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 word, 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 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

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 [6,13,12]. Let Σ_i denote the i-letter alphabet $\{0, 1, \ldots, i-1\}$. 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 $f_{\min} + f_{\max} = 1$. We denote by $\rho(x)$ (resp. $\rho(x^+)$) the minimal letter

Email address: `ochem@labri.fr` (Pascal Ochem).

[1] Research supported in part by COMBSTRU.

frequency in an infinite x-free (resp. (x^+)-free) binary word. Our results are stated in Section 2. The proof technique for negative results is an improved version of the methods given in [13,7] to find lower bounds on the minimal frequency of square occurrences in an infinite binary word. It is detailled in Section 3. Positive results consist in 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 [3] 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 [13] showed that $f_{\min} \in \left[\frac{1780}{6481}, \frac{64}{233}\right] = [0.27464897\ldots, 0.27467811\ldots]$. According to [12], he also proved that $f_{\max} \leq \frac{469}{1201} = 0.39050791\ldots$. We obtain the following results:

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

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

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 [4]. 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 [4] 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 [6] says that infinite $(\beta, 1)$-free binary words have equal letter frequency for $\beta \in \left[2^+, \frac{7}{3}\right]$, i.e. $\rho(2^+) = \rho\left(\frac{7}{3}\right) = \frac{1}{2}$. 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 [5], 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 [4] 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 [10].
Kolpakov et al. [6] proved that the function ρ is discontinuous at $\frac{7}{3}$, more

precisely they obtained that $\rho\left(\frac{7}{3}\right)=\frac{1}{2}$ and $\rho\left(\frac{7}{3}^{+}\right)\le\frac{10}{21}=0.47619047\ldots$. 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)	$\rho\left(\frac{7}{3}^{+}\right)$	$\le$	$\frac{327}{703}$	$=$	*0.4651493599...*
(2)	$\rho\left(\frac{17}{7}\right)$	$>$	$\frac{427}{918}$	$=$	*0.4651416122...*
(3)	$\rho\left(\frac{17}{7}^{+}\right)$	$\le$	$\frac{797}{1722}$	$=$	*0.4628339141...*
(4)	$\rho\left(\frac{5}{2}\right)$	$\ge$	$\frac{54286}{117293}$	$=$	*0.4628238684...*
(5)	$\rho\left(\frac{5}{2}^{+}\right)$	$\le$	$\frac{279}{631}$	$=$	*0.4421553090...*
(6)	$\rho\left(\frac{131}{52}\right)$	$>$	$\frac{107}{242}$	$=$	*0.4421487603...*
(7)	$\rho\left(\frac{131}{52}^{+}\right)$	$\le$	$\frac{191}{432}$	$=$	*0.4421296296...*
(8)	$\rho\left(\frac{43}{17}\right)$	$>$	$\frac{275}{622}$	$=$	*0.4421221865...*
(9)	$\rho\left(\frac{43}{17}^{+}\right)$	$\le$	$\frac{262}{593}$	$=$	*0.4418212479...*
(10)	$\rho\left(\frac{23}{9}\right)$	$>$	$\frac{1063}{2406}$	$=$	*0.4418121363...*
(11)	$\rho\left(\frac{23}{9}^{+}\right)$	$\le$	$\frac{860}{1947}$	$=$	*0.4417051875...*
(12)	$\rho\left(\frac{41}{16}\right)$	$>$	$\frac{322}{729}$	$=$	*0.4417009602...*
(13)	$\rho\left(\frac{41}{16}^{+}\right)$	$\le$	$\frac{79}{179}$	$=$	*0.4413407821...*
(14)	$\rho\left(\frac{18}{7}\right)$	$\ge$	$\frac{79}{179}$	$=$	*0.4413407821...*
(15)	$\rho\left(\frac{18}{7}^{+}\right)$	$\le$	$\frac{272}{617}$	$=$	*0.4408427877...*
(16)	$\rho\left(\frac{631}{245}\right)$	$>$	$\frac{231}{524}$	$=$	*0.4408396947...*
(17)	$\rho\left(\frac{631}{245}^{+}\right)$	$\le$	$\frac{190}{431}$	$=$	*0.4408352668...*
(18)	$\rho\left(\frac{8}{3}\right)$	$>$	$\frac{1885}{4276}$	$=$	*0.4408325538...*
(19)	$\rho\left(\frac{8}{3}^{+}\right)$	$\le$	$\frac{76}{187}$	$=$	*0.4064171123...*
(20)	$\rho\left(\frac{26}{9}\right)$	$>$	$\frac{165}{406}$	$=$	*0.4064039409...*
(21)	$\rho\left(\frac{26}{9}^{+}\right)$	$\le$	$\frac{89}{219}$	$=$	*0.4063926941...*
(22)	$\rho\left(\frac{44}{15}\right)$	$>$	$\frac{675}{1661}$	$=$	*0.4063816978...*
(23)	$\rho\left(\frac{44}{15}^{+}\right)$	$\le$	$\frac{332}{817}$	$=$	*0.4063647491...*
(24)	$\rho(3)$	$>$	$\frac{115}{283}$	$=$	*0.40636042...*
(25)	$\rho(3^{+})$	$\le$	$\frac{523}{1810}$	$=$	*0.2889502762...*

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 k-biprolongable word in L 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 cover. 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 $u \in S$. Then any infinite word $w \in L$ has a decomposition into finite factors $w_0 w_1 w_2 \ldots$ such that $|w_0| = k + \max_{u \in S} |u|$ and $\frac{|w_i|_0}{|w_i|} \geq q$ for every $i \geq 1$.

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 [13], where the suffix cover consist in the empty word, and of the one in [7], where the suffix cover consist in 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 $\{01, 111, 000, 1110, 1010, 0001111000010, 0111101000010, 1110101000010,$ $0111100010\}$. We omit the computer proof that this is indeed a suffix cover for $(\frac{5}{3}, 3)$-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. To construct an infinite word $w \in L$ with a given letter frequency $q \in \mathbb{Q}$, we basically use the method described in [10]. We note $q = \frac{a}{b}$ 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

$$\begin{aligned} m(0) &= 01232103, \\ m(1) &= 01230323, \\ m(2) &= 01210321. \end{aligned}$$

To get the bound $f_{\min} \le \frac{883}{3215}$ 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 $f_{\max} \ge \frac{255}{653}$ 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 [2] 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** [9] 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. $h_{\max}$) is 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 $\frac{7}{3} < e \le 3$. If $|p| < 2q$, then the length of the repetition less than $6q$, so that checking the h-image of every factor of t length 7 is sufficient. If $|p| \ge 2q$, then p contains a full h-image of somme 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 exist 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 $k \geq 5$. In its full generality, this conjecture is still open, although Pansiot [11] proved that $\alpha(4) = \frac{7}{5}$ and Moulin-Ollagnier [8] proved that Dejean's conjecture holds for $5 \leq k \leq 11$. For more information, see [1]. 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 $k \geq 6$.

References

[1] 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.

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

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

[4] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of Repetition Threshold, *Theoret. Comput. Sci.* To appear.

[5] 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.

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

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

[8] 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.

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

[10] 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.

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

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

[13] 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).