Avoidability of long k-abelian repetitions

Michaël Rao, Matthieu Rosenfeld

HAL Id: lirmm-04659496
https://hal-lirmm.ccsd.cnrs.fr/lirmm-04659496

Submitted on 24 Jul 2024

AVOIDABILITY OF LONG k-ABELIAN REPETITIONS

MICHAËL RAO AND MATTHIEU ROSENFELD

ABSTRACT. We study the avoidability of long k-abelian-squares and k-abelian-cubes on binary and ternary alphabets. For $k = 1$, these are Mäkelä's questions. We show that one cannot avoid abelian-cubes of abelian period at least 2 in infinite binary words, and therefore answering negatively one question from Mäkelä. Then we show that one can avoid 3-abelian-squares of period at least 3 in infinite binary words and 2-abelian-squares of period at least 2 in infinite ternary words. Finally, we study the minimum number of distinct k-abelian-squares that must appear in an infinite binary word.

1. INTRODUCTION

Avoidability of structures and patterns has been extensively studied in theoretical computer science since the work of Thue on avoidability of repetitions in words [15]. Thue showed that there are infinitely long ternary words avoiding squares (factors of the form ww where w is a word) and infinitely long binary words avoiding cubes (factors of the form www where w is a word).

The avoidability of abelian repetitions has been studied since a question from Erdős in 1957 [5,6]. A factor uv is an abelian-square if u is a permutation of the letters of v. Erdős asked whether it is possible to avoid abelian-squares in an infinite word over an alphabet of size 4. (Abelian-square-free ternary words have a length of at most 7.) After some intermediary results (alphabet of size 25 by Evdokimov [7] and size 5 by Pleasant [13]), Keränen answered positively Erdős's question by giving a 85-uniform morphism (found with the assistance of a computer) whose fixed-point is abelian-square-free [10]. Moreover, Dekking showed that it is possible to avoid abelian-cubes on a ternary alphabet and abelian-4th-powers over a binary alphabet [3].

Erdős also asked if it is possible to avoid arbitrarily long ordinary squares on binary words. This question was answered positively by Entringer, Jackson and Schatz [4]. Mäkelä asked the following two questions about the avoidability of long abelian-cubes (resp. squares) on a binary (resp. ternary) alphabet:

Question 1 (Mäkelä (see [11])). Can you avoid abelian-cubes of the form uvw where $|u| \geq 2$, over two letters? - You can do this at least for words of length 250.

Question 2 (Mäkelä (see [11])). Can you avoid abelian-squares of the form uv where $|u| \geq 2$ over three letters? - Computer experiments show that you can avoid these patterns at least in words of length 450.

The notion of k-abelian repetition has been introduced recently by Karhumäki *et al.* as a generalization of both repetition and abelian repetition [9]. One can

avoid 3-abelian-squares (resp. 2-abelian-cubes) on ternary (resp. binary) words [14]. Following Erdős's and Mäkelä's questions, one can ask whether it is possible to avoid long k-abelian-powers on binary (resp. ternary) words.

In Section 3, we answer negatively Mäkelä's Question 1, and we propose a new version of the conjecture. In Section 4, we show that one can avoid 3-abelian-squares of period at least 3 in binary words and 2-abelian-squares of period at least 2 in ternary words. In Section 5, we study the minimum number of distinct k-abelian-squares that must appear in an infinite binary word. Finally, in Section 6, we explain the computer searches we use to find the morphisms of Section 4 and Section 5.

2. PRELIMINARIES AND DEFINITIONS

We use terminology and notation of Lothaire [12]. Let Σ be a finite alphabet. For a word $u \in \Sigma^*$ and $a \in \Sigma$, we denote by $|u|_a = |\{i : u[i] = a\}|$ the number of occurrences of the letter a in u. For $u, w \in \Sigma^*$, we denote by $|u|_w = |\{i : u[i : i + |w| - 1] = w\}|$ the number of occurrences of the factor w in u.

Two words u and v are said to be *abelian equivalent*, denoted $u \approx_a v$, if for every $a \in \Sigma$, $|u|_a = |v|_a$, and they are said to be *k-abelian equivalent* (for $k \geq 1$), denoted $u \approx_{a,k} v$, if for every $w \in \Sigma^*$ such that $|w| \leq k$, $|u|_w = |v|_w$. A word $u_1 u_2 \ldots u_n$ is a *k-abelian-n-power* if it is non-empty, and $u_1 \approx_{a,k} u_2 \approx_{a,k} \ldots \approx_{a,k} u_n$. Its *period* is $|u_1|$. Similarly, a *k-abelian-square* (resp. *k-abelian-cube*) is a k-abelian-2-power (resp. k-abelian-3-power). A word is said to be *k-abelian-n-power-free* if none of its factors is a k-abelian-n-power. Note that when $k = 1$, the k-abelian-equivalence is exactly the abelian equivalence, and we ommit the "1-" prefix in this case.

The *Parikh vector* of a word $w \in \Sigma^*$, denoted $\Psi(w)$, is the vector indexed by Σ such that for every $a \in \Sigma$, $\Psi(w)[a] = |w|_a$. Then, by definition, two words u and v are abelian-equivalent if $\Psi(u) = \Psi(v)$. For a set $S \subset \Sigma^*$ and a word $w \in \Sigma^*$, we denote by $\Psi_S(w)$ the vector indexed by S such that for every $s \in S$, $\Psi_S(w)[s] = |w|_s$. We may write $\Psi_k(w)$ instead of $\Psi_{\Sigma^k}(w)$ if Σ is clear in the context.

For all $u \in \Sigma^*$, $i \leq |u|$, let $\text{pref}_i(u)$ be the prefix of size i of u and let $\text{suf}_i(u)$ be the suffix of size i of u. There are equivalent definitions of k-abelian-equivalence (see [9]). Two words of size at most $2k - 1$ are k-abelian equivalent if and only if they are equal. For every two words u and v of size at least $k - 1$, the following conditions are equivalent:

- u and v are k-abelian equivalent (*i.e.* $u \approx_{a,k} v$),
- $\Psi_k(u) = \Psi_k(v)$ and $\text{pref}_{k-1}(u) = \text{pref}_{k-1}(v)$,
- $\Psi_k(u) = \Psi_k(v)$ and $\text{suf}_{k-1}(u) = \text{suf}_{k-1}(v)$.

3. ABELIAN CUBES AND MÄKELÄ'S QUESTION 1

Dekking showed that it is possible to avoid abelian-cubes in an infinite word over a ternary alphabet [3]. More recently, Rao showed that one can avoid 2-abelian-cubes over a binary alphabet [14] and one can check that every word over a binary alphabet of length greater than 9 contains an abelian-cube. The only open question about the avoidability of long k-abelian-cubes on infinite words is then the avoidability of long abelian-cubes over the binary alphabet. This is the subject of Question 1 from Mäkelä: he asked whether one can avoid every abelian-cube of period at least 2 in binary words. We answer negatively this question.

For this, we used a property of Lyndon words that made the exhaustive search much faster. A word $w \in \Sigma^*$ is a *Lyndon word* if for all $u, v \in \Sigma^+$ such that $w = uv$, $w <_{\text{lex}} vu$, where $<_{\text{lex}}$ is the lexicographic order. The well-known Chen-Fox-Lyndon Theorem states that every word may be written uniquely as a concatenation of non-increasing Lyndon words (see for example [12]). In the following, we refer to this decomposition as the Lyndon factorization. A language $L \subset \Sigma^*$ is *factorial* if for every w in L, every factor of w is in L.

Lemma 1. *Any factorial language L with arbitrarily long words contains arbitrarily long Lyndon words or repetitions of arbitrarily large power.*

Proof. Let us assume that there are no arbitrarily long Lyndon words in L. This implies that there is a finite number n of Lyndon words in L and $s \in \mathbb{N}$ such that for every Lyndon word w in L, $|w| \leq s$. Let $w_1, \ldots, w_n \in L^n$ be the Lyndon words of L ordered by decreasing lexicographic order.

Then using the Lyndon factorization for every $w \in L$ there are some Lyndon words $L_1 \geq_{\text{lex}} L_2 \geq_{\text{lex}} \ldots \geq_{\text{lex}} L_d$ such that $w = L_1 \ldots L_d$. The fact that our language is factorial tells us that all the L_i are in L. We get that for every $w \in L$ there are $\alpha_1, \ldots, \alpha_n \in \mathbb{N}$ such that $w = w_1^{\alpha_1} \ldots w_n^{\alpha_n}$. Then $|w| \leq \sum_i |w_i| \times \alpha_i \leq s \times \sum_i \alpha_i \leq s \times n \times \max_{\text{i}}(\alpha_i)$.

Since L contains arbitrarily long words, then for any $t \in \mathbb{N}$ there is a $w \in L$ such that $|w| \geq t \times s \times n$. Let $j \in \{1, \ldots, n\}$ be such that $\alpha_j = \max_{\text{i}}(\alpha_i)$. Thus $\alpha_j \geq t$ and then $(w_j)^t \in L$. Thus we have arbitrarily long powers in L. $\square$

A set of words that avoid certain kinds of abelian repetitions is a factorial language and does not contain arbitrarily large powers. Thus we just need to check that there are no arbitrarily long Lyndon words in it to deduce that this set does not contain arbitrarily long words. The exhaustive search on prefixes of Lyndon words is then much shorter. Figure 1 shows how it helps for the exhaustive search of binary words avoiding abelian-squares of period at least 2. The next proposition answers negatively Question 1.

Proposition 1. *There is no infinite word over a binary alphabet avoiding abelian-cubes of period at least 2.*

We checked using a computer program that there are only finitely many Lyndon words over a binary alphabet avoiding abelian-cubes of period at least 2. The program took approximately 3 hours to find all 2,732,711,352 such Lyndon words and prefixes of Lyndon words. The longest word has a length of 290. Using Lemma 1 we deduce that there is no infinite binary word avoiding abelian-cubes of size at least two. Then we can reformulate the question and ask:

Question 3. Is there a $p \in \mathbb{N}$ such that one can avoid abelian-cubes of period at least p over two letters?

For $p = 3$, we found a word of length 2,500.

4. Avoiding long k-abelian-squares on binary words

It is easy to verify that one cannot avoid squares of period at least 2 over a binary alphabet. Entringer *et al.* showed that it is possible to construct a binary word avoiding squares of period at least 3 [4]. They also showed that every infinite binary word contains arbitrarily long abelian-squares. Thus, one can wonder about the avoidability of large k-abelian-squares. Rao asked the following question:

011 - 0111 — 01110 — 011100 — 0111000 — 01110001 — 011100011

01

010 - 0100 — 01000

0

000 - 0001 ⟨ 00011 — 000111 — 0001110 — 00011100 — 000111000
                00010 — 000100 — 0001000

00

001 ⟨ 0011 — 00111 — 001110 — 0011100 — 00111000 — 001110001 — 0011100011
      0010 — 00100 — 001000

ε

100 - 1000 — 10001 — 100011 — 1000111 — 10001110 — 100011100

10

101 - 1011 — 10111

1

111 - 1110 ⟨ 11100 — 111000 — 1110001 — 11100011 — 111000111
              11101 — 111011 — 1110111

11

110 ⟨ 1100 — 11000 — 110001 — 1100011 — 11000111 — 110001110 — 1100011100
      1101 — 11011 — 110111

011 - 0111 — 01110

01

010

0

000 - 0001 ⟨ 00011 — 000111 — 0001110 — 00011100 — 000111000
              00010 — 000100 — 0001000

00

001 ⟨ 0011 — 00111 — 001110 — 0011100
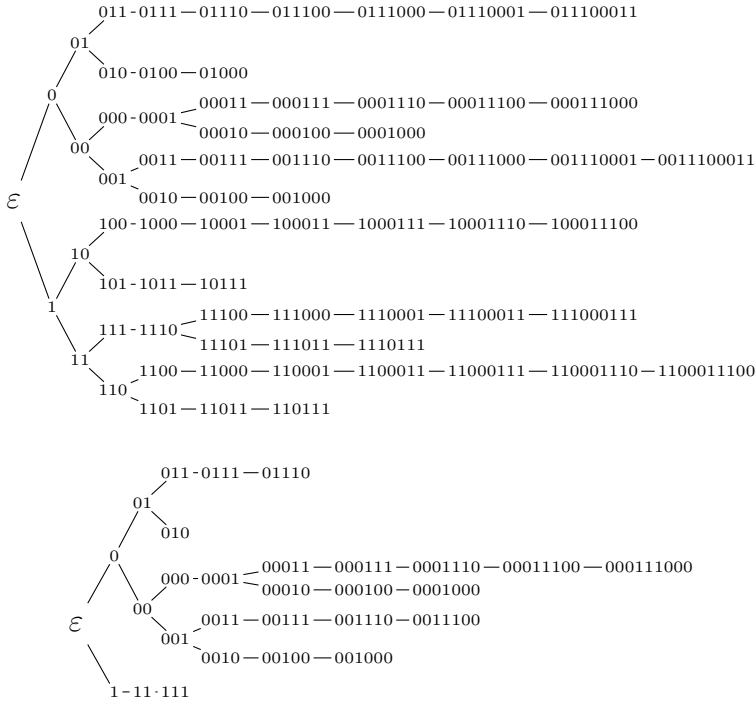   0010 — 00100 — 001000

ε

1 - 11 · 111

FIGURE 1. Top: the exhaustive search of binary words avoiding abelian-squares of period at least 2. Bottom: the same exhaustive search restricted on the prefixes of Lyndon words.

Question 4 (Rao [14]). What is the smallest k (if any) such that arbitrarily long k-abelian-squares can be avoided over a binary alphabet?

Since arbitrarily long abelian squares cannot be avoided in binary words, k is at least 2. In this section, we show that k is at most 3, that is, one can avoid long 3-abelian-squares over a binary alphabet, by giving a morphism whose fixed point avoids 3-abelian-squares of period at least 3.

A morphism h is said to be (p,k)-*abelian-square-free* if for every abelian-square-free word w, $h(w)$ avoids k-abelian-squares of period at least p. Let h be the following morphism:

$$h : \begin{cases} 0 \to & 00001101010, \\ 1 \to & 00011111010, \\ 2 \to & 00110100110, \\ 3 \to & 00111001010. \end{cases}$$

Theorem 1. *The morphism h is $(3,3)$-abelian-square-free.*

Proof. The proof is based on the same idea as the one used by Rao to give sufficient conditions for a morphism to be k-abelian-free [14] which is a generalization of the sufficient conditions given by Carpi for abelian-free-morphisms [1]. We checked the sufficient conditions on h by computer.

In the proof we use the following property: for every $k, i \in \mathbb{N}$ and $u, v \in \Sigma^*$ such that $i < k$, $k-1-i \leq |u|$ and $i \leq |v|$:

(S) $\qquad\qquad \Psi_k(uv) = \Psi_k(u\operatorname{pref}_i(v)) + \Psi_k(\operatorname{suf}_{k-1-i}(u)v).$

Let $A = \{0, 1, 2, 3\}$ and $w \in A^*$. Let us show that if $h(w)$ contains a 3-abelian-square of period at least 3, then w is not abelian-square-free. Note that every image $h(x)$, $x \in A$, starts with the prefix $p = 00$ and ends with the suffix $s = 10$.

We check using a computer that for every $a, b \in A$, $a \neq b$, every 3-abelian-square of $h(ab)$ has period at most 2. So if there is a forbidden 3-abelian-square, it has to be on the image of at least 3 letters. Then there are $a_1, a_2, a_3 \in A$, $x_1, x_2 \in A^*$ and $(u_1, v_1), (u_2, v_2), (u_3, v_3) \in (\{0,1\}^*, \{0,1\}^+)$ such that

- $a_1 x_1 a_2 x_2 a_3$ is a factor of w,
- for every $i \in \{1, 2, 3\}$, $u_i v_i = h(a_i)$,
- $v_1 h(x_1) u_2 \approx_{a,3} v_2 h(x_2) u_3$.

Since $|v_1 u_2 v_2 u_3| \geq 1 + |h(a_2)| \geq 12$, either $|v_1 u_2| \geq 6$ or $|v_2 u_3| \geq 6$. Moreover, $|v_1 h(x_1) u_2| = |v_2 h(x_2) u_3|$, thus for all $i \in \{1, 2\}$, $|v_i h(x_i) u_{i+1}| \geq 6$, and for all $i \in \{1, 2\}$, $|v_i| \geq 2$, $|h(x_i)| \geq 2$ or $|u_{i+1}| \geq 2$. If $|u_{i+1}| \geq 2$, then

$\Psi_3(v_i h(x_i) u_{i+1})$

$\qquad = \Psi_3(v_i 00) + \Psi_3(h(x_i) u_{i+1})$ (using (S) and $\operatorname{pref}_2(h(x_i)u_{i+1}) = 00$),

$\qquad = \Psi_3(v_i 00) + \Psi_3(h(x_i)00) + \Psi_3(u_{i+1})$ (using (S) and $\operatorname{pref}_2(u_{i+1}) = 00$),

$\qquad = \Psi_3(v_i 00) + \Psi_3(h(x_i)00) + \Psi_3(10u_{i+1}) - \Psi_3(1000).$

If $|v_i| \geq 2$ or $|h(x_i)| \geq 2$, we have the same result. So for every $i \in \{1, 2\}$, we get:

(L) $\qquad \Psi_3(v_i h(x_i) u_{i+1}) = \Psi_3(v_i 00) + \Psi_3(h(x_i)00) + \Psi_3(10u_{i+1}) - \Psi_3(1000).$

Let N be the matrix indexed by $\{0, 1\}^3 \times \{0, 1, 2, 3\}$ with $N[w, x] = |h(x)00|_w$.

$$
{}^t N = \begin{pmatrix} 3 & 1 & 2 & 1 & 1 & 2 & 1 & 0 \\ 2 & 1 & 1 & 1 & 1 & 1 & 1 & 3 \\ 1 & 2 & 1 & 2 & 2 & 1 & 2 & 0 \\ 1 & 2 & 2 & 1 & 2 & 1 & 1 & 1 \end{pmatrix}.
$$

For every word w, $\Psi_3(h(w)00) = N\Psi(w)$, thus the equality (L) can be rewritten:

$\qquad \Psi_3(v_i h(x_i) u_{i+1}) = \Psi_3(v_i 00) + N\Psi(x_i) + \Psi_3(10u_{i+1}) - \Psi_3(1000).$

Using $v_1 h(x_1) u_2 \approx_{a,3} v_2 h(x_2) u_3$, we get the following:

(1) $\qquad N(\Psi(x_2) - \Psi(x_1)) = \Psi_3(v_1 00) + \Psi_3(10u_2) - \Psi_3(v_2 00) - \Psi_3(10u_3).$

Let M be the sub-matrix of N made of its rows 1, 2, 3 and 4 (they correspond to the words $000, 001, 010, 011$).

$$
M = \begin{pmatrix} 3 & 2 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 2 & 1 & 1 & 2 \\ 1 & 1 & 2 & 1 \end{pmatrix}.
$$

Let $S = \{000, 001, 010, 011\}$. Then $\Psi_S(w)$ is the sub-vector of $\Psi_3(w)$ made of the rows 1, 2, 3 and 4. We can check that M is non-singular. Thus,

(2) $\qquad \Psi(x_2) - \Psi(x_1) = M^{-1}(\Psi_S(v_1 00) + \Psi_S(10u_2) - \Psi_S(v_2 00) - \Psi_S(10u_3)).$

Let

$$\Psi_S(v_1, u_2, v_2, u_3) = \Psi_S(v_100) + \Psi_S(10u_2) - \Psi_S(v_200) - \Psi_S(10u_3),$$

and

$$\Psi_3(v_1, u_2, v_2, u_3) = \Psi_3(v_100) + \Psi_3(10u_2) - \Psi_3(v_200) - \Psi_3(10u_3).$$

From equation (1), $\Psi_3(v_1, u_2, v_2, u_3)$ is in $\mathrm{Im}(N)$, and from equation (2), $M^{-1}(\Psi_S(v_1, u_2, v_2, u_3))$ is an integer vector. From $v_1 h(x_1) u_2 \approx_{a,3} v_2 h(x_2) u_3$, if we note $p = 00$, $s = 10$ and $k = 3$, we get $\mathrm{pref}_{k-1}(v_1 p) = \mathrm{pref}_{k-1}(v_2 p)$, $\mathrm{suf}_{k-1}(su_2) = \mathrm{suf}_{k-1}(su_3)$.

The following claim is verified using a computer program. There are 4^3 values for the a_i and 11^3 ways of choosing the u_i, v_i for each of them which makes 85,184 cases to check (most of them are eliminated by the prefix and suffix conditions).

Claim 1. For all $a_1, a_2, a_3 \in A$ and $(u_1, v_1), (u_2, v_2), (u_3, v_3) \in (\{0,1\}^*, \{0,1\}^+)$ such that

- $\forall i \in \{1,2,3\}$, $u_i v_i = h(a_i)$,
- $\mathrm{pref}_{k-1}(v_1 p) = \mathrm{pref}_{k-1}(v_2 p)$ and $\mathrm{suf}_{k-1}(su_2) = \mathrm{suf}_{k-1}(su_3)$,
- $\Psi_k(v_1, u_2, v_2, u_3) \in \mathrm{Im}(N)$ and $M^{-1}(\Psi_S(v_1, u_2, v_2, u_3))$ is an integer vector,

there are $(\alpha_1, \alpha_2, \alpha_3) \in \{0,1\}$ such that

$$(E) \qquad M^{-1}(\Psi_S(v_1, u_2, v_2, u_3)) = \alpha_1 \Psi(a_1) - (2\alpha_2 - 1)\Psi(a_2) - (1 - \alpha_3)\Psi(a_3).$$

From the claim we have $(\alpha_1, \alpha_2, \alpha_3) \in \{0,1\}$ such that equation (E) is fulfilled. Now we can introduce x_1', x_2' such that $\forall i \in \{1,2\}$, $x_i' = a_i^{\alpha_i} x_i a_{i+1}^{1-\alpha_{i+1}}$. Then $x_1' x_2'$ is a factor of w and we have the following:

$$\Psi(x_2') - \Psi(x_1') = \Psi(x_2) + (1-\alpha_3)\Psi(a_3) + \alpha_2 \Psi(a_2) - (\Psi(x_1) + (1-\alpha_2)\Psi(a_2) + \alpha_1 \Psi(a_1))$$
$$= M^{-1}(\Psi_S(v_1, u_2, v_2, u_3)) - \alpha_1 \Psi(a_1) + (2\alpha_2 - 1)\Psi(a_2) + (1-\alpha_3)\Psi(a_3),$$
$$\Psi(x_2') - \Psi(x_1') = 0.$$

This implies that there is an abelian-square on w. $\qquad\square$

Using Theorem 1 together with the existence of abelian-square-free words over four letters, we get the following corollary.

Corollary 1. *There is an infinite binary word that avoids 3-abelian-squares of period at least* 3.

Moreover, we can deduce the exponential growth of such words from the exponential growth of abelian-square-free words over four letters [2]. Corollary 1 gives a partial answer to Question 4: there is such a k, and $k \in \{2,3\}$. We can then ask the following question.

Question 5. Can we avoid 2-abelian-squares of period at least p on the binary alphabet, for some $p \in \mathbb{N}$?

Computer experiments show that we can avoid those patterns for $p = 3$ in a word of length 15,000.

2-abelian squares over a ternary alphabet. Rao showed that one can build an infinite word that avoids 3-abelian-squares over a ternary alphabet [14]. The longest 2-abelian-square-free ternary word has a length of 537 [9]. Mäkelä asked whether we can avoid abelian-squares of period at least 2 in ternary words (Question 2). We give the answer to a weaker version of this question, that is one can avoid 2-abelian-squares of period at least 2 over the ternary alphabet. Let

$$h_2 : \begin{cases} 0 \to & 00021, \\ 1 \to & 00111, \\ 2 \to & 01121, \\ 3 \to & 01221. \end{cases}$$

Theorem 2. h_2 is $(2,2)$-abelian-square-free.

Proof. The proof is also done by checking sufficient conditions, similar to those in the proof of Theorem 1. Claim 1 is true for this morphism with $k = 2$, $S = \{00, 01, 02, 11\}$, $s = 1$ and $p = 0$. □

If w is an infinite square-free-abelian word over four letters, $h_2(w)$ is a ternary word which avoids 2-abelian-squares of period at least 2.

5. Minimal number of distinct 3-abelian-squares in infinite binary words

Fraenkel and Simpson showed that there is an infinite binary word containing only the squares 0^2, 1^2, $(01)^2$ [8]. Moreover, every binary infinite word contains at least three distinct squares. It is natural to ask whether this property can be extended to the k-abelian case: Is there a $k \in \mathbb{N}$ such that there is an infinite binary word that contains only 3 distinct k-abelian-squares?

More generally, let $g(k)$ be the minimal number of distinct k-abelian squares that an infinite binary word must contain. Any $(k + 1)$-abelian-square is a k-abelian-square so g is non-increasing. From Fraenkel and Simpson's result, we know that $g(k) \geq 3$ for all k.

Proposition 2. *The morphism h_3 (defined in Table 1) is $(5,3)$-abelian-square-free. Moreover, for every abelian-square-free word w, $h_3(w)$ contains only 3 distinct 5-abelian-squares: 0^2, 1^2 and $(01)^2$.*

Proof. The proof that h_3 is $(5,3)$-abelian-square-free is also done by a computer check, similar to the proof of Theorem 1. This morphism is not uniform, so we need to check images of words of size up to 3 to ensure that the image of a long 5-abelian square starts and ends on images of different letters. □

This proposition together with Keränen's word tells us that for any $k \geq 5$ there is an infinite binary word with only 3 distinct k-abelian-squares, *i.e.*, $g(k) = 3$ for every $k \geq 5$. Propositions 3 and 4 give us that $g(3) = g(4) = 4$.

Proposition 3. *Every word of size more than 87 over the binary alphabet contains at least 4 distinct 4-abelian-squares.*

This was verified by an exhaustive computer search.

TABLE 1. A $(5,3)$-abelian-square-free morphism, with only three distinct 5-abelian-squares: 00, 11 and 0101.

$$h_3 : \begin{cases} \end{cases}$$

$0 \rightarrow$ u1001011000101110001100101100010111001011000111001011100011001011000
1011100101100111000101100011100101110001100101100010111001011000111 0
0101110001100101100011100101100111000110010110001110010111000110010 1
1000101110010110001110010111000110010110001110010110011100010110001 1
1001011000101110010110011100010110001011100101110001100101110001100 1
0110011100010111001011000111001011100011v

$1 \rightarrow$ u0001110010110011100010010110001110010110011100010111001011000111 00
1011100011001011000110010110011100010010110001110010111000101100011
1100101100111000110010110001110010110011100010110001110010110001011 1
0001100101100011100101100111000100101100011100101110001100101100010
1110010110001110010111000100101100011100101100111000110010110001110
0101110001100101100010111001011001110v

$2 \rightarrow$ u0001110010110011100011001011000111001011001110001011100101100011100
1011100011001011000101110010110011100010110001110010110001011100101 1
0011100010111001011000111001011100011001011000101110010110011100010 1
11001011000101110001100101100010111001011001110001011000111001011001
1100011001011000111001011100011001011000101110010110011100010111001 0
11000101110001100101100010111001011001110001 1v

$3 \rightarrow$ u0001110010110001011100011001011000101110010110011100010111000110010
1100011100101100111000100101100011100101110001011000111001011001110
0011001011000111001011001110001011000111001011000101110001100101100 0
1011100101100111000101100011100101110001011001110001100101100011100 1
0110011100010110001110010110001011100101100111000101110010110001110 0
10111000110010110001011100101100111$0v$

Where:

$u =$ 11000110010110001011100101100111000101100011100101100
01011100101100111000101110010110001011100011001011000
11100101100111000110010110001011100101100111000101 1

$v =$ 00101100011100101100111000110010111000101100111000101
11001011000101110001100101110001011001110001100101100
01110010111000101100011100101100111000101100011100101

Proposition 4. *Let*

$$h_4 : \begin{cases} 0 \rightarrow & 000110010100110101100010101000101110101100010 1, \\ 1 \rightarrow & 000110010100110101100111010101110001110101100010 1, \\ 2 \rightarrow & 000110010100111000101000110010110001110101100010 1, \\ 3 \rightarrow & 000110010100111001010100111000101100101011000101. \end{cases}$$

Then h_4 is $(3,3)$-abelian-square-free. Moreover, for every abelian-square-free word w, $h_4(w)$ contains only 4 distinct 3-abelian-squares: 0^2, 1^2, $(01)^2$ and $(10)^2$.

The proof that h_4 is $(3,3)$-abelian-square-free is also done by a computer check, similar to the proof of Theorem 1. One can then check that $(00)^2$ and $(11)^2$ do not appear as factors of any image of a two-letter word.

Finally, using again an exhaustive search we were able to give the lower bound $g(2) \geq 5$.

Proposition 5. *Every word of size more than 92 over the binary alphabet contains at least 5 distinct 2-abelian-squares.*

The following question is a stronger version of Question 5:

Question 6. How many distinct 2-abelian-squares must an infinite binary word contain?

6. Computer searches

The toughest part to prove the existence of an infinite word with a desired property (say, the property P) is to find a morphism with sufficient conditions whose fixed point has the property P. Let $\mathcal{W}$ be the family of finite words with the property P. This part was done by a computer-assisted search, as follows. We look for a morphism whose images share a long common prefix and a long common suffix, to avoid the creation of small forbidden patterns when we concatenate two images. For this, we first selected one "good" factor $w = uv$ (with $|u| = |v|$), where v will be the common prefix, and u the common suffix. To optimize the chance of success, we took a factor w which is supposed to appear often in words in $\mathcal{W}$: we constructed, with a backtrack algorithm, a long random word in $\mathcal{W}$, and selected the factor w among the factors which appear most often in this word. Then we constructed a family $\mathcal{F}_w$ of words in $\mathcal{W}$ with prefix u, suffix v, and with a size of at most a fixed number. (If the search fails, we try again with another w, or with a larger size.) We constructed the graph with vertex set $\mathcal{F}_w$, and edges $\{x, y\}$ such that xyx and yxy are in $\mathcal{W}$. Finally, we checked the sufficient conditions on every morphism which corresponds to a clique of size 4 in the graph.

As an example, for the morphism h_3, which was the most arduous to find among the morphisms presented here, we use the following parameters. We selected the 88 factors with a size of 300 which appear most often among approximately 1.2 million factors we found in the random word. We computed the families $\mathcal{F}_w$ with words of size at most 700. All the families $\mathcal{F}_w$ we computed had a size between 1,000 and 5,000. The graph of the family that gave us the morphism h_3 had 58,680 edges and 1,977 vertices (density of $\sim$0.03). With the right parameters ($|w|$ and length of the elements of $\mathcal{F}_w$) it took half a day to find a good morphism.

The presented approach cannot be directly used to answer positively to Question 5. For every binary word w, $|wx|_{10} = |wx|_{01}$ (where x is the first letter of w). So the matrix N has rank at most 3, and one cannot find an invertible sub-matrix M of size 4. Thus Question 5 is, in spirit, close to Mäkelä's questions (Question 2 and Question 3, the modified version of Question 1).

References

[1] A. Carpi, *On abelian power-free morphisms*, Internat. J. Algebra Comput. **3** (1993), no. 2, 151–167, DOI 10.1142/S0218196793000123. MR1233218 (94g:68087)

[2] A. Carpi, *On the number of abelian square-free words on four letters*, Discrete Appl. Math. **81** (1998), no. 1-3, 155–167, DOI 10.1016/S0166-218X(97)88002-X. MR1492008 (98j:68139)

[3] F. M. Dekking, *Strongly nonrepetitive sequences and progression-free sets*, J. Combin. Theory Ser. A **27** (1979), no. 2, 181–185, DOI 10.1016/0097-3165(79)90044-X. MR542527 (81b:05027)

[4] R. C. Entringer, D. E. Jackson, and J. A. Schatz, *On nonrepetitive sequences*, J. Combinatorial Theory Ser. A **16** (1974), 159–164. MR0332533 (48 #10860)

[5] P. Erdős, *Some unsolved problems*, Michigan Math. J. **4** (1957), 291–300. MR0098702 (20 #5157)

[6] P. Erdős, *Some unsolved problems*, Magyar Tud. Akad. Mat. Kutató Int. Közl. **6** (1961), 221–254. MR0177846 (31 #2106)

[7] A. A. Evdokimov, *Strongly asymmetric sequences generated by a finite number of symbols.* (Russian), Dokl. Akad. Nauk SSSR **179** (1968), 1268–1271. MR0234842 (38 #3156)

[8] A. S. Fraenkel and R. J. Simpson, *How many squares must a binary sequence contain?*, Electron. J. Combin. **2** (1995), Research Paper 2, approx. 9 pp. (electronic). MR1309124 (96b:11023)

[9] J. Karhumaki, A. Saarela, and L. Q. Zamboni, *On a generalization of Abelian equivalence and complexity of infinite words*, J. Combin. Theory Ser. A **120** (2013), no. 8, 2189–2206, DOI 10.1016/j.jcta.2013.08.008. MR3102181

[10] V. Keränen, *Abelian squares are avoidable on 4 letters*, Automata, Languages and Programming (Vienna, 1992), Lecture Notes in Comput. Sci., vol. 623, Springer, Berlin, 1992, pp. 41–52, DOI 10.1007/3-540-55719-9_62. MR1250629 (94j:68244)

[11] Veikko Keränen, *New abelian square-free DT0L-languages over 4 letters.*, Manuscript (2003).

[12] M. Lothaire, *Combinatorics on Words*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1997. MR1475463 (98g:68134)

[13] P. A. B. Pleasants, *Non-repetitive sequences*, Proc. Cambridge Philos. Soc. **68** (1970), 267–274. MR0265173 (42 #85)

[14] M. Rao, *On some generalizations of abelian power avoidability*, Theoret. Comput. Sci. **601** (2015), 39–46, DOI 10.1016/j.tcs.2015.07.026. MR3396334

[15] A. Thue, *Über die gegenseitige lage gleicher teile gewisser zeichenreihen.*, Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania **10** (1912), 1–67.

École Normale Supérieure de Lyon, LIP (UMR 5668 CNRS, ENSL, Inria, UCBL, UDL), 46 allée d'Italie, 69364 Lyon Cedex 07, France

E-mail address, M. Rao: `michael.rao@ens-lyon.fr`

École Normale Supérieure de Lyon, LIP (UMR 5668 CNRS, ENSL, Inria, UCBL, UDL), 46 allée d'Italie, 69364 Lyon Cedex 07, France

E-mail address, M. Rosenfeld: `matthieu.rosenfeld@ens-lyon.fr`