On repetition-free binary words of minimal density

Roman Kolpakov, Gregory Kucherov, Yuri Tarannikov

On repetition-free binary words of minimal density

(Extended abstract)

Roman Kolpakov[*] Gregory Kucherov[†] Yuri Tarannikov[‡]

Abstract

In [12], a notion of minimal proportion (density) of one letter in n-th power-free binary words has been introduced and some of its properties have been proved. In this paper, we proceed with this study and substantially extend some of these results. First, we introduce and analyse a general notion of minimal letter density for any infinite set of words which don't contain a specified set of "prohibited" subwords. We then prove that for n-th power-free binary words, the density function is $\frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \mathcal{O}(\frac{1}{n^5})$ refining the estimate from [12]. Following [12], we also consider a natural generalization of n-th power-free words to x-th power-free words for real argument x. We prove that the minimal proportion of one letter in x-th power-free binary words, considered as a function of x, is discontinuous at all integer points $n \geq 3$. Finally, we give an estimate of the size of the jumps.

1 Introduction

In this paper we continue with the study initiated in [12]. The general problem behind this study can be described as follows. Assume we have specified a set of "prohibited" words $P \subseteq A^*$ and we are interested in the set $F \subseteq A^*$ of words that don't contain words from P as subwords. Words of F are said to *avoid* P. If the set F is infinite, that is there exists an infinite number of words without subwords from P, the set P is called *avoidable*, otherwise it is called *unavoidable*. One might specify, for example, a finite number of prohibited subwords P. Properties of unavoidable finite sets of words were studied in [13]. The set P of prohibited subwords can be infinite, in which case it may be specified by one or several *patterns*, i.e. words composed with variables and possibly with alphabet letters. Pattern avoidability has been subject of many works, and we refer to [7, 6] for an introduction to this area and a survey of known results. One might think of other ways of specifying the set of subwords to be avoided, e.g. as a language specified by a grammar. Note that for any set P of prohibited subwords, the set F of avoiding words is closed under taking subwords, and vice versa, any set F closed under subwords is the set of avoiding words for some P (just take $P = A^* \setminus F$). Therefore, being closed under subwords can be considered as a characterization for the sets of words that can be specified by means of prohibited subwords.

The case when the prohibited subwords are those of the form u^n, for some $n \geq 2$, has been extensively studied. Such subwords are called n-repetitions or n-powers, and words that don't

[*]French-Russian Institute for Informatics and Applied Mathematics, Moscow University, 119899 Moscow, Russia, e-mail: roman@vertex.inria.msu.ru

[†]INRIA-Lorraine/LORIA, 615, rue du Jardin Botanique, B.P. 101, 54602 Villers-lès-Nancy, France, e-mail: kucherov@loria.fr

[‡]Faculty of Discrete Mathematics, Department of Mechanics and Mathematics, Moscow State University, 119899 Moscow, Russia, e-mail: yutaran@nw.math.msu.su

contain such subwords are called n-th power-free. Back in the beginning of the century, Thue proved that there exist infinite 2-nd power-free (square-free) words over the three-letter alphabet, and 3-rd power-free (cube-free) words over the two-letter alphabet [15, 16] (see also [4]). To recast it in terms of pattern avoidability, Thue showed that pattern xx is avoidable on the three-letter alphabet, and pattern xxx is avoidable on the two-letter alphabet. Note that xx is unavoidable on the 2-letter alphabet, and this illustrates the fact that a set of subwords (or patterns) can be avoidable on some alphabet and unavoidable on a smaller alphabet. Paper [2] contains an example of a pattern which is avoidable on four letters but not on three letters. The question whether a given pattern is avoidable on the k-letter alphabet (for a given k) is not known to be decidable (see [9]) even if patterns are composed only of variables. In contrast, the question whether a given pattern is unavoidable on *any* alphabet has been shown decidable in [3, 1].

This paper is motivated by the following general question: If a pattern p is unavoidable on k letters but avoidable on $(k+1)$ letters, what is the minimal proportion (density) of a letter in words over $(k+1)$ letters avoiding p? In other terms, what is the minimal contribution (in terms of relative number of occurrences) of the $(k+1)$-st letter that allows to create words of unbounded length avoiding p? Note that the "minimal proportion" is understood here as the limit minimal proportion as the length of words goes to infinity. An answer to this question would establish a relationship between two properties of different kind: avoidance of a certain pattern (regularity) and proportion of occurrences of a letter.

To the best of our knowledge, minimal density has been first studied in a related paper [12]. However, some work had been done on counting limit densities of subwords in words defined by DOL-systems (cf e.g. [11]). In [12], this study was undertaken for the case of n-th power-free words on the 2-letter alphabet, and some first results were obtained. Here we continue with this analysis, and considerably extend the results of [12]. First, we analyse the very notion of minimal limit proportion (density) of a letter by comparing different possible definitions. In particular, we prove that two natural definitions, through finite and infinite words, lead actually to the same quantity. This confirms the significance of this notion and the interest of studying it. We then analyse the minimal proportion $\rho(n)$ of one letter in n-th power-free binary words. In [12] it has been shown that $\rho(n) = \frac{1}{n} + \mathcal{O}(\frac{1}{n^2})$. Here we obtain a much more precise estimate by computing the first four terms of the asymptotic expansion of $\rho(n)$. Specifically, we show that $\rho(n) = \frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \mathcal{O}(\frac{1}{n^5})$, and we also provide bounds for the residual term $\mathcal{O}(\frac{1}{n^5})$. Then we turn to the analysis of the generalized minimal density $\rho(x)$, defined for all real $x > 2$. This generalization, based on the notion of period of a word, was introduced in [12]. It was shown, in particular, that $\rho(x)$, considered as a real function, is discontinuous, as it admits a jump at $x = \frac{7}{3}$. Here we prove much more, namely that $\rho(x)$ has actually an infinity of discontinuity points, as those are all integer points $n \geq 3$. Futhermore, we give an estimate for $\rho(n+0)$ – the right limit of $\rho(x)$ at integer points $n \geq 3$ – and prove that $\rho(n+0) = \frac{1}{n} - \frac{1}{n^2} + \frac{2}{n^3} - \frac{2}{n^4} + \mathcal{O}(\frac{1}{n^5})$.

As usual, A^* denotes the free monoid over an alphabet A. $u \in A^*$ is a *subword* of $w \in A^*$ if w can be written as $u_1 u u_2$ for some $u_1 u_2 \in A^*$. $|u|$ stands for the length of $u \in A^*$. A^ω stands for the set of *one-way infinite* words, often called w-words, over A, that are defined as mappings $\mathbb{N} \to A$. For $n \in \mathbb{N}$, the word w obtained by concatenating n copies of a word v is called the *n-th power* of v and denoted v^n. A word v is a period of w iff w is a subword of v^n for some $n \in \mathbb{N}$.

2 Minimal density: general definition and properties

In this section we analyse, in a general context, the notion of minimal limit density of a letter in words of an infinite set.

Assume we have an infinite set $F \subseteq A^*$ which is *closed under subwords*, that is if a word w is in F, then any subword of w belongs to F too. As noted in Introduction, the property of being closed under subwords characterizes the class of languages that can be specified by a set of prohibited subwords. As F is infinite and closed under subwords, there exist an infinite word from A^ω such that its every finite subword belongs to F. With interpretation of subword avoidance, this allows to speak about infinite words avoiding the set of subwords. We denote by F^ω the set of all infinite words of A^ω with every finite subword belonging to F.

Let $a \in A$ be a distinguished letter, and we are interested in the minimal limit proportion of a's in words of F of unbounded length. For $w \in F$, define $c_a(w)$ to be the number of occurrences of a in w and $\rho_a(w) = \frac{c(w)}{|w|}$. Denote $F(l) = \{w \in F \mid |w| = l\}$.

Definition 1 *For every $l \in \mathbb{N}$, let $\rho_a(F,l) = \frac{1}{l} \min_{w \in F(l)} c_a(w)$ and $\rho_a(F) = \underline{\lim}_{l \to \infty} \rho_a(F,l)$. $\rho_a(F)$ is called the* minimal (limit) density *of a in F.*

Note that the type of argument of ρ_a will always make it clear if the density of an individual word, or the minimal density is meant.

Obviously, all numbers $\rho_a(F,l)$ belong to $[0,1]$ and therefore $\rho_a(F)$ belongs to $[0,1]$ too. The following two Lemmas clarify the behaviour of the sequence $\{\rho_a(F,l)\}_{l=1}^{\infty}$ with respect to $\rho_a(F)$. They are direct generalizations of Propositions 1,2 from [12] and are given without proof.

Lemma 1 *For every $l \in \mathbb{N}$, $\rho_a(F,l) \leq \rho_a(F)$.*

Lemma 2 $\rho_a(F) = \lim_{l \to \infty} \rho_a(F,l) = \sup_{l \geq 1} \rho_a(F,l)$.

By Lemma 2, the lower limit in Definition 1 can be replaced by the simple limit. Thus, the definition $\rho_a(F) = \lim_{l \to \infty} \min_{w \in F(l)} \frac{c_a(w)}{l}$ is correct and seems to capture in a right way the notion of the minimal density. However, there is another natural way to define the minimal limit density directly in terms of infinite words F^ω, and not as the limit density value for finite words. One may ask if this approach always leads to the same density value or may lead to a different one.

For a word $w \in F \cup F^\omega$, let $w[1:j]$ denotes the prefix of w of length j. The density of letter a in an infinite word $v \in F^\omega$ is naturally defined as the limit $\lim_{j \to \infty} \rho_a(v[1:j])$. Obviously, this limit may not exist. However, below we show that among all words for which this limit exists, there is one that realizes the minimum of these limits, which is equal to $\rho_a(F)$. This confirms that $\rho_a(F)$ is the right quantity caracterizing the limit density.

We define an auxiliary measure $\sigma_a(F,l) = \min_{w \in F(l)} \max_{1 \leq j \leq l} \rho_a(w[1:j])$. The following lemma gives a key argument.

Lemma 3 *For every $l \in \mathbb{N}$, $\rho_a(F,l) \leq \sigma_a(F,l) \leq \rho_a(F)$.*

Proof: It is easily seen that $\sigma_a(F,l) \geq \rho_a(F,l)$. Let us prove that $\sigma_a(F,l) \leq \rho_a(F)$ for all $l \in \mathbb{N}$. Assume that $\sigma_a(F,L) > \rho_a(F)$ for some $L \in \mathbb{N}$. This means that every word $v \in F$ of length at least L has a pefix $v[1:j]$ with $\rho_a(v[1:j]) > \rho_a(F)$. Let $\varepsilon = \min\{\rho_a(v[1:j]) - \rho_a(F)\}$ where minimum is taken over all such prefixes. Take any word $w \in F(N)$ with $N > \frac{2L}{\varepsilon}(\rho_a(F) + \varepsilon)$. Find a decomposition $w = w_1 w_2 \ldots w_m$ such that $|w_j| \leq L$ and $\rho_a(w_j) > \rho_a(F)$ for every j, $1 \leq j \leq m-1$, and $|w_m| < L$. Then $c_a(w) \geq (\rho_a(F) + \varepsilon)(|w| - L)$ and $\rho_a(w) \geq \rho_a(F) + \varepsilon - L(\frac{\rho_a(F)+\varepsilon}{|w|}) \geq \rho_a(F) + \frac{\varepsilon}{2}$. Since w was chosen arbitrarily, this contradicts to $\rho_a(F,N) \leq \rho_a(F)$ (Lemma 1). $\square$

Corollary 1 *The limit* $\lim_{l\to\infty} \sigma_a(n,l)$ *exists and is equal to* $\rho_a(F)$.

Lemma 4 *There exists a word* $v \in F^\omega$ *such that* $\lim_{j\to\infty} \rho_a(v[1:j])$ *exists and is equal to* $\rho_a(F)$.

Proof: From Lemma 3 it follows that for every $l \in \mathbb{N}$, there exists a word $w \in F(l)$ with $\rho(w) = \sigma_a(n,l) \leq \rho_a(F)$, that is $\max_{1\leq j\leq |w|} \rho_a(w[1:j]) \leq \rho_a(F)$. Moreover, every prefix of w verifies the same inequality. Therefore, the set of words w verifying the inequality forms an infinite tree with respect to the prefix relation such that the parent of a word w in the tree is its immediate prefix, obtained by removing the rightmost letter. Since the alphabet A is finite, the tree is finitely branching. By König's Lemma, there exists an infinite path in this tree which defines the infinite word v with $\rho_a(v[1:j]) \leq \rho_a(F)$ for all $j \in \mathbb{N}$. Since $\rho_a(F,j) \leq \rho(v[1:j]) \leq \rho_a(F)$, the result follows from Lemma 2. □

Lemma 5 $\min_{v\in F^\omega} \lim_{j\to\infty} \rho_a(v[1:j]) = \rho_a(F)$, *where minimum is taken for over* $v \in F^\omega$ *for which the limit exists.*

Proof: By Lemma 4, there exists a word $v \in F^\omega$ such that $\lim_{j\to\infty} \rho_a(v[1:j]) = \rho_a(F)$. Therefore, $\inf_{v\in F^\omega} \lim_{j\to\infty} \rho_a(v[1:j]) \leq \rho_a(F)$. On the other hand, since $v[1:j] \in F(j)$, then $\rho_a(v[1:j]) \geq \rho_a(F,j)$, then $\lim_{j\to\infty} \rho_a(v[1:j]) \geq \lim_{j\to\infty} \rho_a(F,j) = \rho_a(F)$ and $\inf_{v\in F^\omega} \lim_{j\to\infty} \rho_a(v[1:j]) \geq \rho_a(F)$. The lemma follows. □

Lemmas 4 and 5 imply that there exists a word $v \in F^\omega$ that realizes the minimal limit $\lim_{j\to\infty} \rho_a(v[1:j])$ among all words of F^ω for which the limit exists. Moreover, this minimum is equal $\rho_a(F)$. To avoid the problem of existence of the limit, we could replace it by the lower limit and define the quanity $\inf_{v\in F^\omega} \underline{\lim}_{j\to\infty} \rho_a(v[1:j])$ where the infimum is taken over *all* words $v \in F^\omega$. The proof of Lemma 5 shows that this value is also equal to $\rho_a(F)$, and the infimum is reached on some word $v \in F^\omega$.

Finally, note that one might suggest yet another, though less natural definition of minimal letter density as the value $\underline{\lim}_{j\to\infty} \min_{v\in F^\omega} \rho_a(v[1:j])$. Using Lemma 4 and arguments similar to the proof of Lemma 5, it is easily shown that the lower limit here can be replaced by the simple limit which is again equal to $\rho_a(F)$.

The equvalence of different definitions gives a strong evidence that $\rho_a(F)$ is an interesting quantity to study. In this paper, we undertake this study for a particular family of sets F – the sets of n-th power-free binary words.

3 Minimal letter density in n-th power-free binary words

Consider an alphabet A. For a natural $n \geq 2$, a word $w \in A^*$ is called *n-th power-free* iff it does not contain a subword which is the n-th power of some non-empty word. We denote $PF(n) \subseteq A^*$ the set of n-th power-free finite words. Words from $PF(2)$ are called *square-free*, and words from $PF(3)$ are called *cube-free*. If $w \in A^*$ does not contain a subword uua, where u is a non-empty word and a is the first letter of u, then w is called *strongly cube-free*. An equivalent property (see [14]) is overlap-freeness – w is *overlap-free* if it does not contain two overlapping occurrences of a non-empty word u. Well known Thue's results [15, 16] state that there exist square-free words of unbounded length on the 3-letter alphabet, and strongly cube-free words of unbounded length on the 2-letter alphabet. An infinite sequence of strongly cube-free words can be constructed by iterating the morphism $h(a) = ab$, $h(b) = ba$, known as Thue-Morse morphism. Note that the

existence of infinite strongly cube-free words on the 2-letter alphabet implies that for that alphabet the set $PF(n)$ is infinite for every $n \geq 3$.

From now on we fix on the binary alphabet $A = \{0, 1\}$. Our goal is to compute, for all $n > 2$, the value $\rho_1(PF(n))$ – minimal density of 1 in the words $PF(n)$. Note that by symmetry, $\rho_1(PF(n)) = \rho_0(PF(n))$, and to simplify the notation, we denote $\rho_1(PF(n))$ (respectively $\rho_1(PF(n), l)$) by $\rho(n)$ (respectively $\rho(n, l)$) in the sequel. Similarly, we will drop the index in $c_1(w)$ and $\rho_1(w)$, and will write $c(w)$ and $\rho(w)$ instead.

In [12] it has been proved that $\rho(n) = \frac{1}{n} + \mathcal{O}(\frac{1}{n^2})$. Here, using a different method, we prove the following more precise estimation, that corresponds to the first four terms in the asymptotic expansion of $\rho(n)$.

Theorem 1 $\rho(n) = \frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \mathcal{O}(\frac{1}{n^5})$.

We first establish the upper bound

$$\rho(n) \leq \frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \frac{C}{n^5}, \tag{1}$$

for all $n \geq 3$ and some positive constant C. The proof is based on the following lemma.

Denote by α_i the word 0^i1.

Lemma 6 *Let $k \geq 3$. For i, j, $0 \leq i, j \leq k$ and $i \neq j$, consider a morphism $h : \{0,1\}^* \to \{0,1\}^*$ defined by $h(0) = \alpha_i$, $h(1) = \alpha_j$. For a word $w \in \{0,1\}^*$, if $w \in PF(k)$ then $h(w) \in PF(k+1)$.*

Proof: First observe that $\{h(0), h(1)\}$ is a prefix code, i.e. the inverse image w of any word $h(w)$ is unique. Furthermore, for any $u \in \{0,1\}^*$, the occurrences of 1 in $h(u)$ delimit the images of individual letters of w. This means that any subword of $h(w)$ which ends with 1 and is preceeded by 1 (or starts at the beginning of $h(w)$) is the image of some subword of w.

To prove the lemma, assume by contradiction that for some $w \in PF(k)$, $h(w)$ contains a subword v^{k+1}. Proceed by case analysis on the number of 1's in v. If v contains no 1's, then v^{k+1} contains at least $k+1$ consecutive 0's which is impossible as $h(w)$ is a concatenation of words α_i, α_j. If v contains one 1, then $v = 0^l10^m$, and $v^{k+1} = 0^l1(0^{l+m}1)^k0^m$. Since $h(w) \in \{\alpha_i, \alpha_j\}^*$, we conclude that $l + m \in \{i, j\}$ and w must contain k consecutive occurrences of the letter $h^{-1}(0^{l+m}1)$. Finally, if v contains s 1's, then $v = 0^l1\alpha_{i_1} \ldots \alpha_{i_{s-1}}0^m$, and $v^{k+1} = 0^l1(\alpha_{i_1} \ldots \alpha_{i_{s-1}}0^{l+m}1)^k0^m$. Again, $l + m \in \{i, j\}$ and w contains the k-th power of the inverse image $h^{-1}(\alpha_{i_1} \ldots \alpha_{i_{s-1}}0^{l+m}1)$. □

Lemma 7 *For every $n \geq 4$,*

$$\rho(n) \leq \frac{1}{n - \rho(n-1)} \tag{2}$$

Proof: For $l \in \mathbb{N}$, take a word $w \in PF(n-1)$ with $|w| = l$ and $\rho(w) = \rho(n-1, l)$. Denote by h the morphism defined by $h(0) = \alpha_{n-1}$, $h(1) = \alpha_{n-2}$. Let $u = h(w)$. By Lemma 6, $u \in PF(n)$. Since $c(u) = |w|$, and $|u| = (n-1)c(w) + n(|w| - c(w)) = n|w| - c(w)$, we have $\rho(n, |u|) \leq \rho(u) = \frac{c(u)}{|u|} = \frac{1}{n-\rho(w)} = \frac{1}{n-\rho(n-1,l)}$. Taking the limit for $l \to \infty$, and then $|u| \to \infty$, we have $\rho(n) \leq \frac{1}{n-\rho(n-1)}$. □

Upper bound (1) is now proved by simple induction on $n \geq 3$. Using the trivial inequality $\rho(3) \leq 1/2$, the base case $n = 3$ can be satisfied by choosing any constant $C \geq 57/2$. To prove the inductive step, we apply Lemma 7. This leads to the inequality

$$\frac{1}{n - \left(\frac{1}{n-1} + \frac{1}{(n-1)^3} + \frac{1}{(n-1)^4} + \frac{C}{(n-1)^5}\right)} \leq \frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \frac{C}{n^5}$$

for $n \geq 4$, which reduces to the polynomial inequality

$$(-3+C)n^6 + (-5C+8)n^5 + (8C-9)n^4 + (2-6C)n^3 + (-3C+3)n^2 + (3C-1)n - (C^2+C) \geq 0$$

After substituting $C = 30$, the routine check shows that the inequality holds for all $n \geq 4$ (just substitute $n-4$ for n, expand and notice that all coefficients get positive). This proves that upper bound (1) holds for $C = 30$.

Note that constant C can be reduced if we take into consideration the next term in the asymptotic expansion. Using a similar argument, it can be shown that

$$\rho(n) \leq \frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \frac{3}{n^5} + \frac{C_1}{n^6}$$

for $C_1 = 90$.

Now we turn to bounding $\rho(n)$ from below, and prove the following lower bound.

$$\rho(n) \geq \frac{n-1}{n^2 - n - 1} \tag{3}$$

for all $n \geq 3$.

Consider an arbitrary finite n-th power-free word w. First, group its letters into blocks $\alpha_i = 0^i 1$, $0 \leq i \leq n-1$. For a technical reason we assume that w does not start with α_{n-1}. If it does, we temporarily remove the first symbol 0. w is uniquely decomposed into a concatenation of α_i's and a suffix of at most $n-1$ 0's. Then, we group occurrences of α_i's into larger blocks $\beta(m,k) = (\alpha_{n-1})^m \alpha_k$, $0 \leq m \leq n-1$, $0 \leq k \leq n-2$. Informally, blocks β are delimited by occurrences of α_i with $i \leq n-2$. Again, w is uniquely decomposed into blocks β and the remaining suffix Q of length at most $n^2 - 1$ ($n-1$ occurrences of α_{n-1} followed by $n-1$ 0's). We proceed by grouping blocks β into yet more large blocks. Let

$$\gamma(l, k_0, k_1, \ldots, k_s) = \beta(l, k_0)\beta(n-1, k_1)\ldots\beta(n-1, k_s) = (\alpha_{n-1})^l \alpha_{k_0} (\alpha_{n-1})^{n-1} \alpha_{k_1} \ldots (\alpha_{n-1})^{n-1} \alpha_{k_s},$$

where $0 \leq l \leq n-2$, $s \geq 0$, $0 \leq k_0, k_1, \ldots, k_s \leq n-2$. Blocks γ are delimited by each occurrence of $\beta(l,k)$ with $l \leq n-2$. Note that since w starts with α_k, $k \leq n-2$, it starts with $\beta(0,k)$ and therefore the first block γ starts at the beginning of w. Thus, the decomposition of w is uniquely defined with a possibly remaining suffix Q of length up to $n^2 - 1$. Taking into account the first possibly removed 0, we have $w = Pw'Q$, where $|P| \leq 1$, $|Q| \leq n^2 - 1$, and w' is uniquely decomposed into blocks γ.

Let us now compute the minimal possible ratio of 1's in blocks γ. Consider a block $\gamma(l, k_0, k_1, \ldots, k_s)$. We distingish two cases:

Case $s \geq 1$**:** We show that $k_j + k_{j+1} \leq n-2$ for every j, $0 \leq j \leq s-1$. Indeed, consider the subword $\alpha_{k_j}(\alpha_{n-1})^{n-1}\alpha_{k_{j+1}}$ of $\gamma(l, k_0, k_1, \ldots, k_s)$. If $k_j + k_{j+1} \geq n-1$ then it has the prefix $(0^{k_j}10^{n-1-k_j})^n$ which contradicts to the n-th power-freeness of w.

Using this observation, we can bound

$$\sum_{j=0}^{s} |\alpha_{k_j}| \leq \begin{cases} \frac{s+1}{2}n & s \text{ impair}, \\ \frac{s}{2}n + (n-1) & s \text{ pair}. \end{cases}$$

Then $|\gamma(l, k_0, k_1, \ldots, k_s)| \leq \frac{s}{2}n + (n-1) + sn(n-1) + ln = s(n^2 - \frac{n}{2}) + nl + n - 1$. Since the number of 1's in $\gamma(l, k_0, k_1, \ldots, k_s)$ is $ns + l + 1$, we have $\rho(\gamma(l, k_0, k_1, \ldots, k_s)) \geq \frac{ns+l+1}{s(n^2-\frac{n}{2})+nl+n-1}$. The right-hand side minimizes when l is maximal ($l = n-2$) and s is minimal ($s = 1$). We then obtain $\rho(\gamma(l, k_0, k_1, \ldots, k_s)) \geq \frac{2n-1}{2n^2-\frac{3n}{2}-1}$.

Case $s = 0$: In this case $\gamma(l, k_0) = \beta(l, k_0)$, $|\gamma(l, k_0)| = ln + k_0 + 1$, and $\rho(\gamma(l, k_0)) = \frac{l+1}{ln+k_0+1}$. The right-hand mininizes when both l and k_0 are maximal ($l = k_0 = n-2$), which gives $\rho(\gamma(l, k_0)) \geq \frac{n-1}{n^2-n-1}$.

The second case gives a smaller bound for all $n \geq 3$ and we conclude that $\rho(\gamma(l, k_0, \ldots, k_s)) \geq \frac{n-1}{n^2-n-1}$. Since w' is a concatenation of blocks γ, this implies $\rho(w') \geq \frac{n-1}{n^2-n-1}$. Returning to w, we have $c(w) \geq c(w') \geq \frac{n-1}{n^2-n-1}|w'| \geq \frac{n-1}{n^2-n-1}(|w| - n^2)$, and then $\rho(w) = \frac{c(w)}{|w|} \geq \frac{n-1}{n^2-n-1}(1 - \frac{n^2}{|w|})$. As w is an arbitrary n-th power-free word, we have $\rho(n, l) \geq \frac{n-1}{n^2-n-1}(1 - \frac{n^2}{l})$ for all l. Taking the limit for l going to infinity, we obtain $\rho(n) \geq \frac{n-1}{n^2-n-1}$. This implies in particular that

$$\rho(n) \geq \frac{1}{n} + \frac{1}{n^3} + \frac{1}{n^4} + \frac{2}{n^5} \tag{4}$$

Lower bound (4) together with upper bound (1) implies Theorem 1.

4 Generalized minimal density function

Following [12], we consider in this Section a natural generalization of function $\rho(n)$ to real arguments. Recall that the exponent of a word w is the ratio $\frac{|w|}{\min |v|}$, where the minimum is taken over all periods v of w. The exponent is a useful notion often used in word combinatorics (see [10, 5, 8]), that generalizes the notion of n-th power. For example, Dejean proved that on the 3-letter alphabet, there exist infinite words that don't contain any subword of exponent more than $\frac{7}{4}$. This strengthens Thue's result on the existence of square-free words (i.e. words without subwords of exponent 2) over the 3-letter alphabet.

Using periods, function $\rho(n)$ can be defined on real numbers in the following way. For a real number x, define $PF(x)$ (resp. $PF(x+\varepsilon)$) to be the set of binary words that do not contain a subword of exponent greater than or equal to (resp. strictly greater than) x.

Note that $PF(2+\varepsilon)$ is precisely the class of strongly cube-free words. For the binary alphabet, the existence of infinite cube-free words implies that $PF(x)$ (resp. $PF(x+\varepsilon)$) is infinite for $x > 2$ (resp. for $x \geq 2$). Using the results of Section 2, values $\rho_1(PF(x))$ and $\rho_1(PF(x+\varepsilon))$ are well-defined for $x > 2$ and $x \geq 2$ respectively. Similar to the previous section, we denote them respectively by $\rho(x)$ and $\rho(x+\varepsilon)$. Notation $\rho(x, l)$ and $\rho(x+\varepsilon, l)$ is defined accordingly. Observe that for natural values of $x > 2$, $\rho(x)$ coincides with $\rho(n)$ studied in the previous section.

4.1 Discontinuity of $\rho(x)$

Note that functions $\rho(x), \rho(x+\varepsilon)$ are non-increasing with values from $[0, \frac{1}{2}]$. This implies the existence, for every $x > 2$, of the right limit $\rho(x+0)$, that verifies $\rho(x+0) = \sup_{y>x} \rho(y)$. The following lemma is from [12].

Lemma 8 *For every $x > 2$, $\rho(x+0) = \rho(x+\varepsilon)$.*

In [12], it has been shown that $\rho(x) = \frac{1}{2}$ for $x \in (2, \frac{7}{3}]$, and then proved that the right limit of $\rho(x)$ at $x = \frac{7}{3}$ is strictly smaller than $\frac{1}{2}$, implying that $\rho(x)$ has a jump to the right of $x = \frac{7}{3}$. Here we complement this result by proving that $\rho(x)$ has an infinite number of discontinuity points We show that, besides $x = 7/3$, the function $\rho(x)$ is discontinuous to the right at all integer points $x \geq 3$. We use the following lemma which is somewhat similar to Lemma 6. Recall that $\alpha_i = 0^i 1$.

Lemma 9 *Let $A = \{a_1, \ldots, a_k\}$ and $n \geq 3$. Let $h : A \to \{0,1\}$ be a morphism such that $h(a_i) = \alpha_{m_i}$, where $m_i \leq n$ for all $1 \leq i \leq k$, and $m_i \neq m_j$ for all $i \neq j$. Then for every $(n-1)$-th power-free word $w \in A^*$, $h(w)$ is $(n+\varepsilon)$-th power-free.*

Proof: Similar to Lemma 6, morphism h is injective, and every subword of $h(w)$ ending with 1 and preceeding by 1 is the image of a subword of w.

Assume that $h(w)$ is not $(n+\varepsilon)$-th power-free. Then it contains a subword $u^n a$ for a non-empty word $u \in \{0,1\}^*$ and a the first letter of u. If u contains no 1's, then $u^n a$ contains at least $n+1$ consecutive 0's, which is impossible as $h(w)$ is a concatenation of α_{m_i}'s, and $m_i \leq n$. Assume that u contains at least one 1, that is $u = 0^p 1 u'$, $p \geq 0$, $u' \in \{0,1\}^*$. Then $u^n = (0^p 1 u')^n = 0^p 1 v^{n-1} u'$ for $v = u' 0^p 1$. By properties of morphism h, each occurrence of v is the image of some subword of w under morphism h. Since this subword is the same for all occurrences of v, then w contains a subword $(h^{-1}(v))^{n-1}$ which contradicts to n-th power-freeness of w. □

Lemma 10 *For every $n \geq 4$,*

$$\rho(n+\varepsilon) \leq \frac{1}{n+1-\rho(n-1)} \tag{5}$$

Proof: Denote $h_n : \{0,1\}^* \to \{0,1\}^*$ the morphism defined by $h_n(0) = \alpha_n$, $h_n(1) = \alpha_{n-1}$. Let w_l be an $(n-1)$-th power-free word of length l with minimal number of 1's ($\rho(w_l) = \rho(n-1, l)$). Clearly, $|h_n(w_l)| = (n+1)(l - c(w_l)) + n c(w_l) = (n+1)l - c(w_l)$, and $c(h_n(w_l)) = l$. By Lemma 9, $h_n(w_l)$ is $(n+\varepsilon)$-th power-free, and we have

$$\rho(n+\varepsilon, |h_n(w_l)|) \leq \rho(h_n(w_l)) = \frac{l}{(n+1)l - c(w_l)} = \frac{1}{n+1-\rho(n-1,l)}$$

By taking the limit for $l \to \infty$ (see Lemma 2), inequality (5) follows. □

Inequality (5) together with the trivial inequality $\rho(n-1) \leq 1/2$ gives $\rho(n+\varepsilon) \leq \frac{1}{n-1/2} < \frac{1}{n}$ for $n \geq 4$. On the other hand, from lower bound (3) it follows that $\rho(n) \geq \frac{n-1}{n^2-n-1} > \frac{1}{n}$. This implies that $\rho(n+0) = \rho(n+\varepsilon) < \rho(n)$, that is $\rho(x)$ has a jump to the right of all integer points $n \geq 4$.

For $n = 3$, inequality (5) does not make sense ($\rho(2)$ is not defined). Therefore, the case $n = 3$ should be analysed separately.

Lemma 11 $\rho(3+\varepsilon) \leq \frac{1}{3}$.

Proof: Take a 3-letter alphabet $A = \{1,2,3\}$. For $w \in A^*$, let $c_i(w)$ $(i = 1,2,3)$ denote the number of occurrences of i in w. For any $l \in \mathbb{N}$, choose a square-free word $w_l \in A^*$ of length l such that $c_1(w) \leq c_2(w) \leq c_3(w)$. Note that for all $l \in \mathbb{N}$, w_l is well-defined, which follows from the existence of infinite square-free words on the 3-letter alphabet. Consider the morphism $h : A^* \to \{0,1\}^*$

defined by $h(1) = 01$, $h(2) = 001$, $h(3) = 0001$. Then $|h(w_l)| = 2c_1(w_l) + 3c_2(w_l) + 4c_3(w_l) = 3l + (c_3(w_l) - c_1(w_l)) \geq 3l$, and $\rho(h(w_l)) \leq \frac{l}{3l} = \frac{1}{3}$. By Lemma 9, word w_l is $(3+\varepsilon)$-th power-free, and then $\rho(3+\varepsilon, |h(w_l)|) \leq \frac{1}{3}$. Taking the limit for $l \to \infty$ and using Lemma 2, we get $\rho(3+\varepsilon) \leq \frac{1}{3}$. $\square$

On the other hand, from lower bound (3) it follows that $\rho(3) \geq \frac{2}{5}$. Therefore, $\rho(x)$ has a jump to the right of $x = 3$.

Putting all together, we obtain

Theorem 2 *$\rho(x)$ is discontinuous to the right of $x = \frac{7}{3}$ as well as to the right of all natural points $n \geq 3$.*

4.2 Estimating $\rho(n + \varepsilon)$

In Section 3 we obtained an estimate of $\rho(n)$, for natural $n \geq 3$ (Theorem 1). Theorem 2 says that $\rho(x)$, considered as function on real argument, has a jump on the right of all these points. In this final part of the paper, we estimate the size of these jumps by estimating the values $\rho(n + \varepsilon)$ for natural $n \geq 3$. Recall that $\rho(n + \varepsilon) = \rho(n + 0)$ by Lemma 8.

We start with proving the lower bound

$$\rho(n + \varepsilon) \geq \frac{n - 1}{n^2 - 2} \tag{6}$$

for all $n \geq 3$. The proof follows closely the proof of lower bound (3) from Section 3. Therefore, we only give a sketch of it, underlining the differences with the proof of Section 3.

Consider a finite $(n+\varepsilon)$-th power-free word w. As in Section 3, we group its letters into blocks $\alpha_i = 0^i 1$, where $0 \leq i \leq n$ (w may contain n-th powers). Again, we assume that w does not start with α_n, otherwise we remove the first 0 into a separate prefix. We now note that under this assumption, w cannot contain a subword $(\alpha_n)^n$. Indeed, since w does not start with α_n, the occurrence of $(\alpha_n)^n$ is preceeded by at least one letter. This latter cannot be 0, as w would then have a subword 0^{n+1} which contradicts to the fact that w does not contain subwords of exponent greater than n. This letter cannot be 1 either, as this would give the subword $(10^n)^n 1$ which again contradicts to the fact that w is $(n + \varepsilon)$-th power-free. Thus, no occurrence $(\alpha_n)^n$ exists.

We then group α_i's into blocks $\beta(m, k) = (\alpha_n)^m \alpha_k$, $0 \leq m \leq n - 1$, $0 \leq k \leq n - 1$, and then further into blocks

$$\gamma(l, k_0, k_1, \ldots, k_s) = \beta(l, k_0)\beta(n - 1, k_1) \ldots \beta(n - 1, k_s) = (\alpha_n)^l \alpha_{k_0} (\alpha_n)^{n-1} \alpha_{k_1} \ldots (\alpha_n)^{n-1} \alpha_{k_s},$$

where $0 \leq l \leq n - 2$, $s \geq 0$, $0 \leq k_0, k_1, \ldots, k_s \leq n - 2$.

We now compute the minimal value of $\rho(\gamma(l, k_0, k_1, \ldots, k_s))$. Consider a block $\gamma(l, k_0, k_1, \ldots, k_s)$. As in Section 3, we distingish two cases:

Case $s \geq 1$: Here we show that $k_j + k_{j+1} \leq n$ for every j, $0 \leq j \leq s-1$. By contradiction, assume that $k_j + k_{j+1} > n$. If $k_j = 0$ then $k_{j+1} > n$ which is a contradiction. Assume $k_j > 0$, and consider the subword $\alpha_{k_j} (\alpha_{n-1})^{n-1} \alpha_{k_{j+1}}$. If $k_j + k_{j+1} > n$, then it has the prefix $(0^{k_j} 10^{n-k_j})^n$ followed by at least one 0. This gives a subword of exponent greater than n which is a contradiction.

Now we can bound $\sum_{j=0}^{s} |\alpha_{k_j}| \leq (\frac{s}{2}+1)n+s$, and then $|\gamma(l, k_0, k_1, \ldots, k_s)| \leq l(n+1)+s(n^2+\frac{n}{2})+n$. Then $\rho(\gamma(l, k_0, k_1, \ldots, k_s)) \geq \frac{l+ns+1}{l(n+1)+s(n^2+\frac{n}{2})+n}$. Again, the right-hand side minimizes when l is maximal ($l = n-2$) and s is minimal ($s = 1$). Finally for this case, $\rho(\gamma(l, k_0, k_1, \ldots, k_s)) \geq \frac{2n-1}{2n^2+\frac{n}{2}-2}$.

Case $s = 0$: In this case $\gamma(l, k_0) = \beta(l, k_0)$, $|\gamma(l, k_0)| = l(n+1) + k_0 + 1$, and $\rho(\gamma(l, k_0)) = \frac{l+1}{l(n+1)+k_0+1}$. The right-hand mininizes when both l and k_0 are maximal ($l = n-2$, $k_0 = n-1$), which gives $\rho(\gamma(l, k_0)) \geq \frac{n-1}{n^2-2}$.

The second case gives a smaller or equal bound for all $n \geq 3$ and we conclude that $\rho(\gamma(l, k_0, \ldots, k_s)) \geq \frac{n-1}{n^2-2}$. Since w is a concatenation of blocks γ (with possibly remaining prefix and suffix of bounded length), this implies equation (6).

Turning to asymptotic expansion of (6), we have

$$\rho(n+\varepsilon) \geq \frac{1}{n} - \frac{1}{n^2} + \frac{2}{n^3} - \frac{2}{n^4} + \mathcal{O}(\frac{1}{n^5}). \tag{7}$$

To obtain the lower bound of $\rho(n+\varepsilon)$ that matches the upper bound (7), it suffices to substitute into inequality (5) the upper bound of $\rho(n-1)$ implied by (1) (instead of trivial upper bound $\rho(n-1) \leq \frac{1}{2}$).

We then get

$$\rho(n+\varepsilon) \leq \frac{1}{n+1-\left(\frac{1}{n-1} + \frac{1}{(n-1)^3} + \frac{1}{(n-1)^4} + \frac{C}{(n-1)^5}\right)} = \frac{1}{n} - \frac{1}{n^2} + \frac{2}{n^3} - \frac{2}{n^4} + \mathcal{O}(\frac{1}{n^5})$$

Together with (7), this gives

Theorem 3 $\rho(n+\varepsilon) = \frac{1}{n} - \frac{1}{n^2} + \frac{2}{n^3} - \frac{2}{n^4} + \mathcal{O}(\frac{1}{n^5})$.

5 Concluding remarks

In this paper we have continued with the study of minimal density function $\rho(x)$, introduced in [12]. We analysed a general definition of this notion for any infinite set of words that don't contain subwords of a specified set. We have proved that different viewpoints lead to equivalent definitions. Then, for the case of repetition-free binary words, we have extended several results of [12]. Specifically, we have given a more precise estimation for the values of $\rho(n)$ at integer points $n \geq 3$, and we proved that $\rho(x)$, considered as a function on real argument, is discontinuous to the right of all integer values of x. Finally, we gave an estimate of values $\rho(n)$ and $\rho(n+\varepsilon)$.

Many questions about minimal density function $\rho(x)$ remain open. Does it have other discontinuities? What are they? Is this function piece-wise constant? All these questions are still to be answered.

Acknowledgements This work was supported by the French-Russian A.M.Liapunov Institut of Applied Mathematics and Informatics at Moscow University. The first and third authors were also supported by the Russian Foundation of Fundamental Research (grant 96–01–01068). We are grateful to Alexandre Ugolnikov for his help and to Vladimir Grebinski for commenting the manuscript.

References

[1] А.И. Зимин. Блокирующие множества термов. *Математический Сборник*, 119(3):363–375, 1982. English Translation: A.I.Zimin, Blocking sets of terms, Math. USSR Sbornik 47 (1984), 353-364.

[2] K. Baker, G. McNulty, and W. Taylor. Growth problems for avoidable words. *Theoret. Comp. Sci.*, 69:319–345, 1989.

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

[4] J. Berstel. Axel thue's work on repetitions in words. Invited Lecture at the 4th Conference on Formal Power Series and Algebraic Combinatorics, Montreal, 1992, June 1992. accessible at `http://www-litp.ibp.fr:80/berstel/`.

[5] J. Berstel and D. Perrin. *Theory of codes.* Academic Press, 1985.

[6] J. Cassaigne. *Motifs évitables et régularités dans les mots.* Thèse de doctorat, Université Paris VI, 1994.

[7] C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook on Formal Languages*, volume I. Springer, Berlin-Heidelberg-New York, 1996.

[8] M. Crochemore and P. Goralcik. Mutually avoiding ternary words of small exponent. *International Journal of Algebra and Computation*, 1(4):407–410, 1991.

[9] J. Currie. Open problems in pattern avoidance. *American Mathematical Monthly*, 100:790–793, 1993.

[10] F. Dejean. Sur un théorème de Thue. *J. Combinatorial Th. (A)*, 13:90–99, 1972.

[11] M. Dekking. On the Thue-Morse measure. *Acta Univ. Carolin. Math. Phis*, 33(2):35–40, 1992.

[12] R. Kolpakov and G. Kucherov. Minimal letter frequency in n-power-free binary words. In I. Privara and P. Ružička, editors, *Proceedings of the 22nd International Symposium on Mathematical Foundations of Computer Science (MFCS), Bratislava (Slovakia)*, volume 1295 of *Lecture Notes in Computer Science*, pages 347–357. Springer Verlag, 1997.

[13] L. Rosaz. Making the inventory of unavoidable sets of words of fixed cardinality. *Theoretical Computer Science*, 1998. to appear.

[14] A. Salomaa. *Jewels of formal language theory.* Computer Science Press, 1986.

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

[16] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 10:1–67, 1912.