j\geq 2} c_j x^j\right) \\ &= \left(\sum_{i\geq 0} a_i x^i\right)\left(1 - mx + \sum_{j\geq 2} c_j x^j\right). \end{aligned}$$

Observe that for $k \geq 1$, we have $[x^k]H(x) = a_k - a_{k-1}m + \sum_j a_{k-j} c_j$. By (2), we have $[x^k]H(x) \geq 0$ for $k \geq 1$. Since $[x^0]H(x) = 1$, the inequality $H \geq 1$ holds, and in particular, $H - 1$ has non-negative coefficients. We conclude that $F = HG = (H-1)G + G \geq G$, as required. □

Theorem 2 bears a certain resemblance to the Goulden–Jackson cluster method [11, Section 2.8], which also produces a formula similar to (1). The cluster method yields an exact enumeration of the words avoiding the set S but requires S to be finite. By contrast, Theorem 2 only gives a lower bound on the number of words avoiding S, but now the set S can be infinite.

Theorem 2 can be viewed as a non-constructive method to show the avoidability of patterns over an alphabet of a certain size. In this sense it is somewhat reminiscent of the probabilistic approach to pattern avoidance using the Lovász local lemma (see [2, 9]). For pattern avoidance it may even be more powerful than the local lemma in certain respects. For instance, Pegden [13] proved that doubled patterns are 22-avoidable using the local lemma, whereas Bell and Goh were able to show 4-avoidability using Theorem 2. Similarly, the reader may find it a pleasant exercise to show using Theorem 2 that there are infinitely many words avoiding xx over a 7-letter alphabet; as far as we are aware, the smallest alphabet size for which the avoidability of xx has been shown using the local lemma is 13 [18].

3 Proof of Theorem 1

To prove Theorem 1 we begin with some lemmas.

Lemma 3. *Let $k \geq 1$ and $m \geq 2$ be integers. If w is a word of length at least m^k over a k-letter alphabet, then w contains a non-empty factor w' such that the number of occurrences of each letter in w' is a multiple of m.*

Proof. Suppose w is over the alphabet $\Sigma = \{1, 2, \ldots, k\}$. Define the map $\psi : \Sigma^* \to \mathbb{N}^k$ that maps a word x to the k-tuple $[|x|_1 \bmod m, \ldots, |x|_k \bmod m]$, where $|x|_a$ denotes the number of occurrences of the letter a in x. For each prefix w_i of length i of w, let $v_i = \psi(w_i)$. Since w has length at least m^k, w has at least $m^k + 1$ prefixes, but there are at most m^k distinct tuples v_i. There exists therefore $i < j$ such that $v_i = v_j$. However, if w' is the suffix of w_j of length $j - i$, then $\psi(w') = v_j - v_i = [0, \ldots, 0]$, and hence the number of occurrences of each letter in w' is a multiple of m. ☐

Lemma 4 ([3])**.** *Let $k \geq 1$ be an integer and let p be a pattern over the pattern alphabet $\{x_1, \ldots, x_k\}$. Suppose that for $1 \leq i \leq k$, the variable x_i occurs $a_i \geq 1$ times in p. Let $m \geq 2$ be an integer and let Σ be an m-letter alphabet. Then for $n \geq 1$, the number of words of length n over Σ that are instances of the pattern p is at most $[x^n]C(x)$, where*

$$C(x) := \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1 + \cdots + i_k} x^{a_1 i_1 + \cdots + a_k i_k}.$$

For the proof of the next result, we essentially follow the approach of Bell and Goh.

Theorem 5. *Let $k \geq 2$ be an integer and let p be a pattern over a k-letter pattern alphabet such that every variable occurring in p occurs at least μ times.*

(a) If $\mu = 3$, then for $n \geq 0$, there are at least 2.94^n words of length n avoiding p over a 3-letter alphabet.

(b) If $\mu = 4$, then for $n \geq 0$, there are at least 1.94^n words of length n avoiding p over a 2-letter alphabet.

Proof. Let $(m, \mu) \in \{(3,3),(2,4)\}$ and let Σ be an m-letter alphabet. Define S to be the set of all words over Σ that are instances of the pattern p. By Lemma 4, the number of words of length n in S is at most $[x^n]C(x)$, where

$$C(x) := \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} x^{a_1 i_1+\cdots+a_k i_k},$$

and for $1 \leq i \leq k$ we have $a_i \geq \mu$. Define

$$B(x) := \sum_{i \geq 0} b_i x^i = (1 - mx + C(x))^{-1},$$

and set $\lambda := m - 0.06$ (this is not necessarily the optimal value for λ). We claim that $b_n \geq \lambda b_{n-1}$ for all $n \geq 0$. This suffices to prove the lemma, as we would then have $b_n \geq \lambda^n$ and the result follows by an application of Theorem 2.

We prove the claim by induction on n. When $n = 0$, we have $b_0 = 1$ and $b_1 = m$. Since $m > \lambda$, the inequality $b_1 \geq \lambda b_0$ holds, as required. Suppose that for all $j < n$, we have $b_j \geq \lambda b_{j-1}$. Since $B = (1 - mx + C)^{-1}$, we have $B(1 - mx + C) = 1$. Hence $[x^n]B(1 - mx + C) = 0$ for $n \geq 1$. However,

$$B(1 - mx + C) = \left(\sum_{i \geq 0} b_i x^i\right)\left(1 - mx + \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} x^{a_1 i_1+\cdots+a_k i_k}\right),$$

so

$$[x^n]B(1 - mx + C) = b_n - b_{n-1}m + \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} b_{n-(a_1 i_1+\cdots+a_k i_k)} = 0.$$

Rearranging, we obtain

$$b_n = \lambda b_{n-1} + (m - \lambda) b_{n-1} - \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} b_{n-(a_1 i_1+\cdots+a_k i_k)}.$$

To show $b_n \geq \lambda b_{n-1}$ it therefore suffices to show

$$(m - \lambda) b_{n-1} - \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} b_{n-(a_1 i_1+\cdots+a_k i_k)} \geq 0. \tag{3}$$

Since $b_j \geq \lambda b_{j-1}$ for all $j < n$, we have $b_{n-i} \leq b_{n-1}/\lambda^{i-1}$ for $1 \leq i \leq n$. Hence

$$
\begin{aligned}
\sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} b_{n-(a_1 i_1+\cdots+a_k i_k)} &\leq \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} m^{i_1+\cdots+i_k} \frac{\lambda b_{n-1}}{\lambda^{a_1 i_1+\cdots+a_k i_k}} \\
&= \lambda b_{n-1} \sum_{i_1 \geq 1} \cdots \sum_{i_k \geq 1} \frac{m^{i_1+\cdots+i_k}}{\lambda^{a_1 i_1+\cdots+a_k i_k}} \\
&= \lambda b_{n-1} \sum_{i_1 \geq 1} \frac{m^{i_1}}{\lambda^{a_1 i_1}} \cdots \sum_{i_k \geq 1} \frac{m^{i_k}}{\lambda^{a_k i_k}} \\
&\leq \lambda b_{n-1} \sum_{i_1 \geq 1} \frac{m^{i_1}}{\lambda^{\mu i_1}} \cdots \sum_{i_k \geq 1} \frac{m^{i_k}}{\lambda^{\mu i_k}} \\
&= \lambda b_{n-1} \left(\sum_{i \geq 1} \frac{m^i}{\lambda^{\mu i}} \right)^k \\
&= \lambda b_{n-1} \left(\frac{m/\lambda^\mu}{1-m/\lambda^\mu} \right)^k \\
&= \lambda b_{n-1} \left(\frac{m}{\lambda^\mu - m} \right)^k \\
&\leq \lambda b_{n-1} \left(\frac{m}{\lambda^\mu - m} \right)^2 .
\end{aligned}
$$

In order to show that (3) holds, it thus suffices to show that

$$m - \lambda \geq \lambda \left(\frac{m}{\lambda^\mu - m} \right)^2 .$$

Recall that $m - \lambda = 0.06$. For $(m, \mu) = (3, 3)$ we have

$$2.94 \left(\frac{3}{2.94^3 - 3} \right)^2 = 0.052677 \cdots \leq 0.06,$$

and for $(m, \mu) = (2, 4)$ we have

$$1.94 \left(\frac{2}{1.94^4 - 2} \right)^2 = 0.052439 \cdots \leq 0.06,$$

as required. This completes the proof of the inductive claim and the proof of the lemma. □

We can now complete the proof of Theorem 1. Let p be a pattern with k variables. If p has length at least 2^k, then by Lemma 3, the pattern p contains a non-empty factor p' such that each variable occurring in p' occurs at least twice. However, Bell and Goh showed that such a p' is 4-avoidable and hence p is 4-avoidable.

Similarly, if p has length at least 3^k (resp. 4^k), then by Lemma 3, the pattern p contains a non-empty factor p' such that each variable occurring in p' occurs at least 3 times (resp. 4 times). If p' contains only one distinct variable, recall that we have already noted in the introduction that the pattern xxx is 2-avoidable (and hence also 3-avoidable). If p' contains at least two distinct variables, then by Theorem 5, the pattern p' is 3-avoidable (resp. 2-avoidable), and hence the pattern p is 3-avoidable (resp. 2-avoidable). This completes the proof of Theorem 1.

Recall that Cassaigne and Roth showed that any pattern p over k variables of length greater than $200 \cdot 5^k$ is 2-avoidable. Their proof is constructive but is rather difficult. We are able to obtain the much better bound of 4^k non-constructively by a somewhat simpler argument. Cassaigne suggests (see the open problem [12, Problem 3.3.2]) that the bound of 3^k in Theorem 1(b) can perhaps be replaced by 2^k and that the bound of 4^k in Theorem 1(c) can perhaps be replaced by $3 \cdot 2^k$. Note that the bound of 2^k in Theorem 1(a) is optimal, since the Zimin pattern on k-variables (see [12, Chapter 3]) has length $2^k - 1$ and is unavoidable.

Acknowledgments

We thank Terry Visentin for some helpful discussions concerning Theorem 2 and the Goulden–Jackson cluster method.

References

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

[2] J. Beck, "An application of Lovász local lemma: there exists an infinite 01-sequence containing no near identical intervals", in *Infinite and Finite Sets* (A. Hajnal et al. eds.), *Colloq. Math. Soc. J. Bolyai* **37**, 1981, pp. 103–107.

[3] J. Bell, T. L. Goh, "Lower bounds for pattern avoidance", *Inform. and Comput.* **205** (2007), 1295–1306.

[4] J. Berstel, "Growth of reptition-free words—a review", *Theoret. Comput. Sci.* **340** (2005), 280–290.

[5] F.-J. Brandenburg, "Uniformly growing k-th power-free homomorphisms", *Theoret. Comput. Sci.* **23** (1983), 69–82.

[6] J. Brinkhuis, "Nonrepetitive sequences on three symbols", *Quart. J. Math. Oxford* **34** (1983), 145–149.

[7] J. Cassaigne, "Unavoidable binary patterns", *Acta Inform.* **30** (1993), 385–395.

[8] J. Cassaigne, *Motifs évitables et régularités dans les mots*, Thèse de doctorat, Université Paris 6, LITP research report TH 94-04.

[9] J. Currie, "Pattern avoidance: themes and variations", *Theoret. Comput. Sci.* **339** (2005), 7–18.

[10] P. Goralčik, T. Vaniček, "Binary patterns in binary words", *Int. J. Algebra Comput.* **1**, 387–391.

[11] I. Goulden, D. Jackson, *Combinatorial Enumeration*, Dover, 2004.

[12] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge, 2002.

[13] W. Pegden, "Highly nonrepetitive sequences: winning strategies from the Local Lemma". Manuscript available at `http://people.cs.uchicago.edu/~wes/seqgame.pdf`.

[14] N. Rampersad, "Avoiding sufficiently large binary patterns", *Bull. Europ. Assoc. Theoret. Comput. Sci.* **95** (2008), 241–245.

[15] P. Roth, "Every binary pattern of length six is avoidable on the two-letter alphabet", *Acta Inform.* **29** (1992), 95–107.

[16] L. Rowen, *Ring Theory. Vol. II*, Pure and Applied Mathematics 128, Academic Press, Boston, 1988.

[17] U. Schmidt, "Avoidable patterns on two letters", *Theoret. Comput. Sci.* **63** (1989), 1–17.

[18] J. Shallit, Unpublished lecture notes.

[19] A. Thue, "Über unendliche Zeichenreihen", *Kra. Vidensk. Selsk. Skrifter. I. Mat. Nat. Kl.* **7** (1906), 1–22.

[20] A. Thue, "Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen", *Kra. Vidensk. Selsk. Skrifter. I. Math. Nat. Kl.* **1** (1912), 1–67.

[21] T. Vaniček, *Unavoidable Words*, Diploma thesis, Charles University, Prague, 1989.

[22] A. I. Zimin, "Blocking sets of terms", *Math. USSR Sbornik* **47** (1984), 353–364.