RAIRO-Theor. Inf. Appl. 46 (2012) 147–163
DOI: 10.1051/ita/2011127

Available online at:
www.rairo-ita.org

ON ABELIAN REPETITION THRESHOLD

Alexey V. Samsonov[1] and Arseny M. Shur[1]

Abstract. We study the avoidance of Abelian powers of words and consider three reasonable generalizations of the notion of Abelian power to fractional powers. Our main goal is to find an Abelian analogue of the repetition threshold, *i.e.*, a numerical value separating k-avoidable and k-unavoidable Abelian powers for each size k of the alphabet. We prove lower bounds for the Abelian repetition threshold for large alphabets and all definitions of Abelian fractional power. We develop a method estimating the exponential growth rate of Abelian-power-free languages. Using this method, we get non-trivial lower bounds for Abelian repetition threshold for small alphabets. We suggest that some of the obtained bounds are the exact values of Abelian repetition threshold. In addition, we provide upper bounds for the growth rates of some particular Abelian-power-free languages.

Mathematics Subject Classification. 68Q70, 68R15.

1. Introduction

The study of avoidable powers of words has more than a centennial history since the paper by Thue [17]. Recall that for any finite word w its 2nd power (or *square*) is just the word ww, denoted by w^2, its 3rd power (or *cube*) is $w^3 = www$, and so on. Further, the notion of integral power of a word can be easily generalized to non-integral powers as follows. If w is a word, $|w|$ is its length, $\beta > 1$ is a number, then w^β is a unique prefix v of the infinite word $www\ldots$, whose length satisfies the conditions $|v|/|w| \geq \beta$, $(|v| - 1)/|w| < \beta$. A word u is β-*free*, if none of its factors, including u itself, is a β-power. A β-power is said to be k-*avoidable* if there

Keywords and phrases. Repetition threshold, formal languages, avoidable repetitions, Abelian powers.

[1] Institute of Mathematics and Computer Science, Ural Federal University, 620083 pr. Lenina, 51 Ekaterinburg, Russia. vonosmas@gmail.com; Arseny.Shur@usu.ru

Article published by EDP Sciences

are infinitely many β-free words (or, equivalently, an infinite β-free word) over the k-letter alphabet, and *k-unavoidable* otherwise.

For any k-letter alphabet ($k \geq 2$), the *repetition threshold* is the number $RT(k)$ which separates k-unavoidable and k-avoidable powers of words. Famous Dejean's conjecture [8] states that $RT(3) = 7/4$, $RT(4) = 7/5$, and $RT(k) = k/(k-1)$ otherwise. The recent proof of this conjecture (see [4,7,14]) closed a whole chapter in combinatorics on words but also opened a way to new challenging problems. One of these problems is to obtain a similar characterization of power-free languages in the Abelian case.

Abelian powers of words were first considered by Erdös [10] as a natural generalization of "usual" powers. The word $w_1 w_2 \dots w_n$ is an *Abelian nth power*, if each of the words $w_2, \dots, w_n$ is an anagram of w_1. Equivalently, one can say that $w_1 = w_2 = \dots = w_n$ in any commutative semigroup, or that $w_1, \dots, w_n$ share the same Parikh vector. It is clear that, in contrast with the usual powers, there are several ways to generalize the notion of Abelian power to fractional exponents. Below we define weak, semistrong and strong Abelian fractional powers and then work with all three definitions.

The avoidability of Abelian integral powers is well studied. It is easy to check that Abelian squares are 3-unavoidable and Abelian cubes are 2-unavoidable. On the other hand, the language of all quaternary Abelian-square-free words is infinite [11] and even has exponential growth [3, 12]. The same is true for ternary Abelian-cube-free and binary Abelian-4-free languages, see [9] for infiniteness and [1, 6] for exponential growth. The corresponding estimates for the number of quaternary Abelian-square-free, ternary Abelian-cube-free, and binary Abelian-4-free words are $\Omega(1{,}02306^n)$, $\Omega(1{,}02930^n)$, and $\Omega(1{,}04427^n)$, respectively. These estimates give very rough lower bounds for the exponential growth rates of these languages. In this paper, we give upper bounds for these growth rates; these bounds look more reliable to represent the actual growth of the studied languages. Our bounds are obtained by implementation of a more general method we present in this paper.

Once a definition of Abelian fractional power is chosen, Abelian repetition threshold can be defined in the same way as the "usual" repetition threshold. We study the values of Abelian repetition threshold for all alphabets and three different definitions of Abelian fractional power. The paper contains both analytic and computer-assisted results.

A method to obtain upper bounds for the growth rates of power-free languages is proposed in [16]. On the base of this method we construct our main instrument, which is a method to obtain upper bounds for the growth rates of *Abelian*-power-free languages.

2. Preliminary considerations

We omit basic definitions on words and languages, assuming that the reader is familiar with them. Let $\Sigma = \{1, \dots, k\}$ be an alphabet and $w \in \Sigma^*$ be an arbitrary

k-ary word. The Parikh vector $\vec{p}(w)$ is the vector of length k whose ith component equals the number of occurrences of the letter i in w, for any $i = 1, \ldots, k$. If $v \in \Sigma^*$, then the notation $\vec{p}(w) \leq \vec{p}(v)$ means that the ith component of $\vec{p}(w)$ is not greater than the ith component of $\vec{p}(v)$, for any $i = 1, \ldots, k$.

A language $L \subseteq \Sigma^*$ is *factorial* if it is closed under taking factors of its words. A word w is *forbidden* for a given language L if it is not a factor of any word from L. The set of all minimal (with respect to the factor order) forbidden words for L is called the *antidictionary* of L. A factorial language $L \subseteq \Sigma^*$ with the antidictionary M satisfies the equalities $L = \Sigma^* - \Sigma^* M \Sigma^*$, $M = \Sigma L \cap L\Sigma \cap (\Sigma^* - L)$. Thus, each antidictionary determines a unique factorial language, which is regular if the antidictionary is also regular (in particular, finite).

Remark 2.1. Infinite regular languages contain arbitrarily big powers of words. If the antidictionary of an infinite language L is finite, then L is regular and cannot be a power-free language. Hence, infinite power-free (or Abelian-power-free) languages have infinite antidictionaries.

The "size" of a language L can be expressed by its *combinatorial complexity*, which is the function $C_L(n) = |\Sigma^n \cap L|$. *Growth rate* of L roughly describes the behaviour of combinatorial complexity and is defined by the equality $\alpha(L) = \limsup_{n\to\infty}(C_L(n))^{1/n}$. If a regular language L is given by a deterministic finite automaton (dfa) $\mathcal{A}$ with the property that each vertex is visited during at least one successful computation, then $\alpha(L)$ coincides with the *Frobenius root* (or *spectral radius*) of the adjacency matrix of $\mathcal{A}$. This result is probably folklore; a short proof of it can be found in [15]. The dfa's with the mentioned property are said to be *consistent*.

2.1. Abelian powers

Let $m \geq 2$ be an integer. As we have said above, an *Abelian m-power* is a word of the form $w_1 w_2 \ldots w_m$, where w_i is an anagram of w_1 for $2 \leq i \leq m$, or, in other words, $\vec{p}(w_1) = \ldots = \vec{p}(w_m)$. Now we extend this definition to the rational numbers in the range $(1, \infty)$. Let $\beta > 1$, $|w_1| = q$, $m = \lfloor \beta \rfloor$, $t = \lceil \{\beta\} q \rceil$, where $\{\beta\}$ stands for the fractional part of β. Consider a word of the form $w = w_1 \ldots w_m v$, where $w_1 \ldots w_m$ is an Abelian m-power and $|v| = t$. Throughout this paper, the terms *root* and *tail* with respect to Abelian power denote the words w_1 and v respectively. Clearly, to call the word w an Abelian β-power we should impose an additional restriction upon the Parikh vector of the tail. We consider three different such restrictions, thus obtaining three definitions of Abelian fractional power. Let $\mathrm{pref}(u, l)$ be the prefix of length l of the word u.

A *weak Abelian β-power* is a word w of the form described above such that $\vec{p}(v) \leq \vec{p}(w_1)$. That is, the tail is a prefix of an anagram of the root.

A *strong Abelian β-power* is a word w of the form described above such that $\vec{p}(v) = \vec{p}(\mathrm{pref}(w_1, t))$. That is, the tail is an anagram of a prefix of the root. However, in this definition we clearly distinguish the root among all words w_i,

since the tail does not depend on the order of letters in w_i for $2 \leq i \leq m$. After swapping w_1 and w_2 the word may lose (or gain) its property of being a strong Abelian power. For example, the word $abc\,cab\,ba$ is a strong Abelian 8/3-power, while the word $cab\,abc\,ba$ is not. To remedy this, we introduce the following definition:

A *semistrong Abelian β-power* is a word w of the form described above such that $\vec{p}(v) \leq \bigvee_{i=\overline{1,m}} \vec{p}(\mathrm{pref}(w_i, t))$, where $\bigvee$ is the operation of taking maximum componentwise. Thus, all w_i's are used symmetrically in the restriction imposed upon the tail, like in the definition of the weak powers.

Remark 2.2.

(1) For integral values of β all three definitions are equivalent to the definition of the integral Abelian power;
(2) for $\beta \leq 2$ the definitions of strong and semistrong Abelian β-powers are equivalent;
(3) every strong Abelian β-power is also a semistrong Abelian β-power, and every semistrong Abelian β-power is also a weak Abelian β-power.

Example 2.3. The word $abc\,cba\,ac$ is a semistrong Abelian (8/3)-power and is also a weak Abelian (8/3)-power, but not a strong Abelian (8/3)-power, because ac is not a permutation of ab. The word $abcaa$ is not even a weak Abelian (5/3)-power, but is a strong, semistrong and weak Abelian (5/4)-power.

Remark 2.4. There are several possible "symmetrizations" of the notion of strong Abelian β-power. The above notion of semistrong Abelian β-power uses the "weakest" restriction on $\vec{p}(v)$. It will be seen in Section 4 that the suggested values of Abelian repetition threshold coincide for strong and semistrong powers. So, if we are aimed at Abelian repetition threshold, it makes no sense to study other symmetrizations.

A common weakness of all studied notions of Abelian fractional powers is the lack of symmetry under reversal: the reversal of an Abelian β-power is not necessarily an Abelian β-power, except for the case of strong/semistrong powers if $\beta \leq 2$. But any reversal-preserving "version" of Abelian fractional power should have two tails (on the left and on the right), and thus does not correlate with the notion of integral Abelian power. For example, an Abelian 3-power can have the form $v_1w_1w_2v_2$ instead of $w_1w_2w_3$ (here the tails v_1 and v_2 have the total length equal to the length of the root). As a result, avoidance properties of such "two-tailed Abelian powers" do not relate to the properties of integral Abelian powers. Since we are interesting just in the avoidance properties, we exclude two-tailed Abelian powers from consideration.

2.2. Abelian-power-free languages

Abelian exponent of a word $w \in \Sigma^*$ is the maximal rational number β such that w is an Abelian β-power. A word w is called *Abelian-β-free* (*Abelian-β^+-free*)

if all its factors have Abelian exponents less than β (respectively, at most β). It is convenient to use only the term β-free, assuming that β belongs to the set of "extended rationals". This set consists of all rational numbers and all such numbers with a plus; the number x^+ covers x in the usual $\leq$ order in a way that the inequalities $y \leq x$ and $y < x^+$ are equivalent. By *Abelian-β-free languages* we mean the languages of *all* Abelian-β-free words over a given alphabet. These languages are factorial and are called *Abelian-power-free* languages. We consider three types of Abelian-power-free languages (weak, semistrong and strong), corresponding to the three definitions of fractional Abelian powers. When necessary, we add the attribute strong (semistrong, weak) to any of the defined notions to specify the type of Abelian power.

To compare these definitions, it is reasonable to compare the sizes of the corresponding Abelian-power-free languages. Thus, we need a method for estimating the growth rates of the Abelian-power-free languages for weak, semistrong and strong Abelian powers. First, recall the method of [16] for power-free languages.

2.3. Estimating growth rates of power-free languages

To obtain the upper bounds for the growth rates of factorial languages one can use languages with finite antidictionary as follows. Let $L \subseteq \Sigma^*$ be a factorial language with the antidictionary M. Consider a family $\{M_i\}$ of finite subsets of M such that

$$M_1 \subseteq M_2 \subseteq \cdots \subseteq M_i \subseteq \cdots \subseteq M, \quad M_1 \cup M_2 \cup \cdots \cup M_i \cup \cdots = M.$$

Denote by L_i the factorial language over Σ with the antidictionary M_i. One has

$$L \subseteq \cdots \subseteq L_i \subseteq \cdots \subseteq L_1, \quad L_1 \cap L_2 \cap \cdots \cap L_i \cap \cdots = L.$$

It is not hard to show that the sequence $\{\alpha(L_i)\}$ decreases and converges to $\alpha(L)$. Since the languages L_i are regular, the number $\alpha(L_i)$ can be found with any degree of precision. Increasing i, one can make the upper bound arbitrarily close to $\alpha(L)$.

Thus, to obtain an upper bound for $\alpha(L)$ one should make three steps. First, build the antidictionary M_i for the chosen i. Second, convert this antidictionary into a consistent deterministic finite automaton (dfa) recognizing L_i. And finally, calculate the number $\alpha(L_i)$.

For a power-free (say, β-free) language these steps can be made as follows. We define M_i to be the set of all minimal forbidden words u^β such that $|u| \leq i$, calculate M_i by some advanced search procedure and store it as a trie. Then we use a modification of the Aho-Corasick algorithm for pattern matching to convert the trie into the consistent dfa recognizing L_i. Finally, we calculate $\alpha(L_i)$ with any prescribed precision by an efficient (linear in the size of the automaton) iterative algorithm. A rather complicated but very useful trick allows one to improve the above scheme, shrinking the sizes of the trie and the automaton by the factor of almost $|\Sigma|$! This trick can be adopted for any factorial language L which is *symmetric* in the sense that $u \in L$ implies $\sigma(u) \in L$ for any permutation σ of the

alphabet. This property allows one to represent any equivalence class of symmetric words by a single word. So, we build only the words from the antidictionary which are lexicographically minimal in their symmetry classes (these words form a "subtrie" of the whole trie) and represent the transitions in the whole automaton by transitions between the vertices of this subtrie. In what follows, the words forming this subtrie are referred to as *lexmin words*.

The mentioned algorithms for the second and third steps can be used for any symmetric factorial language. However, the first step can vary significantly for different subclasses of factorial languages.

In practice, the described method for power-free languages works well enough, allowing one to construct and handle huge automata in a short time. So, if we can efficiently organize the first step for Abelian-power-free languages, we will get a very powerful instrument.

3. Lower bounds for Abelian repetition threshold

Like Dejean in [8], we begin the study of the Abelian repetition threshold with the lower bounds. Let $\Sigma = \{1, \dots, k\}$. We prove uniform lower bounds for both strong and weak Abelian repetition threshold (denoted by $ART_s(k)$ and $ART_w(k)$, respectively). In view of the numerical results of Section 4, it looks highly probable that our bound for $ART_s(k)$ is exact, while the bound for $ART_w(k)$ can be improved. In what follows, we prove:

Theorem 3.1. *$ART_s(k) \geq \frac{k-2}{k-3}$ for all $k \geq 5$.*

Let $w = w_1 \dots w_m \in \Sigma^*$, where $w_1, \dots, w_m$ are letters in Σ. We say that a factor u of the word w is an *l-factor* if $|u| = l$. The jth l-factor is the factor $w_j \dots w_{j+l-1}$. For any j, the jth and $(j+1)$th l-factors of w are called *successive*. We need two lemmas.

Lemma 3.2. *Suppose that $k \geq 4$, $w \in \Sigma^*$ is a strong (=semistrong) Abelian-$\frac{k-2}{k-3}$-free word.*

(1) *Each $(k-2)$-factor of w consists of $(k-2)$ different letters*;
(2) *at least one of any two successive $(k-1)$-factors of w consists of $(k-1)$ different letters.*

Proof.

(1) Let $w = w_1 \dots w_m$. If $w_i = w_j$, $i < j$, then $w_i \dots w_j$ is an Abelian $\frac{j-i+1}{j-i}$-power. Since w is Abelian-$\frac{k-2}{k-3}$-free, we get $j - i \geq k-2$, whence the result;

(2) let $u = w_i \dots w_{i+k-2}$ and $v = w_{i+1} \dots w_{i+k-1}$ be two successive $(k-1)$-factors of w. Since u and v are Abelian-β-free, it follows from (1) that the first $(k-2)$ letters of u are different, as well as the last $(k-2)$ letters of both u and v. Suppose that both u and v contain equal letters. This can happen only if $w_i = w_{i+k-2}$ and $w_{i+1} = w_{i+k-1}$, which implies that the word $w_i \dots w_{i+k-1}$ is a strong Abelian $\frac{k}{k-2}$-power. Since $\frac{k}{k-2} \geq \frac{k-2}{k-3}$ for $k \geq 4$, w is not Abelian-$\frac{k-2}{k-3}$-free, and we get a contradiction. Hence, either u or v consists of different letters. □

According to Lemma 3.2, the k-ary $\frac{k-2}{k-3}$-free words have two types of $(k-1)$-factors: *quasipermutations*, consisting of different letters, and *repeats*, in which the first and the last letter coincide. By *permutation* we mean a k-factor consisting of k different letters.

Remark 3.3. Abelian $\frac{k+1}{k-1}$-powers (respectively, $\frac{k+2}{k}$-powers) are forbidden for the k-ary $\frac{k-2}{k-3}$-free language whenever $k \geq 5$ (respectively, $k \geq 6$).

Lemma 3.4. *For $k \geq 6$, any k-ary Abelian-$\frac{k-2}{k-3}$-free word beginning with a quasipermutation and containing no permutations as factors has length at most $4k-2$.*

Proof. Let w be a word satisfying the conditions of the lemma. Suppose that jth and $(j+1)$th $(k-1)$-factors of w are quasipermutations. Then they consist of the same letters. Hence, the $(j+2)$th $(k-1)$-factor of w is a repeat, otherwise w contains an Abelian $\frac{k+1}{k-1}$-power, which is impossible by Remark 3.3.

Aimed at a contradiction, let $|w| \geq 4k-1$. Then the number of $(k-1)$-factors in w is at least $3k+1$. Hence, w contains at least k repeats and also a quasipermutation to the right of the kth (counting from left to right) of these repeats. Since there are only k possible sets of letters for quasipermutations, w contains two quasipermutations, which consist of the same letters and are not successive. Let ith $(k-1)$-factor z and jth $(k-1)$-factor $\bar{z}$ of w be such quasipermutations such that the difference $j-i>1$ is minimal. If z and $\bar{z}$ overlap in w, they form a factor $xy\bar{x}$ such that $xy=z$ and $y\bar{x}=\bar{z}$. Obviously, $|x| \geq 2$ and $\bar{x}$ is an anagram of x. Thus, $xy\bar{x}$ is an Abelian β-power for some $\beta \geq \frac{k+1}{k-1}$, which is impossible by Remark 3.3.

Now suppose that z and $\bar{z}$ do not overlap in w. Since two successive quasipermutations in w consist of the same letters, by minimality of $j-i$ we get that the $(i+1)$th and the $(j-1)$th $(k-1)$-factors of w are repeats. Consider the word $u = w_i \dots w_{j+k-2} = zy\bar{z}$. By minimality of $j-i$, u contains at most k repeats, and thus, at most $3k$ $(k-1)$-factors. So we have $|u| \leq 4k-2$. The word u is an Abelian $\frac{|u|}{|u|-k+1}$-power. For $k \geq 7$,

$$\frac{|u|}{|u|-k+1} \geq \frac{4k-2}{3k-1} \geq \frac{k-2}{k-3},$$

contradicting to the fact that w is Abelian-$\frac{k-2}{k-3}$-free. A more detailed analysis is needed for $k=6$. We will prove that in this case the word u contains less than $2k=12$ quasipermutations. If this number is achieved, u contains five pairs of successive quasipermutations in addition to z and $\bar{z}$. Then the beginning of u looks as follows, up to the renaming of letters (recall that u contains no permutations). Under each letter, we indicate the type of the $(k-1)$-factor which begins in the position of this letter (q = quasipermutation, r = repeat):

1 2 3 4 5 2 6 3 **5 1 2 3 4** . . .
q r q q r q q r q . . .

We see that u contains another anagram of z (written in boldface), a contradiction with the choice of $\bar{z}$. In fact, we proved that the sequence of $(k-1)$-factors of u has no factor qrqqrqqrq. Thus, the longest sequence of $(k-1)$-factors of u is qrqqrqrqqrqrqqrq. So we have $|u| \leq 20$. Therefore, u is an Abelian $(4/3)$-power, while w is Abelian-$(4/3)$-free by conditions of the lemma. This contradiction concludes the proof. □

Proof of Theorem 3.1. Suppose that the exponent $\frac{k-2}{k-3}$ is Abelian-k-avoidable, and W is a k-ary Abelian-$\frac{k-2}{k-3}$-free infinite word. First we consider the case $k \geq 6$. By Lemmas 3.2, 3.4, infinitely many k-factors of W are permutations. Note that at least one of any three successive k-factors of W is not a permutation (otherwise, W contains an Abelian $\frac{k+2}{k}$-power, which is impossible by Rem. 3.3). So, we can pick a pair of indices i, j such that $i+1 < j$, ith and jth k-factors of W are permutations, and rth k-factor of W is not a permutation whenever $i < r < j$.

If the two chosen permutations overlap in W, they form a factor $xy\bar{x}$ such that $|xy| = k$, $|x| \geq 2$, and $\bar{x}$ is an anagram of x. This factor is an Abelian β-power for some $\beta \geq \frac{k+2}{k}$, which is impossible by Remark 3.3. On the other hand, if these permutations do not overlap in W, then W has a factor $u = xy\bar{x}$ such that x and $\bar{x}$ are permutations (and hence, anagrams of each other). If we delete the first and the last letters of u, we will get a $\frac{k-2}{k-3}$-free word which begins with a quasipermutation and contains no permutations as factors. By Lemma 3.4, $|u| - 2 \leq 4k-2$, *i.e.* $|u| \leq 4k$. Hence, u contains an Abelian β-power for some $\beta \geq 4/3$. This contradiction shows that W cannot exist, so the exponent $\frac{k-2}{k-3}$ is Abelian-k-unavoidable for any $k \geq 6$.

The case $k = 5$ requires a separate analysis. This is a dull (although not very long) case examination. So, we omit it, giving only statistics confirmed by computer search: the antidictionary of the 5-ary $(3/2)$-free language contains only 49 lexmin words, and the maximum length of the root of such a word is 10. □

Now we move to weak Abelian powers. In what follows, we prove

Theorem 3.5. *$ART_w(k) \geq \frac{k}{k-2}$ for all $k \geq 10$.*

The proof of this theorem relies on two lemmas. We say that a letter w_i of the word $w \in \Sigma^*$ is *old* if $w_i = w_j$ for some $j < i$, and *new* otherwise[2].

Lemma 3.6. *Suppose that $k \geq 8$, $w \in \Sigma^*$ is a k-ary weak Abelian-$\frac{k}{k-2}$-free word, $1 \leq l \leq 4$, and w ends with l old letters. Then $2|w| > lk$.*

Proof. All l last letters of w are different, otherwise they form a weak Abelian power of exponent at least $\frac{4}{3}$, which is greater than $\frac{k}{k-2}$ for $k \geq 8$. The last l letters are old, so all of them occur somewhere in the word $w_1 \dots w_{|w|-l}$. Hence, w is a weak Abelian power of exponent $\frac{|w|}{|w|-l} < \frac{k}{k-2}$, whence $2|w| > lk$. □

[2]More rigorous, not letters themselves, but their particular occurrences are old/new; but our loose treatment makes the proof more readable.

Corollary 3.7. *Suppose that $k \geq 8$, $w = pqrs$ is a k-ary weak Abelian-$\frac{k}{k-2}$-free word, $|p| = |r| = \lfloor k/2 \rfloor$, $|q| = |s| = \lceil k/2 \rceil$. Then*

(1) *p contains no old letters*;
(2) *pq contains no two successive old letters*;
(3) *pqr contains no three successive old letters*;
(4) *$w = pqrs$ contains no four successive old letters.*

Proof. If we consider a prefix of $pqrs$ ending with a group of successive old letters, we get the bound for the length of this group from Lemma 3.6. □

Remark 3.8. Suppose that $k \geq 8$, w is a k-ary weak Abelian-$\frac{k}{k-2}$-free word, and $|w| = \lfloor \frac{k}{2} \rfloor$. Then w consists of $\lfloor \frac{k}{2} \rfloor$ different letters.

Lemma 3.9. *Suppose that $k \geq 10$, $w \in \Sigma^*$ is a k-ary weak Abelian-$\frac{k}{k-2}$-free word, and $|w| = 2k$. Then w contains all k letters from Σ^*.*

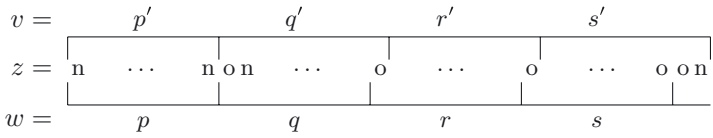
Proof. Choose a word w of length $2k$ in which any new letter appears as late as possible in view of Corollary 3.7. That is, if we write $w = pqrs$, where $|p| = |r| = \lfloor \frac{k}{2} \rfloor$ and $|q| = |s| = \lceil \frac{k}{2} \rceil$, then each new letter in q (in r, in s) follows one (respectively, two and three) old letters in w. We call such a word w *late*. Obviously, it is enough to prove that any late word contains k new letters. We prove this by induction. For the inductive base $k = 10$, the location of new letters in w is given by the following on-*encoding* (o = old, n = new):

$$\text{nnnnn onono onoon ooono}, \tag{3.1}$$

so the required statement holds. We consider "even" ($k = 2l - 1$ to $2l$) and "odd" ($k = 2l$ to $2l + 1$) inductive steps separately. During even step, the length of p and r increases, while the length of q and s increases during odd step.

Even step is easy. Take the on-encoding of a late word $w = pqrs$ of length $2k$ and add n to its beginning and o to its end. If r does not end by two old letters, then we get the on-encoding of a late word of length $2k+2$. Otherwise, to get such an on-encoding we should just move all the n's to the right of these two letters by one symbol to the left (because the first letter of s has become the last letter of r). Anyway, a late word of length $2k + 2 = 4l$ contains $k + 1$ new letters. We also note that the last letter of the late word of length $4l$ is old (this assertion holds for the inductive base as well).

Now we prove the odd step. Let $v = p'q'r's'$ be a late word of length $2k + 2$. We take the on-encoding of a late word $w = pqrs$ of length $2k = 4l$, add o and n to its end, and compare the resulting word $z \in \{\text{o}, \text{n}\}^*$ with the on-encoding of v. As was shown above, the last letter of w is old:



The word z contains $k+1$ n's. Each letter n in z, the position of which corresponds to the part q' (part r', part s') of the word v, follows at least one (respectively, two and three) letters o. The only possible exception is the last letter n in z. So, if w ends with two old letters, then no exception occurs, and the position of jth new letter in v is less than or equal to the position of jth letter n in z for any $j = 1, \ldots, k+1$. As a result, v certainly contains $k+1$ new letters, and we are done. The case when the penultimate letter of w is new implies the exception and is analyzed in the last part of the proof.

Assuming that the on-encoding of w ends with no, we analyze three cases depending on the position j of the first new letter in s. If j is the second position of s, we shift all n's, corresponding to the part s of w, by one symbol to the left in the word z. Since the position $j-1$ in this case corresponds to the part r' of v, the obtained word z' possesses the desired property: each letter n, the position of which corresponds to the part q' (part r', part s') of the word v, follows at least one (respectively, two and three) letters o. As above, we conclude that v contains $k+1$ new letters. If j is the fourth position of s, then $l = |s|$ is odd, because the positions of new letters in s are equal modulo 4. Then q ends with an old letter, implying that the second letter of r is new. We shift all n's, corresponding to the parts r and s of w, by one symbol to the left in the word z. Repeating the above argument, we get that v contains $k+1$ new letters.

It remains to consider the case when j is the third position of s. Then $l = |s|$ is divisible by 4. Hence, the last letter of q is new. Since the last letter of q and the third letter of s are new, the third and the penultimate letters of r are also new. Therefore, $l = |r| = 3l' + 1$. Note that q and s contain $\frac{l}{2}$ and $\frac{l}{4}$ new letters, respectively, while r contains $\frac{l-1}{3}$ new letters. Since $l = k/2 \geq 5$, we have $l \geq 16$. Then we have $\frac{l-1}{3} > \frac{l}{4}$, implying that w contains more than k new letters, a contradiction. The lemma is proved. □

Proof of Theorem 3.5. Let $w = uv$ be a k-ary word, $|u| = 2k$, $v = \lfloor k/2 \rfloor$. Aimed at a contradiction, we assume that w is a weak Abelian-$\frac{k}{k-2}$-free word. By Remark 3.8, all letters in the word v are different. By Lemma 3.9, the word u contains all k letters. Hence, all letters of v are contained in u, so w is a weak Abelian power of exponent $\frac{2k+\lfloor k/2 \rfloor}{2k}$ which is not less than $\frac{k}{k-2}$ for $k \geq 10$. This is impossible as w is weak Abelian-$\frac{k}{k-2}$-free.

So, any k-ary word of length at least $2k + \lfloor k/2 \rfloor$ has a forbidden factor. This means that the language of k-ary weak Abelian-$\frac{k}{k-2}$-free words is finite, implying $ART_w(k) \geq \frac{k}{k-2}$. □

4. Antidictionaries of Abelian-power-free languages

Now we prepare computer-assisted studies. Let $\beta > 1$ be an extended rational number. In order to estimate the growth rate of the Abelian-β-free language by the method described in Section 1.3, we should construct the antidictionary of this language. If the language is infinite, the antidictionary is also infinite by

Remark 2.1, so we actually construct its finite approximation. More precisely, we find all minimal forbidden words whose root is of length at most R for some integer R. Next lemma easily follows from the definitions of Abelian fractional powers.

Lemma 4.1. *If a word $w \in \Sigma^*$ is a weak (semistrong, strong) Abelian β-power and σ is a permutation of Σ, then $\sigma(w)$ is also a weak (respectively, semistrong, strong) Abelian β-power.*

Lemma 4.1 tells us that Abelian-power-free languages are symmetric. Hence, it is not necessary to construct the whole antidictionary; we can construct just the trie of all lexmin words instead. We store all possible roots of lexmin words in an auxiliary queue Q, iterating over roots in the order of increasing length. Note that these roots should be lexicographically minimal words in their symmetry classes. For convenience, we store with each word in the queue the number of different letters in it. For each root r, we try to construct all minimal forbidden words from it and add them to a trie which stores the lexmin words found so far. After that, we try to extend the root by appending letters from Σ to its end and push this longer root into the queue. Then we pop another root and continue the algorithm until all roots of length at most R are handled, see procedure ITERATE below. Recall that $\Sigma = \{1, \dots, k\}$. BUILD is a recursive procedure described below. Call of BUILD at line 4 results in adding all lexmin forbidden words with the root r to the trie.

Procedure ITERATE (outer cycle).

```
01. push (′1′; 1) into Q
02. while Q is not empty
03.       pop (r; t) from Q
04.       BUILD(r, r, p⃗(r))
05.       if |r| < R
06.          for each c ≤ t such that rc has no forbidden suffixes
07.              push (rc; t) into Q
08.          if t < k
09.              push (r(t + 1); t + 1) into Q
10. end while
```

The process of constructing all minimal forbidden words from a given root is quite time-consuming, as the number of these words can be very large. The dependence of their lengths on the length of the root is also not trivial, as the following lemma shows.

Lemma 4.2. *If $w \in \Sigma^*$ is a minimal forbidden word for an Abelian-β-free language and the length of its root is equal to R, then*

(1) $|w| = \lceil R \cdot \beta \rceil$ *for weak Abelian powers*;
(2) $\lceil R \cdot \beta \rceil \le |w| \le R \cdot \lceil \beta \rceil$ *for strong or semistrong Abelian powers, and this interval cannot be shortened in general.*

Proof. A minimal forbidden word for an Abelian-β-free language is at least an Abelian β-power, so the lower bound for $|w|$ follows from definitions. To prove (1), note that each prefix of a weak Abelian power is also a weak Abelian power with the same root. So, if the length of the word w exceeds $\lceil R \cdot \beta \rceil$, w will have a forbidden proper prefix. This note also proves the upper bound for (2) in view of Remark 2.2 (1). In the case of strong and semistrong Abelian powers the situation is more difficult, because not all their prefixes are Abelian powers. Thus, the Abelian square *abcde bdaec* of length 10 is also a minimal forbidden word for the strong (= semistrong) Abelian-(7/5)-free language, as one can check directly. This example proves that the upper bound in (2) is sharp. □

We construct lexmin forbidden words from a root recursively, appending letters to the end of a current word one by one until one of three exit conditions is fulfilled: current word is a lexmin forbidden word, current word is too long, or current word has a forbidden proper suffix. At each step we maintain the vector $\vec{p}$ of unused letters. When we append a letter c to our word, we decrease the corresponding component $[\vec{p}]_c$ of $\vec{p}$. But if by appending a letter we finish appending an anagram of the root, the next letter can again be any letter in the root, so we make $\vec{p}$ equal to the Parikh vector of the root. The formal description of this construction is given by procedure BUILD whose arguments are the current word w, the root r, and the current vector of unused letters $\vec{p}$. IS_ABELIAN is a subroutine which checks whether w is a (weak, semistrong, strong) Abelian power with the root r.

Procedure BUILD$(w, r, \vec{p})$ (constructing forbidden words).

```
01. if IS_ABELIAN(w, r)
02.    if |w| satisfies Lemma 4.2
03.       add w to the trie
04.    exit procedure
05. if |w| equals the upper bound from Lemma 4.2
06.    exit procedure
07. if |w| mod |r| = 0
08.    p = p(r)
09. for each letter c such that [p]_c > 0
10.     u = wc
11.     q = p
12.     [q]_c = [q]_c − 1
13.     if u has no forbidden proper suffixes
14.        BUILD(u, r, q)
15. end for
```

If the conditions at lines 1 and 2 hold, then the word w is an Abelian power with the root r and its Abelian exponent is at least β. The word w has no forbidden proper factors because (a) r is Abelian-β-free by construction, see Procedure ITERATE, and (b) the condition at line 13 is checked on the previous steps of recursion

for each prefix of w which is longer than r. So, w is indeed a minimal forbidden word and the operation at line 3 is justified.

We can easily maintain Parikh vectors of all prefixes of w. Therefore, we are able to calculate the Parikh vector of any factor of this word, subtracting the Parikh vectors of corresponding prefixes. This is enough to implement the subroutine IS_ABELIAN in constant time for each of three definitions of Abelian fractional power, assuming that $\lfloor\beta\rfloor$ is a constant.

Finally, we have to check the condition at line 13 efficiently. The following lemmas are useful in special cases:

Lemma 4.3. *If β is an integer and $w \in \Sigma^*$ is an Abelian β-power, then the reversal of w is also an Abelian β-power.*

Lemma 4.4. *Suppose that $u, v, w \in \Sigma^*$, u and v are suffixes of w, and $\bar{u}$ ($\bar{v}$) is the shortest suffix of w, which is a non-trivial weak Abelian power with the tail u (respectively, v). If $|u| < |v|$, then $|\bar{u}| \le |\bar{v}|$.*

Proof. Let $\bar{v} = \hat{v}v$. Then $\vec{p}(\hat{v}) \ge \vec{p}(v) \ge \vec{p}(u)$. It means that $\bar{v}$ is a weak Abelian power with the tail u by definition. But $\bar{u}$ is the shortest suffix of w with such property, so $|\bar{u}| \le |\bar{v}|$. □

Now we can formulate the main lemma, which estimates the running time of the procedure that looks for the forbidden suffixes. We consider the size of the alphabet as a constant.

Lemma 4.5. *Let $w \in \Sigma^*$, $|w| = n$, $\beta > 1$. During our algorithm we can check whether w has a suffix with Abelian exponent at least β:*

(1) *in $O(n)$ time, if β is an integer or if $\beta < 2$ and we consider weak Abelian powers;*
(2) *in $O(n^2)$ time in all other cases.*

Proof. If β is an integer, then at the moment we consider w we can be sure that all shorter forbidden words are already stored in the trie. According to Lemma 4.3, we can check prefixes of the reversal of the word w instead of suffixes of w. So, we read the word w backwards and replace the obtained word "on the fly" by the lexicographically minimal equivalent word. To do this replacement in $O(n)$ time, we store the images of replaced letters in an array of constant size. If the image of the current letter is undefined so far, it is set to the biggest existent image plus one. The obtained lexmin word is then considered as an input word for the current trie. If the trie reaches a terminal state, the forbidden prefix is detected. If the trie cannot reach such a state, then the input word has no forbidden prefixes. This check also takes linear time, so we are done with the integral powers.

Now suppose that β is not an integer and w has a forbidden suffix. Then its length is determined by the lengths of its tail and its root. We just try all possible values of these lengths (each of them does not exceed $n/\lfloor\beta\rfloor$) and check if the resulting suffix is actually a long enough Abelian power. This check is made in

constant time by the subroutine IS_ABELIAN we described earlier, thus giving us quadratic time overall. If $1 < \beta < 2$ and we consider weak Abelian powers, then, according to Lemma 4.4, the length of a suffix cannot decrease as we increase the length of its tail (note that Lem. 4.4 cares only about *shortest* Abelian powers with the given tail). So, in this case the whole procedure takes linear time. □

5. Numerical results

We present some of the results on the growth rates of Abelian power-free languages. These results are related to integral Abelian powers and fractional Abelian powers that appear to be close to the Abelian repetition threshold. In most cases our algorithm allowed us to build any antidictionary M_i such that the corresponding deterministic finite automaton (dfa) can be stored in the 2Gb memory of a PC. It means that our method in general proved efficient.

But we discovered that the sequences of upper bounds converge to the growth rate of the target languages extremely slow. For example, to get an upper bound for the growth rate of the binary "usual" cube-free language, one may take $R = 7$ and build a dfa with 246 vertices; the bound given by this automaton deviates by less than 0.001 from the precise value of the growth rate. On the other hand, we have a numerical evidence that any of our non-zero upper bounds for Abelian-power-free languages is more than 0.01 away from the actual value. It appears that lots of words contain only long forbidden Abelian powers. For the case of quaternary Abelian-square-free words, this phenomenon was also observed by Keränen [13].

Table 1 contains the calculated upper bounds for the growth rates of Abelian-β-free languages for integral numbers β. Using the described method, we were able to build antidictionaries containing millions of words. Although the obtained bounds are not very close to exact values, they can be used to compare the relative "sizes" of Abelian-power-free languages, for example, the ternary cube-free language seems to be much larger than binary 4-free one.

Table 2 is devoted to weak Abelian-power-free languages. The results listed in the table and some additional heuristics show that languages avoiding weak Abelian powers can have extremely large antidictionaries and still be finite. For example, we suppose that the languages of binary Abelian-$(11/3)^+$-free words and ternary Abelian-$(17/7)^+$-free words are finite, which makes finding the exact value of Abelian repetition threshold for small alphabets quite a challenging task.

Table 3 describes the Abelian-power-free languages for semistrong Abelian powers. The antidictionary of Abelian-square-free language over ternary alphabet contains only 7 lexmin words, but the ternary Abelian-2^+-free language seems to be exponential in case of semistrong Abelian powers. To justify Remark 2.4, it is enough to compare this table with Table 4.

Our last Table 4 describes strong Abelian-power-free languages. Here we can notice substantial gaps between the growth rates (and between the sizes of recognizing automata) that are due to the "allowance" of short factors with Abelian exponent β. For example, the allowance of a word aba results in a huge gap of more

TABLE 1. Abelian-power-free languages for integral powers. The columns contain (from left to right): size of the alphabet, avoided exponent, maximum length of a root, number of lexmin words, number of vertices in the dfa built from the trie of lexmin words, upper bound for the growth rate.

$\lvert\Sigma\rvert$	Exponent	Root	Words	Vertices	Growth rate (upper bound)
2	4	12	8 767 762	41 571 476	1.374164
3	3	8	6 015 458	15 187 934	2.371237
4	2	20	4 221 881	23 653 900	1.444344
5	2	10	6 420 827	16 081 994	3.227410

TABLE 2. Weak Abelian-power-free languages. The explanations for the columns are the same as in Table 1.

$\lvert\Sigma\rvert$	Exponent	Root	Words	Vertices	Growth rate (upper bound)
2	11/3	—	13 029	—	0.000000
2	$(11/3)^+$	25	22 440 239	52 403 705	1.055275
3	12/5	—	556 403	—	0.000000
3	17/7	24	6 128 640	7 500 382	1.081150
3	$(17/7)^+$	25	26 586 251	32 980 908	1.108788
4	$(13/7)^+$	—	>2 105 968	—	0.000000
4	15/8	31	1 408 013	7 038 325	1.101764
4	$(15/8)^+$	32	8 069 429	39 242 238	1.143995
5	5/3	—	18 150	—	0.000000
5	$(5/3)^+$	16	18 285 948	31 618 837	1.685733
6	3/2	—	434	—	0.000000
6	$(3/2)^+$	16	7 229 321	9 525 614	1.493064
7	3/2	—	>130 486	—	0.000000
7	$(3/2)^+$	11	7 658 103	4 658 332	3.580134
8	7/5	—	2 710 432	—	0.000000
8	10/7	16	28 069 579	23 580 728	1.506660
8	$(10/7)^+$	14	29 486 766	16 420 964	2.179884
9	4/3	—	>343 077	—	0.000000
9	$(4/3)^+$	14	16 240 204	6 943 950	2.471691

TABLE 3. Semistrong Abelian-power-free languages. The explanations for the columns are the same as in Table 1.

$\lvert\Sigma\rvert$	Exponent	Root	Words	Vertices	Growth rate (upper bound)
2	11/3	—	14 120	—	0.000000
2	$(11/3)^+$	21	12 706 824	42 549 418	1.137926
3	2^+	20	5 923 132	19 078 136	1.266646

TABLE 4. Strong Abelian-power-free languages. The explanations for the columns are the same as in Table 1.

$\lvert\Sigma\rvert$	Exponent	Root	Words	Vertices	Growth rate (upper bound)
2	18/5	—	>1 378 225	—	0.000000
2	11/3	34	10 199 584	57 443 698	1.020862
2	$(11/3)^+$	16	16 393 356	67 193 118	1.232531
3	2^+	13	12 113 648	37 673 788	1.645532
4	9/5	—	25 684	—	0.000000
4	$(9/5)^+$	30	3 116 205	17 942 923	1.175199
5	3/2	—	49	—	0,000000
5	$(3/2)^+$	13	6 999 188	21 445 164	2.334839
6	$(4/3)^+$	21	9 153 227	36 992 553	1.482134
7	$(5/4)^+$	22	2 957 234	12 260 556	1.472652
8	$(6/5)^+$	22	3 165 822	10 035 601	1.551357

than 2 in the growth rate for the Abelian-$(3/2)^+$-free language over the 5-letter alphabet. Unlike results for the weak Abelian powers, the data in Table 4 provides a strong evidence that the corresponding exponents coincide with the actual values of Abelian repetition threshold for strong Abelian powers, so we are able to formulate the following:

Conjecture 5.1. The Abelian repetition threshold for strong and semistrong Abelian powers is given by

$$ART_s(k) = \begin{cases} 11/3, & k = 2, \\ 2, & k = 3, \\ 9/5, & k = 4, \\ (k-2)/(k-3), & k \geq 5. \end{cases}$$

REFERENCES

[1] A. Aberkane, J.D. Currie and N. Rampersad, The number of ternary words avoiding Abelian cubes grows exponentially. *J. Integer Seq.* **7** (2004) 13 (electronic only).
[2] F.-J. Brandenburg, Uniformly growing k-th power free homomorphisms. *Theoret. Comput. Sci.* **23** (1983) 69–82.
[3] A. Carpi, On the number of Abelian square-free words on four letters. *Discrete Appl. Math.* **81** (1998) 155–167.
[4] A. Carpi, On Dejean's conjecture over large alphabets. *Theoret. Comput. Sci.* **385** (2007) 137–151.
[5] M. Crochemore, F. Mignosi and A. Restivo, Automata and forbidden words. *Inf. Process. Lett.* **67** (1998) 111–117.
[6] J.D. Currie, The number of binary words avoiding Abelian fourth powers grows exponentially. *Theoret. Comput. Sci.* **319** (2004) 441–446.
[7] J.D. Currie and N. Rampersad, A proof of Dejean's conjecture. *Math. Comput.* **80** (2011) 1063–1070.

[8] F. Dejean, Sur un théorème de Thue. *J. Comb. Th. (A)* **13** (1972) 90–99.
[9] F.M. Dekking, Strongly non-repetitive sequences and progression-free sets. *J. Comb. Th. (A)* **27** (1979) 181–185.
[10] P. Erdös, Some unsolved problems. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** (1961) 221–264.
[11] V. Keränen, Abelian squares are avoidable on 4 letters, in *Proc. ICALP'92. Lect. Notes Comput. Sci.* **623** (1992) 41–52.
[12] V. Keränen, A powerful abelian square-free substitution over 4 letters. *Theoret. Comput. Sci.* **410** (2009) 3893–3900.
[13] V. Keränen, Combinatorics on words – suppression of unfavorable factors in pattern avoidance. *TMJ* **11** (2010). Available at http://www.mathematica-journal.com/issue/v11i3/Keranen.html consulted in November 2011.
[14] M. Rao, Last cases of Dejean's conjecture. *Theoret. Comput. Sci.* **412** (2011) 3010–3018; *Combinatorics on Words (WORDS 2009), 7th International Conference on Words.*
[15] A.M. Shur, Comparing complexity functions of a language and its extendable part. *RAIRO-Theor. Inf. Appl.* **42** (2008) 647–655.
[16] A. M. Shur, Growth rates of complexity of power-free languages. *Theoret. Comput. Sci.* **411** (2010) 3209–3223.
[17] A. Thue, Über unendliche Zeichenreihen. *Kra. Videnskab. Selsk. Skrifter. I. Mat. Nat. Kl. Christiana* **7** (1906) 1–22.

Communicated by G. Richomme.
Received November 2, 2010. Accepted October 10, 2011.