Letter frequency in infinite repetition-free words

Pascal Ochem [a,1]

[a] *LaBRI, Université Bordeaux 1,*
351 cours de la Libération, 33405 Talence Cedex, France

Abstract

We estimate the extremal letter frequency in infinite words over a finite alphabet avoiding some repetitions. For ternary square-free words, we improve the bounds of Tarannikov on the minimal letter frequency, and prove that the maximal letter frequency is $\frac{255}{653}$. Kolpakov et al. have studied the function ρ such that $\rho(x)$ is the minimal letter frequency in an infinite binary x-free word. In particular, they have shown that ρ is discontinuous at $\frac{7}{3}$ and at every integer at least 3. We answer one of their question by providing some other points of discontinuity for ρ. Finally, we propose stronger versions of Dejean's conjecture on repetition threshold in which unequal letter frequencies are required.

Key words: Combinatorics on words, repetitions, letter frequency

1 Introduction

A *square* is a repetition of the form xx, where x is a nonempty word; an example in English is `hotshots`.
Let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k-1\}$. It is easy to see that every word of length ≥ 4 over Σ_2 must contain a square, so squares cannot be avoided in infinite binary words. However, Thue showed [18,19,1] that there exist infinite words over Σ_3 that avoid squares.

An interesting variation is to consider avoiding fractional powers. For $\alpha \geq 1$ a rational number, we say that y is an α-power if we can write $y = x^n x'$ with

Email address: `ochem@labri.fr` (Pascal Ochem).
[1] This research was supported by The European Research Training Network COMBSTRU-HPRN-CT-2002-00278.

x' a prefix of x and $|y| = \alpha|x|$. For example, the French word `entente` is a $\frac{7}{3}$-power and the English word `tormentor` is a $\frac{3}{2}$-power. For real $\alpha > 1$, we say that a word *is α-free* (resp. *is α^+-free*) if it contains no factor that is a α'-power for any rational $\alpha' \geq \alpha$ (resp. $\alpha' > \alpha$).

We study the extremal frequencies of a letter in factorial languages defined by an alphabet size and a set of forbidden repetitions. Given such a language, we denote by $f_{\min}$ (resp. $f_{\max}$) the minimal (resp. maximal) letter frequency in an infinite word that belong to the language L. Letter frequencies have been mainly studied in [10,16,17]. We consider here the frequency of the letter 0. Let $|w|_0$ denote the number of occurrences of 0 in the finite word w. So the letter frequency in w is $\frac{|w|_0}{|w|}$. A negative result is either a lower bound on $f_{\min}$ or an upper bound on $f_{\max}$. Notice that for binary words, we only need to consider $f_{\min}$ since $f_{\min} + f_{\max} = 1$.

Our results are stated in Section 2. The proof technique for negative results is an improved version of the methods given in [11] to find lower bounds on the minimal frequency of occurrences of squares infinite binary words. It is detailed in Section 3. Positive results consist of uniform morphisms that can produce infinite words in L with a given letter frequency. The method used to find such morphisms is explained in Section 4. In Section 5, we make a conjecture related to Dejean's conjecture [7] involving unequal letter frequencies. The C sources of the programs and the morphisms used in this paper are available at: `http://dept-info.labri.fr/~ochem/morphisms/`.

2 Statement of main results

For ternary square-free words, Tarannikov [17] showed that $f_{\min} \in \left[\frac{1780}{6481}, \frac{64}{233}\right] = [0.27464897\cdots, 0.27467811\cdots]$. According to [16], he also proved that $f_{\max} \leq \frac{469}{1201} = 0.39050791\cdots$. We obtain the following results:

Theorem 1 *For ternary square-free words, we have*

(1) $f_{\min} \in \left[\frac{1000}{3641}, \frac{883}{3215}\right] = [0.27464982\cdots, 0.27465007\cdots]$.
(2) $f_{\max} = \frac{255}{653} = 0.39050535\cdots$.

A (β, n)-repetition is a repetition with prefix size n and exponent β. The notions of (β, n)-freeness and (β^+, n)-freeness are introduced in [8]. A word is said to be (β, n)-free (resp. (β^+, n)-free) if it contains no (β', n')-repetition such that $n' \geq n$ and $\beta' \geq \beta$ (resp. $\beta' > \beta$). We construct in [8] an infinite $\left(\frac{8}{5}^+, 3\right)$-free binary word.

Theorem 2 *For $(\frac{5}{3}, 3)$-free binary words, we have $f_{\min} = \frac{1}{2}$.*

Theorem 2 implies that infinite $(\beta, 3)$-free binary words have equal letter frequency for $\beta \in \left[\frac{8}{5}^+, \frac{5}{3}\right]$. A similar result in [10] says that infinite $(\beta, 1)$-free binary words have equal letter frequency for $\beta \in \left[2^+, \frac{7}{3}\right]$. It is noticeable that these two cases of equal letter frequency have different kind of growth function. Karhumäki and Shallit have shown that the growth function of $\frac{7}{3}$-free binary words is polynomial [9], whereas the growth function of $(\frac{8}{5}^+, 3)$-free binary words is exponential. To see this, notice that the 992-uniform morphism $h : \Sigma_4^* \to \Sigma_2^*$ given in [8] produces a $\left(\frac{8}{5}^+, 3\right)$-free binary word $h(w)$ for every $\frac{7}{5}^+$-free word $w \in \Sigma_4^*$, and that an exponential lower bound on the number of 4-ary $\frac{7}{5}^+$-free words is shown in [14].

Let $\rho(x)$ (resp. $\rho(x^+)$) denote the minimal letter frequency in an infinite x-free (resp. (x^+)-free) binary word. By the previous discussion, we thus have $\rho(2^+) = \rho\left(\frac{7}{3}\right) = \frac{1}{2}$. Kolpakov et al. [10] proved that the function ρ is discontinuous at every integer value at least 3 and at $\frac{7}{3}$, more precisely they obtained that $\rho\left(\frac{7}{3}^+\right) \leq \frac{10}{21} = 0.47619047\cdots < \frac{1}{2} = \rho\left(\frac{7}{3}\right)$.
The next result provides 11 new points of discontinuity for ρ in the range $\left[\frac{7}{3}^+, 3\right]$, namely $\frac{17}{7}$, $\frac{5}{2}$, $\frac{131}{52}$, $\frac{43}{17}$, $\frac{23}{9}$, $\frac{41}{16}$, $\frac{18}{7}$, $\frac{631}{245}$, $\frac{8}{3}$, $\frac{26}{9}$, and $\frac{44}{15}$. It also exhibits a new constant segment: $\rho\left(\frac{41}{16}^+\right) = \rho\left(\frac{18}{7}\right) = \frac{79}{179}$.

Theorem 3

(1)	$\rho\left(\frac{7}{3}^+\right)$	$\leq$	$\frac{327}{703}$	$=$	$0.4651493599\cdots$
(2)	$\rho\left(\frac{17}{7}\right)$	$>$	$\frac{427}{918}$	$=$	$0.4651416122\cdots$
(3)	$\rho\left(\frac{17}{7}^+\right)$	$\leq$	$\frac{797}{1722}$	$=$	$0.4628339141\cdots$
(4)	$\rho\left(\frac{5}{2}\right)$	$\geq$	$\frac{54286}{117293}$	$=$	$0.4628238684\cdots$
(5)	$\rho\left(\frac{5}{2}^+\right)$	$\leq$	$\frac{279}{631}$	$=$	$0.4421553090\cdots$
(6)	$\rho\left(\frac{131}{52}\right)$	$>$	$\frac{107}{242}$	$=$	$0.4421487603\cdots$
(7)	$\rho\left(\frac{131}{52}^+\right)$	$\leq$	$\frac{191}{432}$	$=$	$0.4421296296\cdots$
(8)	$\rho\left(\frac{43}{17}\right)$	$>$	$\frac{508}{1149}$	$=$	$0.4421235857\cdots$
(9)	$\rho\left(\frac{43}{17}^+\right)$	$\leq$	$\frac{262}{593}$	$=$	$0.4418212479\cdots$
(10)	$\rho\left(\frac{23}{9}\right)$	$>$	$\frac{1063}{2406}$	$=$	$0.4418121363\cdots$
(11)	$\rho\left(\frac{23}{9}^+\right)$	$\leq$	$\frac{860}{1947}$	$=$	$0.4417051875\cdots$
(12)	$\rho\left(\frac{41}{16}\right)$	$>$	$\frac{519}{1175}$	$=$	$0.4417021277\cdots$

$$
\begin{array}{llcll}
(13) & \rho\left(\frac{41}{16}^{+}\right) & \leq & \frac{79}{179} & = 0.4413407821\cdots \\
(14) & \rho\left(\frac{18}{7}\right) & \geq & \frac{79}{179} & = 0.4413407821\cdots \\
(15) & \rho\left(\frac{18}{7}^{+}\right) & \leq & \frac{272}{617} & = 0.4408427877\cdots \\
(16) & \rho\left(\frac{631}{245}\right) & > & \frac{1196}{2713} & = 0.4408403981\cdots \\
(17) & \rho\left(\frac{631}{245}^{+}\right) & \leq & \frac{190}{431} & = 0.4408352668\cdots \\
(18) & \rho\left(\frac{8}{3}\right) & > & \frac{3431}{7783} & = 0.4408325838\cdots \\
(19) & \rho\left(\frac{8}{3}^{+}\right) & \leq & \frac{76}{187} & = 0.4064171123\cdots \\
(20) & \rho\left(\frac{26}{9}\right) & > & \frac{165}{406} & = 0.4064039409\cdots \\
(21) & \rho\left(\frac{26}{9}^{+}\right) & \leq & \frac{89}{219} & = 0.4063926941\cdots \\
(22) & \rho\left(\frac{44}{15}\right) & > & \frac{675}{1661} & = 0.4063816978\cdots \\
(23) & \rho\left(\frac{44}{15}^{+}\right) & \leq & \frac{332}{817} & = 0.4063647491\cdots \\
(24) & \rho\left(3\right) & > & \frac{115}{283} & = 0.4063604240\cdots \\
(25) & \rho\left(3^{+}\right) & \leq & \frac{523}{1810} & = 0.2889502762\cdots
\end{array}
$$

3 Method for negative results

Let L be a factorial language. A word w is said to be k-biprolongable in L if there exists a word $lwr \in L$ such that $|l| = |r| = k$. A *suffix cover* of L is a set S of finite words in L such that every finite word that is k-biprolongable in L and of length at least $\max_{u\in S} |u|$ has a suffix that belongs to S, for some finite number k. Taking $k = 20$ is sufficient for every negative result in this paper. For a word $u \in S$, let

$$A_u(q) = \left\{ w \in L \mid uw \in L \text{ and for every prefix } w' \text{ of } w,\ \frac{|w'|_0}{|w'|} < q \right\}.$$

Lemma 4 *Let L be a factorial language and S one of its suffix covers. Let $q \in \mathbb{Q}$. If $A_u(q)$ is finite for every word $u \in S$, then $f_{\min} \geq q$.*

PROOF. Assume $A_u(q)$ is finite for every word $u \in S$. We show that every right-infinite word $w \in L$ has a decomposition into finite factors $w = pv_0v_1v_2v_3\cdots$ such that $|p| = k$, $|v_0| = \max_{u\in S} |u|$, and $\frac{|v_i|_0}{|v_i|} \geq q$ for every $i \geq 1$. Notice that for every $i \geq 0$, the factor $f_i = v_0\cdots v_i$ is k-biprolongable in L and is such that $|f_i| \geq \max_{u\in S} |u|$. Thus, for every $i \geq 0$, f_i has a suffix $s_i \in S$, and since $A_{s_i}(q)$ is finite, there exists a finite factor v_{i+1} at the right of f_i such that $\frac{|v_{i+1}|_0}{|v_{i+1}|} \geq q$.

Lemma 4 enables us to obtain bounds of the form $f_{\min} \geq q$ by choosing an explicit suffix cover and checking by computer that every set $A_u(q)$ is finite. It is easy to see that Lemma 4 and the definition of $A_u(q)$ can be modified to provide bounds of the form $f_{\min} > q$, $f_{\max} \leq q$, or $f_{\max} < q$. This method is a natural generalization of the one in [17], where the suffix cover consists of the empty word, and of the one in [11], where the suffix cover consists of all binary words of length three.

Since we study here the frequency of the letter 0 in repetition-free words, every letter other than 0 play the same role. Let us say that two words u and u' in Σ_s are equivalent if and only if u can be obtained from u' by a permutation of the letters in $\Sigma_s \setminus \{0\}$. Notice that for two equivalent words u and u', $A_{u'}(q)$ is finite if and only if $A_u(q)$ is finite. We define the reduced suffix cover of a suffix cover S as the quotient of S by this equivalence relation.

To prove the negative part of Theorem 1.1 we used the reduced suffix cover $\{1, 01210, 0210, 2010\}$, the computation took about 20 days on a XEON 2.2Gh.
For Theorem 1.2 we used the reduced suffix cover $\{0, 01, 021, 0121\}$.
For Theorem 2 we used the suffix cover $\{01, 11, 000, 11100, 0100, 1110, 1010,$ $0001111000010, 0111101000010, 1110101000010, 0111100010\}$.
A computer check shows that this is indeed a suffix cover for 20-biprolongable $(\frac{5}{3}, 3)$-free binary words. The negative statements of Theorem 3 (even items) were obtained using the suffix cover $\{1, 10, 100\}$.

4 Method for positive results

Let L be a factorial language over Σ_s^*. To construct an infinite word $w \in L$ with a given letter frequency $q \in \mathbb{Q}$, we basically use the method described in [14]. We write $q = \frac{a}{b}$ with a coprime to b. For increasing values of k, we look for a $(k \times b)$-uniform morphism $h : \Sigma_e^* \to \Sigma_s^*$ producing (infinite) words in L such that $|h(i)|_0 = k \times a$ for every $i \in \Sigma_e$.

Consider the 8-uniform morphism $m : \Sigma_3^* \longrightarrow \Sigma_4^*$ defined by

$$m(0) = 01232103,$$
$$m(1) = 01230323,$$
$$m(2) = 01210321.$$

To get the bound $f_{\min} \leq \frac{883}{3215}$ in Theorem 1, we found a square-free morphism $h_{\max} : \Sigma_3^* \longrightarrow \Sigma_3^*$ such that $h_{\max} = m_{\max} \circ m$ where $m_{\max} : \Sigma_4^* \longrightarrow \Sigma_3^*$ is a 3215-uniform morphism. To get the bound $f_{\max} \geq \frac{255}{653}$ in Theorem 1, we found a square-free morphism $h_{\min} : \Sigma_3^* \longrightarrow \Sigma_3^*$ such that $h_{\min} = m_{\min} \circ m$ where

$m_{\min} : \Sigma_4^* \longrightarrow \Sigma_3^*$ is a 9142-uniform morphism ($9142 = 14 \times 653$). We need a result of Crochemore [6] saying that a uniform morphism is square-free if the image of every square-free word of length 3 is square-free. The software **mreps** [13] written by Kucherov et al. can test if a word is square-free in linear time. We used it to prove that $h_{\min}$ and $h_{\max}$ are square-free by checking that $h_{\min}(w)$ and $h_{\max}(w)$ are square-free, where $w = 010201210120212$ is square-free and contains every ternary square-free words of length 3 as factors. Checking the image of w is faster than checking the images of the 12 ternary square-free words of length 3 because **mreps** runs in linear time. Since the morphisms $h_{\min}$ (resp. $h_{\max}$) are square-free, we obtain an exponential lower bound for ternary square-free words with letter frequency $\frac{883}{3215}$ (resp. $\frac{255}{653}$).

Let t denote the Thue-Morse word, i.e. the fixed point of $0 \mapsto 01,\ 1 \mapsto 10$. A uniform morphism $h : \Sigma_i^* \to \Sigma_k^*$ is said to be *synchronizing* if for any $a, b, c \in \Sigma_i$ and $s, r \in \Sigma_k$, if $h(ab) = rh(c)s$, then either $r = \varepsilon$ and $a = c$ or $s = \varepsilon$ and $b = c$. For each positive statement in Theorem 3 (odd items), we provide a q-uniform synchronizing morphism $h : \Sigma_2^* \longrightarrow \Sigma_2^*$ such that $h(t)$ has the desired properties of repetition-freeness and letter frequency. Suppose $h(t)$ contains a forbidden repetition of prefix p and exponent $\frac{7}{3} < e \leq 3$. If $|p| < 2q$, then the length of the repetition is less than $6q$, so that checking the h-image of every factor t of length 7 is sufficient. If $|p| \geq 2q$, then p contains a full h-image of some letter, so $|p|$ is a multiple of q by the synchronizing property. Thus we only need to check that $h(0)$ and $h(1)$ do not have too large common prefixes and suffixes, or equivalently check the words h(1001) and h(0110).

5 Dejean's conjecture and letter frequencies

The *repetition threshold* is the least exponent $\alpha = \alpha(k)$ such that there exists an infinite (α^+)-free word over Σ_k. Dejean proved that $\alpha(3) = \frac{7}{4}$. She also conjectured that $\alpha(4) = \frac{7}{5}$ and $\alpha(k) = \frac{k}{k-1}$ for $k \geq 5$. This conjecture is now "almost" solved: Pansiot [15] proved that $\alpha(4) = \frac{7}{5}$ and Moulin-Ollagnier [12] proved that Dejean's conjecture holds for $5 \leq k \leq 11$. Recently, Currie and Mohammad-Noori [5] also proved the cases $12 \leq k \leq 14$, and Carpi [2] settled the cases $k \geq 38$. For more information, see [4]. Based on numerical evidences, we propose the following conjecture which implies Dejean's conjecture.

Conjecture 5

(1) For every $k \geq 5$, there exists an infinite $\left(\frac{k}{k-1}^+\right)$-free word over Σ_k with letter frequency $\frac{1}{k+1}$.

(2) For every $k \geq 6$, there exists an infinite $\left(\frac{k}{k-1}^+\right)$-free word over Σ_k with letter frequency $\frac{1}{k-1}$.

It is easy to see that the values $\frac{1}{k+1}$ and $\frac{1}{k-1}$ in Conjecture 5 would be best possible. For $\left(\frac{5}{4}^+\right)$-free words over Σ_5, we obtain $f_{\max} < \frac{103}{440} = 0.23409090\cdots < \frac{1}{4}$ using the reduced suffix cover $\{0, 01, 012, 0123, 012341, 401234, 4301234\}$. That is why Conjecture 5.2 is stated with $k \geq 6$. Recently, we proved [3] Conjecture 5.1 with $k = 5$ and Conjecture 5.2 with $k = 6$.

References

[1] J. Berstel. *Axel Thue's Papers on Repetitions in Words: a Translation.* Number 20 in Publications du Laboratoire de Combinatoire et d'Informatique Mathématique. Université du Québec à Montréal, February 1995.

[2] A. Carpi. On the repetition threshold for large alphabets, *MFCS 2006.*

[3] J. Chalopin and P. Ochem. Dejean's conjecture and letter frequency, *Mons Days of Theoretical Computer Science 2006.*

[4] C. Choffrut and J. Karhumäki. Combinatorics of words, In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, Vol. 1, pp. 329–438. Springer-Verlag, 1997.

[5] J.D. Currie and M. Mohammad-Noori. Dejean's conjecture and sturmian words, *Eur. J. Combin.* - To appear.

[6] M. Crochemore. Sharp characterizations of squarefree morphisms. *Theoret. Comput. Sci.* **18** (1982), 221–226.

[7] F. Dejean. Sur un théorème de Thue, *J. Combin. Theory. Ser. A* **13** (1972), 90–99.

[8] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of repetition threshold, *Theoret. Comput. Sci.* **345** (2005), 359–369.

[9] J. Karhumäki and J.O. Shallit. Polynomial versus exponential growth in repetition-free binary words. *J. Combin. Theory. Ser. A* **105(2)** (2004), 335–347.

[10] R. Kolpakov, G. Kucherov, and Y. Tarannikov. On repetition-free binary words of minimal density, *Theoret. Comput. Sci.* **218** (1999), 161–175.

[11] G. Kucherov, P. Ochem, and M. Rao. How many square occurrences must a binary sequence contain ? *Electron. J. Comb.* **10(1)** (2003), #R12.

[12] J. Moulin-Ollagnier. Proof of Dejean's conjecture for alphabets with $5, 6, 7, 8, 9, 10$ and 11 letters, *Theoret. Comput. Sci.* **95** (1992), 187–205.

[13] `http://mreps.loria.fr/`

[14] P. Ochem. A generator of morphisms for infinite words, In *Proceedings of the Workshop on Word Avoidability, Complexity, and Morphisms*, Turku, Finland, July 17 2004. LaRIA Technical Report 2004-07, pp. 9–14. *RAIRO - Theoretical Informatics and Applications* - To appear.

[15] J.-J. Pansiot. A propos d'une conjecture de F. Dejean sur les répétitions dans les mots, *Disc. Appl. Math.* **7** (1984), 297–311.

[16] C. Richard and U. Grimm. On the entropy and letter frequencies of ternary square-free words, *Electron. J. Comb.* **11** (2004), #R14

[17] Y. Tarannikov. The minimal density of a letter in an infinite ternary square-free word is 0.2746..., *J. Integer Sequences* 5(2):Article 02.2.2 (2002).

[18] A. Thue. Über unendliche Zeichenreihen, *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

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