Avoidability of circular formulas

Guilhem Gamard[a], Pascal Ochem[a,b], Gwenaël Richomme[a,c], Patrice Séébold[a,c]

[a] *LIRMM, Université de Montpellier and CNRS, France*
[b] *CNRS*
[c] *Université Paul-Valéry Montpellier 3*

Abstract

Clark has defined the notion of n-avoidance basis which contains the avoidable formulas with at most n variables that are closest to be unavoidable in some sense. The family C_i of circular formulas is such that $C_1 = AA$, $C_2 = ABA.BAB$, $C_3 = ABCA.BCAB.CABC$ and so on. For every $i \leqslant n$, the n-avoidance basis contains C_i. Clark showed that the avoidability index of every circular formula and of every formula in the 3-avoidance basis (and thus of every avoidable formula containing at most 3 variables) is at most 4. We determine exactly the avoidability index of these formulas.

1. Introduction

A *pattern* p is a non-empty finite word over an alphabet $\Delta = \{A, B, C, \ldots\}$ 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 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 information on pattern avoidability, we refer to Chapter 3 of Lothaire's book [8]. See also this book for basic notions in Combinatorics on Words.

A variable that appears only once in a pattern is said to be *isolated*. Following Cassaigne [3], we associate to a pattern p the *formula* f obtained by replacing every isolated variable in p by a dot. The factors between the dots are called *fragments*.

An *occurrence* of a formula f in a word w is a non-erasing morphism $h : \Delta^* \to \Sigma^*$ such that the h-image of every fragment of f is a factor of w. As for patterns, the avoidability index $\lambda(f)$ of a formula f is the size of the smallest alphabet allowing the existence of an infinite word containing no occurrence of f. Clearly, if a formula f is associated to a pattern p, every word avoiding f also avoids p, so $\lambda(p) \leqslant \lambda(f)$. Recall that an infinite word is *recurrent* if every finite factor appears infinitely many times. If there exists an infinite word over Σ avoiding p, then there exists an infinite recurrent word over Σ avoiding p. This recurrent word also avoids f, so that $\lambda(p) = \lambda(f)$. Without loss of generality,

a formula is such that no variable is isolated and no fragment is a factor of another fragment.

Cassaigne [3] began and Ochem [9] finished the determination of the avoidability index of every pattern with at most 3 variables. A *doubled* pattern contains every variable at least twice. Thus, a doubled pattern is a formula with exactly one fragment. Every doubled pattern is 3-avoidable [10]. A formula is said to be *binary* if it has at most 2 variables. The avoidability index of every binary formula has been recently determined [11]. We say that a formula f is *divisible* by a formula f' if f does not avoid f', that is, there is a non-erasing morphism h such that the image of every fragment of f' by h is a factor of a fragment of f. If f is divisible by f', then every word avoiding f' also avoids f and thus $\lambda(f) \leqslant \lambda(f')$. Moreover, the reverse f^R of a formula f satisfies $\lambda(f^R) = \lambda(f)$. For example, the fact that $ABA.AABB$ is 2-avoidable implies that $ABAABB$ and $BAB.AABB$ are 2-avoidable. See Cassaigne [3] and Clark [4] for more information on formulas and divisibility.

Clark [4] has introduced the notion of *n-avoidance basis* for formulas, which is the smallest set of formulas with the following property: for every $i \leqslant n$, every avoidable formula with i variables is divisible by at least one formula with at most i variables in the n-avoidance basis.

From the definition, it is not hard to obtain that the 1-avoidance basis is $\{AA\}$ and the 2-avoidance basis is $\{AA, ABA.BAB\}$. Clark obtained that the 3-avoidance basis is composed of the following formulas:

- AA
- $ABA.BAB$
- $ABCA.BCAB.CABC$
- $ABCBA.CBABC$
- $ABCA.CABC.BCB$
- $ABCA.BCAB.CBC$
- $AB.AC.BA.CA.CB$

The following properties of the avoidance basis are derived.

- The n-avoidance basis is a subset of the $(n+1)$-avoidance basis.
- The n-avoidance basis is closed under reverse. (In particular, $ABCA.BCAB.CBC$ is the reverse of $ABCA.CABC.BCB$.)
- Two formulas in the n-avoidance basis with the same number of variables are incomparable by divisibility. (However, AA is divisible $AB.AC.BA.CA.CB$.)
- The n-avoidance basis is computable.

The *circular formula* C_t is the formula over $t \geqslant 1$ variables $A_0, \ldots, A_{t-1}$ containing the t fragments of the form $A_i A_{i+1} \ldots A_{i+t}$ such that the indices are taken modulo t. Thus, the first three formulas in the 3-avoidance basis, namely $C_1 = AA$, $C_2 = ABA.BAB$, and $C_3 = ABCA.BCAB.CABC$, are also the first three circular formulas. More generally, for every $t \leqslant n$, the n-avoidance basis contains C_t.

It is known that $\lambda(AA) = 3$ [12], $\lambda(ABA.BAB) = 3$ [3], and $\lambda(AB.AC.BA.CA.CB) = 4$ [1]. Actually, $AB.AC.BA.CA.CB$ is avoided by the fixed point $b_4 = 0121032101230321\ldots$ of the morphism given below.

$$\begin{aligned} 0 &\mapsto 01 \\ 1 &\mapsto 21 \\ 2 &\mapsto 03 \\ 3 &\mapsto 23 \end{aligned}$$

Clark [4] obtained that b_4 also avoids C_i for every $i \geqslant 1$, so that $\lambda(C_i) \leqslant 4$ for every $i \geqslant 1$. He also showed that the avoidability index of the other formulas in the 3-avoidance basis is at most 4. Our main results finish the determination of the avoidability index of the circular formulas (Theorem 1) and the formulas in the 3-avoidance basis (Theorem 4).

2. Conjugacy classes and circular formulas

In this section, we determine the avoidability index of circular formulas.

Theorem 1. *$\lambda(C_3) = 3$. $\forall i \geqslant 4$, $\lambda(C_i) = 2$.*

We consider a notion that appears to be useful in the study of circular formulas. A *conjugacy class* is the set of all the conjugates of a given word, including the word itself. The length of a conjugacy class is the common length of the words in the conjugacy class. A word contains a conjugacy class if it contains every word in the conjugacy class as a factor. Consider the uniform morphisms given below.

$$\begin{aligned} g_2(0) &= 0000101001110110100 \\ g_2(1) &= 0011100010100111101 \\ g_2(2) &= 0000111000010110100 \\ g_2(3) &= 0011110110100111101 \end{aligned} \qquad \begin{aligned} g_3(0) &= 0010 \\ g_3(1) &= 1122 \\ g_3(2) &= 0200 \\ g_3(3) &= 1212 \end{aligned} \qquad \begin{aligned} g_6(0) &= 01230 \\ g_6(1) &= 24134 \\ g_6(2) &= 52340 \\ g_6(3) &= 24513 \end{aligned}$$

Lemma 2.

- *The word $g_2(b_4)$ avoids every conjugacy class of length at least 5.*
- *The word $g_3(b_4)$ avoids every conjugacy class of length at least 3.*
- *The word $g_6(b_4)$ avoids every conjugacy class of length at least 2.*

Proof. We only detail the proof for $g_2(b_4)$, since the proofs for $g_3(b_4)$ and $g_6(b_4)$ are similar. Notice that g_2 is 19-uniform. First, a computer check shows that $g_2(b_4)$ contains no conjugacy class of length i with $5 \leqslant i \leqslant 55$ (i.e., $2 \times 19 + 17$).

Suppose for contradiction that $g_2(b_4)$ contains a conjugacy class of length at least 56 (i.e., $2 \times 19 + 18$). Then every element of the conjugacy class contains a factor $g_2(ab)$ with $a, b \in \Sigma_4$. In particular, one of the elements of the conjugacy class can be written as $g_2(ab)s$. The word $g_2(b)sg_2(a)$ is also a factor of $g_2(b_4)$. A computer check shows that for every letters α, β, and γ in Σ_4 such that $g_2(\alpha)$ is a factor of $g_2(\beta\gamma)$, $g_2(\alpha)$ is either a prefix or a suffix of $g_2(\beta\gamma)$. This implies that s belongs to $g_2(\Sigma_4^+)$.

Thus, the conjugacy class contains a word $w = g_2(\ell_1\ell_2\dots\ell_k) = x_1x_2...x_{19k}$. Consider the conjugate $\tilde{w} = x_7x_8\dots x_{19k}x_1x_2x_3x_4x_5x_6$. Observe that the prefixes of length 6 of $g_2(0)$, $g_2(1)$, $g_2(2)$, and $g_2(3)$ are different. Also, the suffixes of length 12 of $g_2(0)$, $g_2(1)$, $g_2(2)$, and $g_2(3)$ are different. Then the prefix $x_7\dots x_{19}$ and the suffix $x_1\dots x_6$ of $\tilde{w}$ both force the letter ℓ_1 in the pre-image. That is, b_4 contains $\ell_1\ell_2\dots\ell_k\ell_1$. Similarly, the conjugate of w that starts with the letter $x_{19(r-1)+7}$ implies that b_4 contains $\ell_r\dots\ell_k\ell_1\dots\ell_r$. Thus, b_4 contains an occurrence of the formula C_k. This is a contradiction since Clark [4] has shown that b_4 avoids every circular formula C_i with $i \geqslant 1$. □

Notice that if a word contains an occurrence of C_i, then it contains a conjugacy class of length at least i. Thus, a word avoiding every conjugacy class of length at least i also avoids every circular formula C_t with $t \geqslant i$. Moreover, $g_2(b_4)$ contains no occurrence of C_4 such that the length of the image of every variable is 1. By Lemma 2, this gives the next result, which proves Theorem 1.

Corollary 3. *The word $g_3(b_4)$ avoids every circular formula C_i with $i \geqslant 3$. The word $g_2(b_4)$ avoids every circular formula C_i with $i \geqslant 4$.*

3. Remaining formulas in the 3-avoidance basis

In this section, we prove the following result which completes the determination of the avoidability index of the formulas in the 3-avoidance basis.

Theorem 4. $\lambda(ABCBA.CBABC) = 2$. $\lambda(ABCA.CABC.BCB) = 3$.

Notice that $\lambda(ABCBA.CBABC) = 2$ implies the well-known fact that $\lambda(ABABA) = 2$. It also implies that $\lambda(ABCBABC) = 2$, which was first obtained in [6].

For both formulas, we give a uniform morphism m such that for every $\left(\frac{5}{4}^+\right)$-free word $w \in \Sigma_5^*$, the word $m(w)$ avoids the formula. Since there exist exponentially many $\left(\frac{5}{4}^+\right)$-free words over Σ_5 [7], there exist exponentially many words avoiding the formula. The proof that the formula is avoided follows the method in [9].

To avoid $ABCBA.CBABC$, we use this 15-uniform morphism:

$$\begin{aligned}
m_{15}(0) &= 001111010010110\\
m_{15}(1) &= 001110100101110\\
m_{15}(2) &= 001101001011110\\
m_{15}(3) &= 000111010001011\\
m_{15}(4) &= 000110100001011
\end{aligned}$$

First, we show that the m_{15}-image of every $\left(\frac{5}{4}^+\right)$-free word w is $\left(\frac{97}{75}^+, 61\right)$-free, that is, $m_{15}(w)$ contains no repetition with period at least 61 and exponent strictly greater than $\frac{97}{75}$. By Lemma 2.1 in [9], it is sufficient to check this property for every $\left(\frac{5}{4}^+\right)$-free word w such that $|w| < \frac{2\times\frac{97}{75}}{\frac{97}{75}-\frac{5}{4}} < 60$. Consider a potential occurrence h of $ABCBA.CBABC$ and write $a = |h(A)|$, $b = |h(B)|$, $c = |h(C)|$. Suppose that $a + b \geqslant 61$. The factor $h(BAB)$ is then a repetition with period $a + b \geqslant 61$, so that its exponent satisfies $\frac{a+2b}{a+b} \leqslant \frac{97}{75}$. This gives $53b \leqslant 22a$. Similarly, BCB implies $53b \leqslant 22c$, $ABCBA$ implies $53a \leqslant 22(2b+c)$, and $CBABC$ implies $53c \leqslant 22(a + 2b)$. Summing up these inequalities gives $53a + 106b + 53c \leqslant 44a + 88b + 44c$, which is a contradiction. Thus, we have $a + b \leqslant 60$. By symmetry, we also have $b + c \leqslant 60$. Using these inequalities, we check exhaustively that $h(w)$ contains no occurrence of $ABCBA.CBABC$.

To avoid $ABCA.CABC.BCB$ and its reverse $ABCA.BCAB.CBC$ simultaneously, we use this 6-uniform morphism:

$$\begin{aligned} m_6(0) &= 021210 \\ m_6(1) &= 012220 \\ m_6(2) &= 012111 \\ m_6(3) &= 002221 \\ m_6(4) &= 001112 \end{aligned}$$

We check that the m_6-image of every $\left(\frac{5}{4}^+\right)$-free word w is $\left(\frac{13}{10}^+, 25\right)$-free. By Lemma 2.1 in [9], it is sufficient to check this property for $\left(\frac{5}{4}^+\right)$-free word w such that $|w| < \frac{2\times\frac{13}{10}}{\frac{13}{10}-\frac{5}{4}} = 52$.

Let us consider the formula $ABCA.CABC.BCB$. Suppose that $b + c \geqslant 25$. Then $ABCA$ implies $7a \leqslant 3(b + c)$, $CABC$ implies $7c \leqslant 3(a + b)$, and BCB implies $7b \leqslant 3c$. Summing up these inequalities gives $7a+7b+7c \leqslant 3a+6b+6c$, which is a contradiction. Thus $b + c \leqslant 24$. Suppose that $a \geqslant 23$. Then $ABCA$ implies $a \leqslant \frac{3}{7}(b + c) \leqslant \frac{72}{7} < 23$, which is a contradiction. Thus $a \leqslant 22$. For the formula $ABCA.BCAB.CBC$, the same argument holds except that the roles of B and C are switched, so that we also obtain $b + c \leqslant 24$ and $a \leqslant 22$. Then we check exhaustively that $h(w)$ contains no occurrence of $ABCA.CABC.BCB$ and no occurrence of $ABCA.BCAB.CBC$.

It can be noticed that arguments using repetition to forbid patterns has also been used in [5]

4. Concluding remarks

A major open question is whether there exist avoidable formulas with arbitrarily large avoidability index. If such formulas exist, some of them necessarily belong to the n-avoidance basis for increasing values of n. With the example of circular formulas, Clark noticed that belonging to the n-avoidance basis

and having many variables does not imply a large avoidability index. Our results strengthen this remark and show that the n-avoidance basis contains a 2-avoidable formula on t variables for every $3 \leqslant t \leqslant n$.

A formula f is *nice* if for every variable X of f there exists a fragment of f that contains X at least twice. This notion generalizes the notion of doubled pattern, which corresponds to a nice formula with one fragment. Notice that every formula in the 3-avoidance basis is nice except $AB.AC.BA.CA.CB$. Thus, our results imply that the nice formulas in the 3-avoidance basis are 3-avoidable. Is every nice formula 3-avoidable?

Concerning conjugacy classes, we propose the following conjecture:

Conjecture 5. *There exists an infinite word in Σ_5^* that avoids every conjugacy class of length at least 2.*

Associated to the results in Lemma 2, this would give the smallest alphabet that allows to avoid every conjugacy class of length at least i, for every i.

References

References

[1] K. A. Baker, G. F. McNulty, and W. Taylor. Growth problems for avoidable words. *Theoret. Comput. Sci.*, 69(3):319–345, 1989.

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

[3] J. Cassaigne. *Motifs évitables et régularité dans les mots.* PhD thesis, Université Paris VI, 1994.

[4] R. J. Clark. *Avoidable formulas in combinatorics on words.* PhD thesis, University of California, Los Angeles, 2001. Available at http://www.lirmm.fr/~ochem/morphisms/clark_thesis.pdf

[5] J.D. Currie and V. Linek. Avoiding Patterns in the Abelian Sense. *Canadian J. Math.*, 53:696–714, 2001.

[6] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of repetition threshold. *Theoret. Comput. Sci.*, 92(2):71–76, 2004.

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

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

[9] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theor. Inform. Appl.*, 40:427–441, 2006.

[10] P. Ochem. Doubled patterns are 3-avoidable. *Electron. J. Combin.*, **23(1)** (2016), #P1.19.

[11] P. Ochem and M. Rosenfeld. Avoidability of formulas with two variables. *Electron. J. Combin.* **24(4)** (2017), #P4.30.

[12] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7:1–22, 1906.

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