Doubled patterns are 3-avoidable

Pascal Ochem
LIRMM, Université de Montpellier, CNRS
Montpellier, France
ochem@lirmm.fr

Submitted: June, 2015; Accepted: XX; Published: XX
Mathematics Subject Classifications: 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 f of w such that $f = h(p)$ where $h : \Delta^* \to \Sigma^*$ is a non-erasing morphism. A pattern p is said to be k-avoidable if there exists an infinite word over a k-letter alphabet that avoids p. A pattern is said to be doubled if no variable occurs only once. Doubled patterns with at most 3 variables and patterns with at least 6 variables are 3-avoidable. We show that doubled patterns with 4 and 5 variables are also 3-avoidable.

Keywords: Word; Pattern avoidance.

1 Introduction

A pattern p is a non-empty word over an alphabet $\Delta = \{A, B, C, \dots\}$ of capital letters called *variables*. An *occurrence* of p in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that $h(p)$ is a factor of w. The avoidability index $\lambda(p)$ of a pattern p is the size of the smallest alphabet Σ such that there exists an infinite word w over Σ containing no occurrence of p. Bean, Ehrenfeucht, and McNulty [2] and Zimin [13] characterized unavoidable patterns, i.e., such that $\lambda(p) = \infty$. We say that a pattern p is t-avoidable if $\lambda(p) \leqslant t$. For more informations on pattern avoidability, we refer to Chapter 3 of Lothaire's book [8].

It follows from their characterization that every unavoidable pattern contains a variable that occurs once. Equivalently, every doubled pattern is avoidable. Our result is that :

Theorem 1. *Every doubled pattern is* 3*-avoidable.*

Let $v(p)$ be the number of distinct variables of the pattern p. For $v(p) \leqslant 3$, Cassaigne [5] began and I [9] finished the determination of the avoidability index of every pattern with at most 3 variables. It implies in particular that every avoidable pattern with at most 3 variables is 3-avoidable. Moreover, Bell and Goh [3] obtained that every doubled pattern p such that $v(p) \geqslant 6$ is 3-avoidable.

Therefore, as noticed in the conclusion of [10], there remains to prove Theorem 1 for every pattern p such that $4 \leqslant v(p) \leqslant 5$. In this paper, we use both constructions of infinite words and a non-constructive method to settle the cases $4 \leqslant v(p) \leqslant 5$.

Recently, Blanchet-Sadri and Woodhouse [4] and Ochem and Pinlou [10] independently obtained the following.

Theorem 2 ([4, 10])**.** *Let p be a pattern.*

(a) If p has length at least $3 \times 2^{v(p)-1}$ then $\lambda(p) \leqslant 2$.

(b) If p has length at least $2^{v(p)}$ then $\lambda(p) \leqslant 3$.

As noticed in these papers, if p has length at least $2^{v(p)}$ then p contains a doubled pattern as a factor. Thus, Theorem 1 implies Theorem 2.(b).

2 Extending the power series method

In this section, we borrow an idea from the entropy compression method to extend the power series method as used by Bell and Goh [3], Rampersad [12], and Blanchet-Sadri and Woodhouse [4].

Let us describe the method. Let $L \subset \Sigma_m^*$ be a factorial language defined by a set F of forbidden factors of length at least 2. We denote the factor complexity of L by $n_i = L \cap \Sigma_m^i$. We define L' as the set of words w such that w is not in L and the prefix of length $|w|-1$ of w is in L. For every forbidden factor $f \in F$, we choose a number $1 \leqslant s_f \leqslant |f|$. Then, for every $i \geqslant 1$, we define an integer a_i such that

$$a_i \geqslant \max_{u \in L} \left| \left\{ v \in \Sigma_m^i \mid uv \in L',\ uv = bf,\ f \in F,\ s_f = i \right\} \right|.$$

We consider the formal power series $P(x) = 1 - mx + \sum_{i \geqslant 1} a_i x^i$. If $P(x)$ has a positive real root x_0, then $n_i \geqslant x_0^{-i}$ for every $i \geqslant 0$.

Let us rewrite that $P(x_0) = 1 - mx_0 + \sum_{i \geqslant 1} a_i x_0^i = 0$ as

$$m - \sum_{i \geqslant 1} a_i x_0^{i-1} = x_0^{-1} \tag{1}$$

Since $n_0 = 1$, we will prove by induction that $\frac{n_i}{n_{i-1}} \geqslant x_0^{-1}$ in order to obtain that $n_i \geqslant x_0^{-i}$ for every $i \geqslant 0$. By using (1), we obtain the base case: $\frac{n_1}{n_0} = n_1 = m \geqslant x_0^{-1}$. Now, for every length $i \geqslant 1$, there are:

- m^i words in Σ_m^i,
- n_i words in L,

- at most $\sum_{1\leqslant j\leqslant i} n_{i-j}a_j$ words in L',
- $m(m^{i-1}-n_{i-1})$ words in $\Sigma_m^i \setminus \{L\cup L'\}$.

This gives $n_i+\sum_{1\leqslant j\leqslant i} n_j a_{i-j}+m(m^{i-1}-n_{i-1})\geqslant m^i$, that is, $n_i\geqslant mn_{i-1}-\sum_{1\leqslant j\leqslant i} n_{i-j}a_j$.

$$\begin{array}{rcll} \frac{n_i}{n_{i-1}} & \geqslant & m-\sum_{1\leqslant j\leqslant i} a_j \frac{n_{i-j}}{n_{i-1}} & \\ & \geqslant & m-\sum_{1\leqslant j\leqslant i} a_j x_0^{j-1} & \text{By induction} \\ & \geqslant & m-\sum_{j\geqslant 1} a_j x_0^{j-1} & \\ & = & x_0^{-1} & \text{By (1)} \end{array}$$

The power series method used in previous papers [3, 4, 12] corresponds to the special case such that $s_f=|f|$ for every forbidden factor. Our condition is that $P(x)=0$ for some $x>0$ whereas the condition in these papers is that every coefficient of the series expansion of $\frac{1}{P(x)}$ is positive. The two conditions are actually equivalent. The result in [11] concerns series of the form $S(x)=1+a_1x+a_2x^2+a_3x^3+\dots$ with real coefficients such that $a_1<0$ and $a_i\geqslant 0$ for every $i\geqslant 2$. It states that every coefficient of the series $1/S(x)=b_0+b_1x+b_2x^2+b_3x^3+\dots$ is positive if and only if $S(x)$ has a positive real root x_0. Moreover, we have $b_i\geqslant x_0^{-i}$ for every $i\geqslant 0$.

The entropy compression method as developped by Gonçalves, Montassier, and Pinlou [6] uses a condition equivalent to $P(x)=0$. The benefit of the present method is that we get an exponential lower bound on the factor complexity. It is not clear whether it is possible to get such a lower bound when using entropy compression for graph coloring, since words have a simpler structure than graphs.

3 Applying the method

In this section, we show that some doubled patterns on 4 and 5 variables are 3-avoidable. Given a pattern p, every occurrence f of p is a forbidden factor. With an abuse of notation, we denote by $|A|$ the length of the image of the variable A of p in the occurrence f. This notation is used to define the length s_f.

Let us first consider doubled patterns with 4 variables. We begin with patterns of length 9, so that one variable, say A, appears 3 times. We set $s_f=|f|$. Using the obvious upper bound on the number of pattern occurrences, we obtain

$$\begin{array}{rcl} P(x) & = & 1-3x+\sum_{a,b,c,d\geqslant 1} 3^{a+b+c+d}x^{3a+2b+2c+2d} \\ & = & 1-3x+\sum_{a,b,c,d\geqslant 1} (3x^3)^a (3x^2)^b (3x^2)^c (3x^2)^d \\ & = & 1-3x+\left(\sum_{a\geqslant 1}(3x^3)^a\right)\left(\sum_{b\geqslant 1}(3x^2)^b\right)\left(\sum_{c\geqslant 1}(3x^2)^c\right)\left(\sum_{d\geqslant 1}(3x^2)^d\right) \\ & = & 1-3x+\left(\frac{1}{1-3x^3}-1\right)\left(\frac{1}{1-3x^2}-1\right)\left(\frac{1}{1-3x^2}-1\right)\left(\frac{1}{1-3x^2}-1\right) \\ & = & 1-3x+\left(\frac{1}{1-3x^3}-1\right)\left(\frac{1}{1-3x^2}-1\right)^3 \\ & = & \frac{1-3x-9x^2+24x^3+36x^4-54x^5-108x^6+243x^8+162x^9-243x^{10}}{(1-3x^3)(1-3x^2)^3}. \end{array}$$

Then $P(x)$ admits $x_0=0.3400\dots$ as its smallest positive real root. So, every doubled pattern p with 4 variables and length 9 is 3-avoidable and there exist at least $x_0^{-n}>2.941^n$

ternary words avoiding p. Notice that for patterns with 4 variables and length at least 10, every term of $\sum_{a,b,c,d\geqslant 1} 3^{a+b+c+d}x^{3a+2b+2c+2d}$ in $P(x)$ gets multiplied by some positive power of x. Since $0 < x < 1$, every term is now smaller than in the previous case. So $P(x)$ admits a smallest positive real root that is smaller than $0.3400\ldots$ Thus, these patterns are also 3-avoidable.

Now, we consider patterns with length 8, so that every variable appears exactly twice. If such a pattern has $ABCD$ as a prefix, then we set $s_f = \frac{|f|}{2} = |A| + |B| + |C| + |D|$. So we obtain $P(x) = 1 - 3x + \sum_{a,b,c,d\geqslant 1} x^{a+b+c+d} = 1 - 3x + \left(\frac{1}{1-x} - 1\right)^4$. Then $P(x)$ admits $0.3819\ldots$ as its smallest positive real root, so that this pattern is 3-avoidable.

Among the remaining patterns, we rule out patterns containing an occurrence of a doubled pattern with at most 3 variables. Also, if one pattern is the reverse of another, then they have the same avoidability index and we consider only one of the two. Thus, there remain the following patterns: $ABACBDCD$, $ABACDBDC$, $ABACDCBD$, $ABCADBDC$, $ABCADCBD$, $ABCADCDB$, and $ABCBDADC$.

Now we consider doubled patterns with 5 variables. Similarly, we rule out every pattern of length at least 11 with the method by setting $s_f = |f|$. Then we check that $P(x) = 1 - 3x + \sum_{a,b,c,d,e\geqslant 1} 3^{a+b+c+d+e}x^{3a+2b+2c+2d+2e} = 1 - 3x + \left(\frac{1}{1-3x^3} - 1\right)\left(\frac{1}{1-3x^2} - 1\right)^4$ has a positive real root.

We also rule out every pattern of length 10 having ABC as a prefix. We set $s_f = |f| - |ABC| = |A| + |B| + |C| + 2|D| + 2|E|$. Then we check that $P(x) = 1 - 3x + \sum_{a,b,c,d,e\geqslant 1} 3^{d+e}x^{a+b+c+2d+2e} = 1 - 3x + \left(\frac{1}{1-x} - 1\right)^3\left(\frac{1}{1-3x^2} - 1\right)^2$ has a positive real root.

Again, we rule out patterns containing an occurrence of a doubled pattern with at most 4 variables and patterns whose reversed pattern is already considered. Thus, there remain the following patterns: $ABACBDCEDE$, $ABACDBCEDE$, and $ABACDBDECE$.

4 Sporadic doubled patterns

In this section, we consider the 10 doubled patterns on 4 and 5 variables whose 3-avoidability has not been obtained in the previous section.

We define the *avoidability exponent* $AE(p)$ of a pattern p as the largest real x such that every x-free word avoids p. This notion is not pertinent e.g. for the pattern $ABWBAXACYCAZBC$ studied by Baker, McNulty, and Taylor [1], since for every $\epsilon > 0$, there exists a $(1+\epsilon)$-free word containing an occurrence of that pattern. However, $AE(p) > 1$ for every doubled pattern. To see that, consider a factor $A\ldots A$ of p. If an x-free word contains an occurrence of p, then the image of this factor is a repetition such that the image of A cannot be too large compared to the images of the variables occurring between the As in p. We have similar constraints for every variable and this set of constraints becomes unsatisfiable as x decreases towards 1. We present one way of obtaining the avoidability exponent for a doubled pattern p of length $2v(p)$. We construct the $v(p) \times v(p)$ matrix M such that $M_{i,j}$ is the number of occurrences of the variable X_j between the two occurrences of the variable X_i. We compute the largest eigenvalue β of M and then we have $AE(p) = 1 + \frac{1}{\beta+1}$. For example if $p = ABACDCBD$, then we get $M = \begin{bmatrix} 0&1&0&0 \\ 1&0&0&1 \\ 0&2&0&1 \\ 0&1&1&0 \end{bmatrix}$, $\beta = 1.9403\ldots$, and $AE(p) = 1 + \frac{1}{\beta+1} = 1.3400\ldots$. The avoidability exponents of the 10 patterns considered in this section range from $AE(ABCADBDC) = 1.292893219$ to $AE(ABACBDCD) = 1.381966011$. For each pattern p among the 10, we give a uniform morphism $m : \Sigma_5^* \to \Sigma_2^*$ such that for every $\left(\frac{5}{4}^+\right)$-free word $w \in \Sigma_5^*$, we have that $m(w)$ avoids p. The proof that p is avoided follows the method in [9]. Since there exist exponentially many $\left(\frac{5}{4}^+\right)$-free words over Σ_5 [7], there exist exponentially many binary words avoiding p.

- $AE(ABACBDCD) = 1.381966011$, 17-uniform morphism

$$\begin{aligned} 0 &\mapsto 00000111101010110 \\ 1 &\mapsto 00000110100100110 \\ 2 &\mapsto 00000011100110111 \\ 3 &\mapsto 00000011010101111 \\ 4 &\mapsto 00000011001001011 \end{aligned}$$

- $AE(ABACDBDC) = 1.333333333$, 33-uniform morphism

$$\begin{aligned} 0 &\mapsto 000000101101000111111011001010111 \\ 1 &\mapsto 000000100110100001111101001010111 \\ 2 &\mapsto 000000010110100001111111010010111 \\ 3 &\mapsto 000000010011010100011111010010111 \\ 4 &\mapsto 000000010011001000001111010010111 \end{aligned}$$

- $AE(ABACDCBD) = 1.340090632$, 28-uniform morphism

$$\begin{aligned} 0 &\mapsto 0000101010001110010000111111 \\ 1 &\mapsto 0000001111010001101001111111 \\ 2 &\mapsto 0000001101000011110100100111 \\ 3 &\mapsto 0000001011110000110100111111 \\ 4 &\mapsto 0000001010111100100001111111 \end{aligned}$$

- $AE(ABCADBDC) = 1.292893219$, 21-uniform morphism

$$\begin{aligned} 0 &\mapsto 000011101101011111010 \\ 1 &\mapsto 000010110100100111101 \\ 2 &\mapsto 000001101110100101111 \\ 3 &\mapsto 000001101011001111111 \\ 4 &\mapsto 000000110111010111111 \end{aligned}$$

- $AE(ABCADCBD) = 1.295597743$, 22-uniform morphism

$$\begin{aligned}
0 &\mapsto 0000011011010100011111\\
1 &\mapsto 0000011010101001001111\\
2 &\mapsto 0000001101100100111111\\
3 &\mapsto 0000001010110000111111\\
4 &\mapsto 0000000110101001110111
\end{aligned}$$

- $AE(ABCADCDB) = 1.327621756$, 26-uniform morphism

$$\begin{aligned}
0 &\mapsto 00000011110010101011000111\\
1 &\mapsto 00000011010111111001011011\\
2 &\mapsto 00000010011111101001110111\\
3 &\mapsto 00000001001111110001010111\\
4 &\mapsto 00000001000111111001010111
\end{aligned}$$

- $AE(ABCBDADC) = 1.302775638$, 33-uniform morphism

$$\begin{aligned}
0 &\mapsto 000000101111110011000110011111101\\
1 &\mapsto 000000101111001000001100111111101\\
2 &\mapsto 000000011011111001100000100111101\\
3 &\mapsto 000000011010101011000001001111101\\
4 &\mapsto 000000010111110010101010011111011
\end{aligned}$$

- $AE(ABACBDCEDE) = 1.366025404$, 15-uniform morphism

$$\begin{aligned}
0 &\mapsto 001011011110000\\
1 &\mapsto 001010100111111\\
2 &\mapsto 000110010011000\\
3 &\mapsto 000011111111100\\
4 &\mapsto 000011010101110
\end{aligned}$$

- $AE(ABACDBCEDE) = 1.302775638$, 18-uniform morphism

$$\begin{aligned}
0 &\mapsto 000010110100100111\\
1 &\mapsto 000010100111111111\\
2 &\mapsto 000000110110011111\\
3 &\mapsto 000000101010101111\\
4 &\mapsto 000000000111100111
\end{aligned}$$

- $AE(ABACDBDECE) = 1.320416579$, 22-uniform morphism

$$\begin{aligned}
0 &\mapsto 0000001111110001011011\\
1 &\mapsto 0000001111100100110101\\
2 &\mapsto 0000001111100001101101\\
3 &\mapsto 0000001111001001011100\\
4 &\mapsto 0000001111000010101100
\end{aligned}$$

5 Simultaneous avoidance of doubled patterns

Bell and Goh [3] have also considered the avoidance of multiple patterns simultaneously and ask (question 3) whether there exist an infinite word over a finite alphabet that avoids every doubled pattern. We give a negative answer.

A word w is *n-splitted* if $|w| \equiv 0 \pmod{n}$ and every factor w_i such that $w = w_1w_2\ldots w_n$ and $|w_i| = \frac{|w|}{n}$ for $1 \leqslant i \leqslant n$ contains every letter in w. An n-splitted pattern is defined similarly. Let us prove by induction on k that every word $w \in \Sigma_k^{n^k}$ contains an n-splitted factor. The assertion is true for $k = 1$. Now, if the word $w \in \Sigma_k^{n^k}$ is not itself n-splitted, then by definition it must contain a factor w_i that does not contain every letter of w. So we have $w_i \in \Sigma_{k-1}^{n^{k-1}}$. By induction, w_i contains an n-splitted factor, and so does w.

This implies that for every fixed n, every infinite word over a finite alphabet contains n-splitted factors. Moreover, an n-splitted word is an occurrence of an n-splitted pattern such that every variable has a distinct image of length 1. So, for every fixed n, the set of all n-splitted patterns is not avoidable by an infinite word over a finite alphabet.

Notice that if $n \geqslant 2$, then an n-splitted word (resp. pattern) contains a 2-splitted word (resp. pattern) and a 2-splitted word (resp. pattern) is doubled.

6 Conclusion

Our results answer settles the first of two questions of our previous paper [10]. The second question is whether there exists a finite k such that every doubled pattern with at least k variables is 2-avoidable. As already noticed [10], such a k is at least 5 since, e.g., $ABCCBADD$ is not 2-avoidable.

Acknowledgments

I am grateful to Narad Rampersad for comments on a draft of the paper, to Vladimir Dotsenko for pointing out the result in [11], and to Andrei Romashchenko for translating this paper.

References

[1] K.A. Baker, G.F. McNulty, and W. Taylor. Growth problems for avoidable words, *Theoret. Comput. Sci.* **69** (1989), 319–345.

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

[3] J. Bell, T. L. Goh. Exponential lower bounds for the number of words of uniform length avoiding a pattern. *Inform. and Comput.* **205** (2007), 1295-1306.

[4] F. Blanchet-Sadri, B. Woodhouse. Strict bounds for pattern avoidance. *Theor. Comput. Sci.* **506** (2013), 17–27.

[5] J. Cassaigne. Motifs évitables et régularité dans les mots. Thèse de Doctorat, Université Paris VI, Juillet 1994.

[6] D. Gonçalves, M. Montassier, and A. Pinlou. Entropy compression method applied to graph colorings. arXiv:1406.4380

[7] R. Kolpakov and M. Rao: On the number of Dejan words over alphabets of 5, 6, 7, 8, 9 and 10 letters *Theor. Comput. Sci.* **412(46)** (2011), 6507–6516.

[8] M. Lothaire. Algebraic Combinatorics on Words. *Cambridge Univ. Press* (2002).

[9] P. Ochem. A generator of morphisms for infinite words. *RAIRO: Theoret. Informatics Appl.* **40** (2006) 427–441.

[10] P. Ochem and A. Pinlou. Application of entropy compression in pattern avoidance. *Electron. J. Combinatorics.* **21(2)** (2014), #RP2.7.

[11] D. I. Piotkovskii. On the growth of graded algebras with a small number of defining relations. *Uspekhi Mat. Nauk.* **48:3(291)** (1993), 199–200.

[12] N. Rampersad. Further applications of a power series method for pattern avoidance. *Electron. J. Combinatorics.* **18(1)** (2011), #P134.

[13] A.I. Zimin. Blocking sets of terms. *Math. USSR Sbornik* **47(2)** (1984) 353–364. English translation.