Avoidability of palindrome patterns

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

Matthieu Rosenfeld
LIP, ENS de Lyon, CNRS, UCBL
Université de Lyon
France
matthieu.rosenfeld@ens-lyon.fr

Submitted: May 17, 2020; Accepted: Dec 17, 2020; Published: Jan 15, 2021

Abstract

We characterize the formulas that are avoided by every α-free word for some $\alpha > 1$. We show that the avoidable formulas whose fragments are of the form XY or XYX are 4-avoidable. The largest avoidability index of an avoidable palindrome pattern is known to be at least 4 and at most 16. We make progress toward the conjecture that every avoidable palindrome pattern is 4-avoidable.

Mathematics Subject Classifications: 68R15

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 (a morphism is *non-erasing* if the image of every letter is non-empty). 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. Since there is no risk of confusion, $\lambda(p)$ will be simply called the index of p.

A variable that appears only once in a pattern is said to be *isolated*. Following Cassaigne [5], we associate a pattern p with 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 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 with a pattern p, every

*The authors were partially supported by the ANR project CoCoGro (ANR-16-CE40-0005).

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 and that any infinite factorial language contains a recurrent word [8, Proposition 5.1.13]. 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.

Let us define the types of formulas we consider in this paper. A pattern is *doubled* if it contains every variable at least twice. Thus it is a formula with only one pattern. A formula f is *nice* if for every variable X of f, there exists a fragment of f that contains X at least twice. Notice that a doubled pattern is a nice pattern. A formula is an *xyx-formula* if every fragment is of the form XYX, i.e., the fragment has length 3 and the first and third variable are the same. A formula is *hybrid* if every fragment has length 2 or is of the form XYX. Thus, an xyx-formula is a hybrid formula.

In Section 3, we consider the avoidance of nice formulas. In Section 4, we find some formulas f such that every recurrent word avoiding f over $\Sigma_{\lambda(f)}$ is equivalent to a well-known morphic word. In Section 5, we consider the avoidance of xyx-formulas and hybrid formulas. In Section 6, we consider the avoidance of patterns that are palindromes.

2 Preliminaries

Given a pattern p, the Zimin operator constructs the pattern $Z(p) = pXp$ where X is a variable that is not contained in p. For every fixed t, $Z^t(p)$ denotes the pattern obtained by applying t times the Zimin operator to p. Notice that a recurrent word avoids $Z^t(p)$ if and only if it avoids p.

We say that a formula f *divides* a formula f' if every recurrent word avoiding f also avoids f'. We denote by $f \preceq f'$ the fact that f divides f'. By previous discussion, $p \preceq Z^t(p)$ and $Z^t(p) \preceq p$ for every pattern p. The basic case of divisibility is that $f \preceq f'$ if f' contains an occurrence f, that is, if there exists a non-erasing morphism h such that the h-image of every fragment of f is a factor of a fragment of f'. Another case of divisibility obtained by transitivity: in order to obtain $f \preceq p$, it is sufficient to prove $f \preceq Z^t(p)$, since $Z^t(p) \preceq p$. We use this trick in the proof of Lemma 6 and Theorem 17. Of course, divisibility is related to avoidability: if $f \preceq f'$, then $\lambda(f) \geqslant \lambda(f')$.

Let $\Sigma_k = \{0, 1, \ldots, k-1\}$ denote the k-letter alphabet. We denote by Σ_k^n the k^n words of length n over Σ_k.

The operation of *splitting* a formula f on a fragment ϕ consists in replacing ϕ by two fragments, namely the prefix and the suffix of length $|\phi|-1$ of ϕ. A formula f is *minimally avoidable* if splitting any fragment of f gives an unavoidable formula. The set of every minimally avoidable formula with at most n variables is called the n-avoidance basis.

The *adjacency graph* $AG(f)$ of the formula f is the bipartite graph such that

- for every variable X of f, $AG(f)$ contains the two vertices X_L and X_R,
- for every (possibly equal) variables X and Y, there is an edge between X_L and Y_R if and only if XY is a factor of f.

We say that a set S of variables of f is *free* if for all $X, Y \in S$, X_L and Y_R are in distinct connected components of $AG(f)$. A formula f is said to reduce to f' if it is obtained by deleting all the variables of a free set from f, discarding any empty word fragment. A formula is *reducible* if there is a sequence of reductions to the empty formula. Finally, a *locked* formula is a formula having no free set.

Theorem 1 ([3]). *A formula is unavoidable if and only if it is reducible.*

Let us define here the following well-known pure morphic words. To specify a morphism $m : \Sigma_s \to \Sigma_e$, we use the notation $m = m(\mathtt{0})/m(\mathtt{1})/\cdots/m(s-1)$. Assuming a morphism $m : \Sigma_s \to \Sigma_s$ is such that $m(\mathtt{0})$ starts with $\mathtt{0}$, the *fixed point* of m is the right infinite word $m^\omega(\mathtt{0})$.

- b_2 is the fixed point of 01/10.
- b_3 is the fixed point of 012/02/1.
- b_4 is the fixed point of 01/03/21/23.
- b_5 is the fixed point of 01/23/4/21/0

We also consider the morphic words $v_3 = M_1(b_5)$ and $w_3 = M_2(b_5)$, where $M_1 =$ 012/1/02/12/ε and $M_2 =$ 02/1/0/12/ε. The languages of each of these words have been studied in the literature. Let us first recall the following characterization of b_3, v_3, and w_3. We say that two infinite words are *equivalent* if they have the same set of factors.

Theorem 2 ([1, 16]).

- *Every ternary square-free recurrent word avoiding 010 and 212 is equivalent to b_3.*
- *Every ternary square-free recurrent word avoiding 010 and 020 is equivalent to v_3.*
- *Every ternary square-free recurrent word avoiding 121 and 212 is equivalent to w_3.*

Interestingly, these three words can be characterized in terms of a forbidden distance between consecutive occurrences of one letter.

Theorem 3.

- *Every ternary square-free recurrent word such that the distance between consecutive occurrences of 1 is not 3 is equivalent to b_3.*
- *Every ternary square-free recurrent word such that the distance between consecutive occurrences of 0 is not 2 is equivalent to v_3.*
- *Every ternary square-free recurrent word such that the distance between consecutive occurrences of 0 is not 4 is equivalent to w_3.*

Proof.

- Another characterization for b_3 is that every ternary square-free recurrent word avoiding 1021 and 1021 is equivalent to b_3 [1]. This rules out the possibility that the distance between two occurrences of 1 is 3.

- Since v_3 avoids 010 and 020, the distance between two occurrences of 0 is at least 3.

- Since w_3 avoids 121 and 212, the distance between consecutive occurrences of 0 is at most 3. □

The word b_4 is also known to avoid large families of formulas.

Theorem 4 ([2])**.** *Every locked formula is avoided by* b_4*.*

Theorem 5 ([5, Proposition 1.13])**.** *If every fragment of an avoidable formula* f *has length 2, then* b_4 *avoids* f*.*

Theorem 5 will be extended to hybrid formulas, see Theorem 21 in Section 5.

Let us give here a result that will be needed in various parts of the paper.

Lemma 6. $ABA.ACA.ABCA.ACBA.ABCBA \preceq AA$.

Proof. Indeed, $Z^2(AA) = AABAACAABAA$ contains the occurrence $A \to A$, $B \to ABA$, $C \to ACA$ of $ABA.ACA.ABCA.ACBA.ABCBA$. □

Thus, if w is a recurrent word that avoids a formula dividing $ABA.ACA.ABCA.ACBA.ABCBA$, then w is square-free.

Recall that the repetition threshold $RT(n)$ is the smallest real number α such that there exists an infinite a^+-free word over Σ_n. The proof of Dejean's conjecture established that $RT(2) = 2$, $RT(3) = \frac{7}{5}$, $RT(4) = \frac{7}{4}$, and $RT(n) = \frac{n}{n-1}$ for every $n \geqslant 5$. An infinite $RT(n)^+$-free word over Σ_n is called a Dejean word.

3 Nice formulas

All the nice formulas considered so far in the literature are also 3-avoidable. This includes doubled patterns [12], circular formulas [9], the nice formulas in the 3-avoidance basis [9], and the minimally nice ternary formulas in Table 1 [15].

Theorem 7 ([9, 15])**.** *Every nice formula with at most* 3 *variables is* 3*-avoidable.*

We have a risky conjecture that would generalize both Theorem 7 and the 3-avoidability of doubled patterns.

Conjecture 8. Every nice formula is 3-avoidable.

Theorem 19 in Section 5 shows that there exist infinitely many nice formulas with index 3. It means that Conjecture 8 would be best possible and it contrasts with the case of doubled patterns, since we expect that there exist only finitely many doubled patterns with index 3 [12, 13]. In this section, we make progress toward Conjecture 8 by proving that every nice formula is avoidable and we explain how to get an upper bound on the index of a given nice formula.

3.1 The avoidability exponent

Let us consider a useful tool in pattern avoidance that has been defined in [12] and already used implicitly in [11]. The *avoidability exponent* $AE(p)$ of a pattern p is the largest real α such that every α-free word avoids p. We extend this definition to formulas. The corresponding notion for the avoidance of patterns in the abelian setting has also been considered [7].

Let us show that $AE(ABCBA.CBABC) = \frac{4}{3}$. Suppose for contradiction that a $\frac{4}{3}$-free word contains an occurrence h of $ABCBA.CBABC$. We write $y = |h(Y)|$ for every variable Y. The factor $h(ABCBA)$ is a repetition with period $|h(ABCB)|$. So we have $\frac{a+b+c+b+a}{a+b+c+b} < \frac{4}{3}$. This simplifies to $2a < 2b + c$. Similarly, $CBABC$ gives $2c < a + 2b$, BAB gives $2b < a$, and BCB gives $2b < c$. Summing up these four inequalities gives $2a + 4b + 2c < 2a + 4b + 2c$, which is a contradiction. On the other hand, the word `0123420156786587683420123 4` is $\left(\frac{4}{3}^{+}\right)$-free and contains the occurrence $A \to$ `01`, $B \to$ `2`, $C \to$ `34` of $ABCBA.CBABC$.

As a second example, we obtain that $AE(ABCDBACBD) = 1.246266172\ldots$. When we consider a repetition uvu in an α-free word, we derive that $\frac{|uvu|}{|uv|} < \alpha$, which gives $\beta|u| < |v|$ with $\alpha = 1+\frac{1}{\beta+1}$. We consider an occurrence h of the pattern. The maximal repetitions in $ABCDBACBD$ are $ABCDBA$, $BCDB$, $BACB$, $CDBAC$, and $DBACBD$. They imply the following inequalities.

$$\begin{cases} \beta a \leqslant 2b + c + d \\ \beta b \leqslant c + d \\ \beta b \leqslant a + c \\ \beta c \leqslant a + b + d \\ \beta d \leqslant a + 2b + c \end{cases}$$

We look for the smallest β such that this system has no solution. Notice that a and d play symmetric roles. Thus, we can set $a = d$ and simplify the system.

$$\begin{cases} \beta a \leqslant a + 2b + c \\ \beta b \leqslant a + c \\ \beta c \leqslant 2a + b \end{cases}$$

Then β is the largest eigenvalue of the matrix $\begin{bmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$ that corresponds to the latter system. So $\beta = 3.060647027\ldots$ is the largest root of the characteristic polynomial $x^3 - x^2 - 5x - 4$. Then $\alpha = 1 + \frac{1}{\beta+1} = 1.246266172\ldots$

This matrix approach is a convenient trick to use when possible. It was used in particular for some doubled patterns such that every variable occurs exactly twice [12]. It may fail if the number of inequalities is strictly greater than the number of variables or if the formula contains a repetition uvu such that $|u| \geqslant 2$. In any case, we can fix a rational value to β and ask a computer algebra system whether the system of inequalities is solvable. Then we can get arbitrarily good approximations of β (and thus α) by a dichotomy method.

Of course, the avoidability exponent is related to divisibility.

Lemma 9. *If $f \preceq g$, then $AE(f) \leqslant AE(g)$.*

The avoidability exponent depends on the repetitions induced by f. We have $AE(f) = 1$ for formulas such as $f = AB.BA.AC.CA.BC$ or $f = AB.BA.AC.BC.CDA.DCD$ that do not have enough repetitions. That is, for every $\varepsilon > 0$, there exists a $(1+\varepsilon)$-free word that contains an occurrence of f.

Let us investigate formulas with non-trivial avoidability exponent, that is, $AE(f) > 1$. To show that a nice formula has a non-trivial avoidability exponent (see Lemma 10), we first introduce a notion of minimality for nice formulas similar to the notion of minimally avoidable for general formulas. A nice formula f is *minimally nice* if there exists no nice formula g such that $v(g) \leqslant v(f)$ and $g \prec f$. Alternatively, splitting a minimally nice formula on any of its fragments leads to a non-nice formula. The following property of every minimally nice formula is easy to derive. If a variable V appears as a prefix of a fragment ϕ, then

- V is also a suffix of ϕ (since otherwise we can split on ϕ and obtain a nice formula),
- ϕ contains exactly two occurrences of V (since otherwise we can remove the prefix letter V from ϕ and obtain a nice formula),
- V is neither a prefix nor a suffix of any fragment other than ϕ (since otherwise we can remove this prefix/suffix letter V from the other fragment and obtain a nice formula),
- Every fragment other than ϕ contains at most one occurrence of V (since otherwise we can remove the prefix letter V from ϕ and obtain a nice formula).

Lemma 10. *If f is a nice formula with $v(f) \geqslant 3$, then $AE(f) \geqslant 1 + \frac{1}{2v(f)-3}$.*

Proof. First remark that if a word uvu is $\left(1+\frac{1}{2v(f)-3}\right)$-free then $2|u| + |v| < (|u| + |v|)\left(1+\frac{1}{2v(f)-3}\right)$ which implies $(2v(f)-4)|u| < |v|$.

Suppose that f contradicts the lemma. Then there exists a $\left(1+\frac{1}{2v(f)-3}\right)$-free word w containing an occurrence h of f. Let X be a variable of f such that $|h(X)| \geqslant |h(Y)|$ for every variable Y. Since f is nice, f contains a factor of the form XPX where P is a sequence of variables that does not contain X. Remark that $v(P) \leqslant v(f) - 1$.

For any variable Z, let $|P|_Z$ be the number of occurences of Z in P. Let Y be the variable that maximizes $|h(Y)| \times |P|_Y$, that is, $|h(W)| \times |P|_W \leqslant |h(Y)| \times |P|_Y$ for every variable W in P. We have

$$|h(P)| = \sum_{W \in Var(P)} |h(W)| \times |P|_W \leqslant (v(f)-1)|h(Y)| \times |P|_Y \leqslant (v(f)-1)|h(X)| \times |P|_Y.$$

If $|P|_Y = 1$, then $|h(P)| \leqslant (v(f)-1)|h(X)|$ and the exponent of $|h(XPX)|$ is at least $\frac{(v(f)+1)|h(X)|}{v(f)|h(X)|} = 1 + \frac{1}{v(f)}$, which is a contradiction.

If $|P|_Y \geqslant 2$, then the number of letters of $h(P)$ that do not belong to an occurence of $h(Y)$ is at most

$$\sum_{W \in Var(P)\setminus\{Y\}} |h(W)| \times |P|_W \leqslant (v(f)-2)|h(Y)| \times |P|_Y.$$

Thus there exist two occurences of $h(Y)$ in $h(P)$ that are separated by at most $\frac{(v(f)-2)|h(Y)|\times|P|_Y}{|P|_Y-1}$ letters. Since $h(P)$ is $\left(1+\frac{1}{2v(f)-3}\right)$-free, we obtain

$$(2v(f)-4)|h(Y)| < \frac{(v(f)-2)|h(Y)| \times |P|_Y}{|P|_Y - 1}.$$

This can be simplified to

$$(2v(f)-4)(|P|_Y - 1) < (v(f)-2) \times |P|_Y$$

and finally

$$|P|_Y < \frac{2v(f)-4}{v(f)-2} = 2\,,$$

which is a contradiction. □

The circular formulas studied in [9] show that $AE(f)$ can be as low as $1+(v(f))^{-1}$. Moreover, our example $AE(ABCDBACBD) = 1.246266172\ldots$ shows that lower avoidability exponents exist among nice formulas with at least 4 variables.

We will describe below a method to construct infinite words avoiding a formula. This method can be applied if and only if the formula f satisfies $AE(f) > 1$. So we are interested in characterizing the formulas f such that $AE(f) > 1$. By Theorems 9 and 10, if f is a formula such that there exists a nice formula g satisfying $g \preceq f$, then $AE(f) > 1$. Now we prove that the converse also holds, which gives the following characterization.

Theorem 11. *A formula f satisfies $AE(f) > 1$ if and only if there exists a nice formula g such that $g \preceq f$.*

Proof. What remains to prove is that for every formula f that is not divisible by a nice formula and for every $\varepsilon > 0$, there exists an infinite $(1+\varepsilon)$-free word w containing an occurrence of f, such that the size of the alphabet of w only depends on f and ε.

First, we consider the equivalent pattern p obtained from f by replacing every dot by a distinct variable that does not appear in f. We will actually construct an occurrence of p. Then we construct a family f_i of pseudo-formulas as follows. We start with $f_0 = p$. To obtain f_{i+1} from f_i, we choose a variable that appears at most once in every fragment of f_i. This variable is given the alias name V_i and every occurrence of V_i is replaced by a dot. We say that f_i is a pseudo-formula since we do not try to normalize f_i, that is, f_i can contain consecutive dots and f_i can contain fragments that are factors of other fragments. However, we still have a notion of fragment for a pseudo-formula. Since f is not divisible by a nice formula, this process ends with the pseudo-formula $f_{v(p)}$ with no variable and

$|p|$ consecutive dots. The goal of this process is to obtain the ordering V_0, V_1, …, $V_{v(p)-1}$ on the variables of p.

The image of every V_i is a finite factor w_i of a Dejean word over an alphabet of $\lfloor\varepsilon^{-1}\rfloor+2$ letters, so that w_i is $(1+\varepsilon)$-free. The alphabets are disjoint: if $i \neq j$, then w_i and w_j have no common letter. Finally, we define the length of w_i as follows: $\left|w_{v(p)-1}\right| = 1$ and $|w_i| = \lfloor\varepsilon^{-1}\rfloor \times |p| \times |w_{i+1}|$ for every i such that $0 \leqslant i \leqslant v(p)-2$. Let us show by contradiction that the constructed occurrence h of p is $(1+\varepsilon)$-free. Consider a repetition xyx of exponent at least $1+\varepsilon$ that is maximal, that is, which cannot be extended to a repetition with the same period and larger exponent. Since every w_i is $(1+\varepsilon)$-free and since two matching letters must come from distinct occurrences of the same variable, then $x = h(x')$ and $y = h(y')$ where x' and y' are factors of p. Our ordering of the variables of p implies that y' contains a variable V_i such that $i < j$ for every variable V_j in x'. Thus, $|y| \geqslant |w_i| = \lfloor\varepsilon^{-1}\rfloor \times |p| \times |w_{i+1}| \geqslant \lfloor\varepsilon^{-1}\rfloor \times |x|$, which contradicts the fact that the exponent of xyx is at least $1+\varepsilon$.

To obtain the infinite word w, we can insert our occurrence of p into a bi-infinite $(1+\varepsilon)$-free word over an alphabet of $\lfloor\varepsilon^{-1}\rfloor+2$ new letters. So w is an infinite $(1+\varepsilon)$-free word over an alphabet of $v(p)\left(\lfloor\varepsilon^{-1}\rfloor+2\right)+1$ letters which contains an occurrence of f. □

By Lemma 10, every nice formula is avoidable since it is avoided by a Dejean word over a sufficiently large alphabet. Thus, if a formula is nice and minimally avoidable, then it is minimally nice. This is the case for every formula in the 3-avoidance basis, except $AB.AC.BA.CA.CB$. However, a minimally nice formula is not necessarily minimally avoidable. Indeed, we have shown [15] that the set of minimally nice ternary formulas consists of the nice formulas in the 3-avoidance basis, together with the minimally nice formulas in Table 1 that can be split to $AB.AC.BA.CA.CB$.

- $ABA.BCB.CAC$
- $ABCA.BCAB.CBAC$ and its reverse
- $ABCA.BAB.CAC$
- $ABCA.BAB.CBC$ and its reverse
- $ABCA.BAB.CBAC$ and its reverse
- $ABCBA.CABC$ and its reverse
- $ABCBA.CAC$

Table 1: The minimally nice ternary formulas that are not minimally avoidable.

3.2 Avoiding a nice formula

Recall that a nice formula f is such that $AE(f) > 1$. We consider the smallest integer s such that $RT(s) < AE(f)$. Thus, every Dejean word over Σ_s avoids f, which already gives $\lambda(f) \leqslant s$. Recall that a morphism is q-uniform if the image of every letter has length q. Also, a uniform morphism $h : \Sigma_s^* \to \Sigma_e^*$ is *synchronizing* if for any $a, b, c \in \Sigma_s$ and $v, w \in \Sigma_e^*$, if $h(ab) = vh(c)w$, then either $v = \varepsilon$ and $a = c$ or $w = \varepsilon$ and $b = c$. For increasing values of q, we look for a q-uniform morphism $h : \Sigma_s^* \to \Sigma_e^*$ such that $h(w)$ avoids f for every $RT(s)^+$-free word $w \in \Sigma_s^\ell$, where ℓ is given by Lemma 12 below. Recall that a word is (β^+, n)-free if it contains no repetition with exponent strictly greater than β and period at least n.

Lemma 12. *[11] Let $\alpha, \beta \in \mathbb{Q}$, $1 < \alpha < \beta < 2$ and $n \in \mathbb{N}^*$. Let $h : \Sigma_s^* \to \Sigma_e^*$ be a synchronizing q-uniform morphism (with $q \geqslant 1$). If $h(w)$ is (β^+, n)-free for every α^+-free word w such that $|w| < \max\left(\frac{2\beta}{\beta-\alpha}, \frac{2(q-1)(2\beta-1)}{q(\beta-1)}\right)$, then $h(w)$ is (β^+, n)-free for every (finite or infinite) α^+-free word w.*

Given such a candidate morphism h, we use Lemma 12 to show that for every $RT(s)^+$-free word $w \in \Sigma_s^*$, the image $h(w)$ is (β^+, n)-free. The pair (β, n) is chosen such that $RT(s) < \beta < AE(f)$ and n is the smallest possible for the corresponding β. If $\beta < AE(f)$, then every occurrence h of f in a (β^+, t)-free word is such that the length of the h-image of every variable of f is upper bounded by a function of n and f only. Thus, the h-image of every fragment of f has bounded length and we can check that f is avoided by inspecting a finite set of factors of words of the form $h(w)$.

3.3 The number of fragments of a minimally avoidable formula

Interestingly, the notion of (minimally) nice formula is helpful in proving the following.

Theorem 13. *The only minimally avoidable formula with exactly one fragment is AA.*

Proof. A formula with one fragment is a doubled pattern. Since it is minimally avoidable, it is a minimally nice formula. By the properties of minimally nice formulas discussed above, the unique fragment of the formula is either AA or is of the form ApA such that p does not contain the variable A. Thus, p is a doubled pattern such that $p \prec ApA$, which contradicts that ApA is minimally avoidable. □

By contrast, the family of *two-birds* formulas, which consists of $ABA.BAB$, $ABCBA.CBABC$, $ABCDCBA.DCBABCD$, and so on, shows that there exist infinitely many minimally avoidable formulas with exactly two fragments. Every two-birds formula is nice. Let us check that every two-birds formula $AB \cdots X \cdots BA.X \cdots A \cdots X$ is minimally avoidable. Since the two fragments play symmetric roles, it is sufficient to split on the first fragment. We obtain the formula $AB \cdots X \cdots B.B \cdots X \cdots BA.X \cdots A \cdots X$ which divides the pattern $B \cdots X \cdots BAB \cdots X \cdots B = Z(B \cdots X \cdots B)$. This pattern is equivalent to $B \cdots X \cdots B$, which is unavoidable. Thus, every two-birds formula is indeed minimally avoidable.

Concerning the index of two-birds formulas, we have seen that $\lambda(ABA.BAB) = 3$ and $\lambda(ABCBA.CBABC) = 2$ [9]. Computer experiments suggest that larger two-birds formulas are easier to avoid.

Conjecture 14. Every two-birds formula with at least 3 variables is 2-avoidable.

4 Characterization of some famous morphic words

Our next result gives characterizations of w_3, up to renaming, that use just one formula. Then we give similar characterizations of b_3 and b_2. Let $\sigma = \mathtt{1/2/0}$ be the morphism that cyclically permutes Σ_3.

Theorem 15. *Let $f_h = ABA.BCB.ACA$, $f_e = ABA.ABCBA.ACA.ACB.BCA$, and let f be such that $f_h \preceq f \preceq f_e$. Every ternary recurrent word avoiding f is equivalent to w_3, $\sigma(w_3)$, or $\sigma^2(w_3)$.*

Proof. Using Cassaigne's algorithm [4], we have checked that w_3 avoids f_h. By divisibility, w_3 avoids f.

Let w be a ternary recurrent word avoiding f. By Lemma 6, w is square-free.

Let $v = \mathtt{210201202101201021}$. A computer check shows that no infinite ternary word avoids f_e, squares, v, $\sigma(v)$, and $\sigma^2(v)$. So, without loss of generality, w contains v. If w contains $\mathtt{121}$, then w contains the occurrence $A \to \mathtt{1}$, $B \to \mathtt{2}$, $C \to \mathtt{0}$ of f_e. Similarly, if w contains $\mathtt{212}$, then w contains the occurrence $A \to \mathtt{2}$, $B \to \mathtt{1}$, $C \to \mathtt{0}$ of f_e. Thus, w avoids squares, $\mathtt{121}$, and $\mathtt{212}$. By Theorem 2, w is equivalent to w_3.

By symmetry, every ternary recurrent word avoiding f is equivalent to w_3, $\sigma(w_3)$, or $\sigma^2(w_3)$. □

Theorem 16. *Let f be such that*

- $ABCA.ABA.ACA \preceq f \preceq ABCA.ABA.ACA.ACB.CBA$,
- $ABCA.ABA.BCB.AC \preceq f \preceq ABCA.ABA.ABCBA.ACB$, *or*
- $ABCA.ABA.BCB.CBA \preceq f \preceq ABCA.ABA.ABCBA.ACB$.

Every ternary recurrent word avoiding f is equivalent to b_3, $\sigma(b_3)$, or $\sigma^2(b_3)$.

Proof. Using Cassaigne's algorithm [4], we have checked that b_3 avoids $ABCA.ABA.ACA$, $ABCA.ABA.BCB.AC$, and $ABCA.ABA.BCB.CBA$. By divisibility, b_3 avoids f. Let w be a ternary recurrent word avoiding f. By Lemma 6, w is square-free.

Let $v = \mathtt{20210121020120}$. A computer check shows that no infinite ternary word avoids $ABCA.ABA.ACA.ACB.CBA$ (resp. $ABCA.ABA.ABCBA.ACB$), squares, v, $\sigma(v)$, and $\sigma^2(v)$.

So, without loss of generality, w contains v. If w contains $\mathtt{010}$, then w contains the occurrence $A \to \mathtt{0}$, $B \to \mathtt{1}$, $C \to \mathtt{2}$ of $ABA.ACA.ABCA.ACBA.ABCBA$. Similarly, if w contains $\mathtt{212}$, then w contains the occurrence $A \to \mathtt{2}$, $B \to \mathtt{1}$, $C \to \mathtt{0}$ of

$ABA.ACA.ABCA.ACBA.ABCBA$. Thus, w avoids squares, 010, and 212. By Theorem 2, w is equivalent to b_3.

By symmetry, every ternary recurrent word avoiding f is equivalent to b_3, $\sigma(b_3)$, or $\sigma^2(b_3)$. □

Notice that Theorem 16 is a complement to [15, Theorem 2] in which we gave a disjoint set of formulas with the same property. The difference between Theorem 16 and [15, Theorem 2] is that a different occurrence of f shows that f divides $Z^n(AA)$.

Theorem 17. *Let $f_h = AABCAA.BCB$, $f_e = AABCAAB.AABCAB.AABCB$, and let f be such that $f_h \preceq f \preceq f_e$. Every binary recurrent word avoiding f is equivalent to b_2.*

Proof. Using Cassaigne's algorithm [4], we have checked that b_2 avoids f_h. First, $f_e \preceq AAA$ because $Z(AAA) = AAABAAA$ contains the occurrence $A \to A$, $B \to A$, $C \to B$ of f_e. Second, $f_e \preceq ABABA$ because $Z(ABABA) = ABABACABABA$ contains the occurrence $A \to AB$, $B \to A$, $C \to C$ of f_e.

Thus, every recurrent word avoiding f_e also avoids AAA and $ABABA$, which means that it is overlap-free. Finally, it is well-known that every binary recurrent word that is overlap-free is equivalent to b_2. □

5 xyx-formulas

Recall that every fragment of an xyx-formula is of the form XYX. We associate to an xyx-formula F the directed graph $\overrightarrow{G}$ such that every variable corresponds to a vertex and $\overrightarrow{G}$ contains the arc $\overrightarrow{XY}$ if and only if F contains the fragment XYX. We will also denote by G the underlying simple graph of $\overrightarrow{G}$.

Lemma 18. *Let F_1 and F_2 be xyx-formulas associated to $\overrightarrow{G_1}$ and $\overrightarrow{G_2}$. If there exists a homomorphism $\overrightarrow{G_1} \to \overrightarrow{G_2}$, then $F_1 \preceq F_2$.*

Proof. Since both digraph homomorphism and formula divisibility are transitive relations, we only need to consider the following two cases. If G_1 is a subgraph of G_2, then F_1 is obtained from F_2 by removing some fragments. So every occurrence of F_2 is also an occurrence of F_1 and thus $F_1 \preceq F_2$. If G_2 is obtained from G_1 by identifying the vertices u and v, then F_2 is obtained from F_1 by identifying the variables U and V. So every occurrence of F_2 is also an occurrence of F_1 and thus $F_1 \preceq F_2$. □

For every i, let T_i be the xyx-formula corresponding to the directed circuit $\overrightarrow{C_i}$ of length i, that is, $T_1 = AAA$, $T_2 = ABA.BAB$, $T_3 = ABA.BCB.CAC$,
$T_4 = ABA.BCB.CDC.DAD$, and so on. More formally, T_i is the formula with i variables $A_0, \ldots, A_{i-1}$ which contains the i fragments of length three of the form $A_jA_{j+1}A_j$ such that the indices are taken modulo i. Notice that T_i is a nice formula.

Theorem 19. *For every $i \geqslant 2$, $\lambda(T_i) = 3$.*

Proof. We use Lemma 12 to show that the image of every $(7/4^+)$-free word over Σ_4 by the following 58-uniform morphism is $(3/2, 3)$-free.

0 → 0012211002201021120022100112201002112001022011002211201022
1 → 0012210022010211220010221120011022010021122011002211201022
2 → 0011221002201021122001102201002112001022110012200211201022
3 → 0011221002201021120011022010021122001022110012200211201022

In these words, the factor 010 is the only occurrence m of ABA such that $|m(A)| \geqslant |m(B)|$. This implies that these ternary words avoid T_i for every $i \geqslant 1$, so that $\lambda(T_i) \leqslant 3$.

To show that $\lambda(T_i) \geqslant 3$, we consider the xyx-formula $H = ABA.BAB.ACA.CBC$ associated to the directed graph $\overrightarrow{D_3}$ on 3 vertices and 4 arcs that contains a circuit of length 2 and a circuit of length 3. Standard backtracking shows that $\lambda(H) > 2$, and even the stronger result that $\lambda(ABAB.ACA.CAC.BCB.CBC) > 2$.

For every $i \geqslant 2$, the circuit $\overrightarrow{C_i}$ admits a homomorphism to $\overrightarrow{D_3}$. By Lemma 18, this means that $T_i \preceq H$, which implies that $\lambda(T_i) \geqslant \lambda(H) \geqslant 3$. □

Theorem 20. *For every $i \geqslant 1$, b_4 avoids T_i.*

Proof. Suppose for contradiction that there exist i and n such that $m^n(0)$ contains an occurrence h of T_i. Further assume that n is minimal. Notice that in b_4, every even (resp. odd) letter appears only at even (resp. odd) positions. Thus, for every fragment XYX of T_i, the period $|h(XY)|$ of the repetition $h(XYX)$ must be even. This implies that $|h(X)|$ and $|h(Y)|$ have the same parity. By contagion, the lengths of the images of all the variables of T_i have the same parity. Now we proceed to a case analysis.

- Every $|h(X)|$ is even.
  - Every $h(X)$ starts with 0 or 2. By taking the pre-image by m of every $h(X)$, we obtain an occurrence of T_i that is contained in $m^{n-1}(0)$. This contradicts the minimality of n.
  - Every $h(X)$ starts with 1 or 3. Notice that in b_4, the letter 1 (resp. 3) is in position 1 (mod 4) (resp. 3 (mod 4)). $m^n(0)$ contains the occurrence h' of T_i such that $h'(X)$ is obtained from $h(X)$ by adding to the rigth the letter 1 or 3 depending on its position modulo 4 and by removing the first letter. Since is also contained in $m^n(0)$ and every $h'(X)$ starts with 0 or 2, h' satisfies the previous subcase.
- Every $|h(X)|$ is odd. It is not hard to check that every factor uvu in b_4 with $|v| = 1$ satisfies $v \in \{1, 3\}$ and $u \in \{0, 2\}$. So $|h(X)| \geqslant 3$ for every variable X of T_i. Let $X_1, \cdots, X_i$ be the variables of T_i. Up to a shift of indices, we can assume that j and the first and last letters of $h(X_j)$ have the same parity. We construct the occurrence h' of T_i as follows. If j is odd, then $h'(X_j)$ is obtained by removing the first letter of $h(X_j)$. If j is even, then $h'(X_j)$ is obtained by adding to the right the letter 1 or 3 depending on its position modulo 4. Since h' is also contained in $m^n(0)$ and every $|h'(X)|$ is even, h' satisfies the previous case. □

Our next result generalizes Theorems 5 and 20. Recall that every fragment of a hybrid formula has length 2 or is of the form XYX.

Theorem 21. *Every avoidable hybrid formula is avoided by* b_4.

Proof. Let f be a hybrid formula. If f contains a locked formula or a formula T_i, then b_4 avoids f by Theorems 4 and 20. If f contains neither a locked formula nor a formula T_i, then we show that f is unavoidable. By induction and by theorem 1 it is sufficient to show that f is reducible to a hybrid formula containing neither a locked formula nor a formula T_i. Since f is not locked, f contains a free set of variables and thus f has a free singleton $\{X\}$. If f contains a fragment YXY, then $\{Y\}$ is also a free singleton of f. Using this argument iteratively, we end up with a free singleton $\{Z\}$ such that f contains no fragment TZT, since f contains no formula T_i.

So we can assume that f contains a free singleton $\{Z\}$ and no fragment TZT. Thus, deleting every occurrence of Z from f gives an hybrid sub-formula containing neither a locked formula nor a formula T_i. By induction, f is unavoidable. □

So the index of an avoidable xyx-formula is at most 4 and we have seen examples of xyx-formulas with index 3 in Theorems 15 and 19. The next results give an xyx-formula with index 4 and an xyx-formula with index 2 that is not divisible by AAA.

Theorem 22. $\lambda(ABA.BCB.DCD.DED.AEA) = 4$.

Proof. By Theorem 21, $ABA.BCB.DCD.DED.AEA$ is 4-avoidable.
Notice that $ABA.BCB.DCD.DED.AEA \preceq ABA.BCB.ACA$ via the homomorphism $A \to A$, $B \to B$, $C \to C$, $D \to B$, $E \to C$. Moreover, w_3 contains the occurrence $A \to \texttt{0}$, $B \to \texttt{1}$, $C \to \texttt{02}$, $D \to \texttt{01}$, $E \to \texttt{2}$ of $ABA.BCB.DCD.DED.AEA$. By Theorem 15, the formula is not 3-avoidable. □

Theorem 23. *The fixed point of* `001/011` *avoids the* xyx*-formula associated to the directed graph on* 4 *vertices with all the* 12 *arcs.*

Proof. We use again Cassaigne's algorithm. □

6 Palindrome patterns

Mikhailova [10] has considered the index of an avoidable pattern that is a palindrome and proved that it is at most 16. She actually constructed a morphic word over Σ_{16} that avoids every avoidable palindrome pattern.

We make a distinction between the largest index $\mathcal{P}_w$ of an avoidable palindrome pattern and the smallest alphabet size $\mathcal{P}_s$ allowing an infinite word avoiding every avoidable palindrome pattern. We obtained [15] the lower bound

$$\lambda(ABCADACBA) = \lambda(ABCA.ACBA) = 4,$$

so that $4 \leqslant \mathcal{P}_w \leqslant \mathcal{P}_s \leqslant 16$.

The following result is a slight improvement to $\lambda(ABCA.ACBA) = 4$ that is not related to palindromes.

Theorem 24. $\lambda(ABCA.ACBA.ABCBA) = 4$.

Proof. By Lemma 6, every recurrent word avoiding $ABCA.ACBA.ABCBA$ is square-free. A computer check shows that no infinite ternary square-free word avoids the occurrences h of $ABCA.ACBA.ABCBA$ such that $|h(A)| = 1$, $|h(B)| \leqslant 2$, and $|h(C)| \leqslant 3$. $\square$

Let us give necessary conditions on a palindrome pattern P so that $5 \leqslant \lambda(P) \leqslant 16$.

1. The length of P is odd and the central variable of P is isolated. Indeed, otherwise P would be a doubled pattern and thus 3-avoidable [12].
2. No variable of P appears both at an even and an odd position. Indeed, if P had a variable that appears both at an even and an odd position, then P would be divisible by a formula in the family AA, $ABCA.ACBA$, $ABCDEA.AEDCBA$, $ABCDEFGA.AGFEDCBA$, ... Such formulas (with an odd number of variables) are locked and thus are avoided by b_4 by Theorem 4. So P would be 4-avoidable.

We have found three patterns/formulas satisfying these conditions (see Theorem 25), but they seem to be 2-avoidable. We use again Cassaigne's algorithm with simple pure morphic words to ensure that they are 4-avoidable. Let z_3 be the fixed point of `01/2/20`.

Theorem 25.

1. *$ADBDCDAD.DADCDBDA$ is avoided by b_4.*
2. *$ABCDADC.CDADCBA$ is avoided by z_3.*
3. *$ABACDBAC.CABDCABA$ is avoided by z_3 and b_4.*

7 Discussion

Let us briefly mention the things that we have attempted to do in this paper, without success.

- Find a result similar to Theorems 15 and 16 for v_3, the morphic word avoiding squares, `010`, and `020`.
- Improve Theorem 23 by showing that some xyx-formula on 4 variables and fewer fragments is 2-avoidable.
- Show that the xyx-formula associated to the transitive tournament on 5 vertices is 2-avoidable.

References

[1] G. Badkobeh and P. Ochem. Characterization of some binary words with few squares. *Theor. Comput. Sci.* **588** (2015), 73–80.

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

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

[4] J. Cassaigne. *An Algorithm to Test if a Given Circular HD0L-Language Avoids a Pattern. IFIP Congress*, pages 459–464, 1994.

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

[6] 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

[7] J. Currie and V. Linek. Avoiding patterns in the Abelian sense. *Canadian Journal of Mathematics*, 53:696–714, 2001.

[8] Pytheas Fogg. Substitutions in Dynamics, Arithmetics and Combinatorics. Springer Science & Business Media, 2002.

[9] G. Gamard, P. Ochem, G. Richomme, and P. Séébold. Avoidability of circular formulas. *Theor. Comput. Sci.*, 726:1–4, 2018.

[10] I. Mikhailova. On the avoidability index of palindromes. *Matematicheskie Zametki.*, 93(4):634–636, 2013.

[11] P. Ochem. A generator of morphisms for infinite words. *RAIRO - Theoret. Informatics Appl.*, 40:427–441, 2006.

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

[13] P. Ochem and A. Pinlou. Application of entropy compression in pattern avoidance. *Electron. J. Comb.* **21(2)** (2014), #P2.7.

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

[15] P. Ochem and M. Rosenfeld. On some interesting ternary formulas. *Electron. J. Combin.*, 26(1):#P1.12, 2019.

[16] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, (1977), 413–478.