Theoretical Informatics and Applications Will be set by the publisher
Informatique Théorique et Applications

DEJEAN'S CONJECTURE AND LETTER FREQUENCY

Jérémie Chalopin[1] and Pascal Ochem[2]

Abstract. We prove two cases of a strong version of Dejean's conjecture involving extremal letter frequencies. The results are that there exist an infinite $\left(\frac{5}{4}^+\right)$-free word over a 5 letter alphabet with letter frequency $\frac{1}{6}$ and an infinite $\left(\frac{6}{5}^+\right)$-free word over a 6 letter alphabet with letter frequency $\frac{1}{5}$.

1991 Mathematics Subject Classification. 68R15.

1. Introduction

We consider the extremal frequencies of a letter in factorial languages defined by an alphabet size and a set of forbidden repetitions. Given such a language L, we denote by $f_{\min}$ (resp. $f_{\max}$) the minimal (resp. maximal) letter frequency in an infinite word that belongs to L. Extremal letter frequencies have been mainly studied in [5,7,10,11]. Let Σ_i denote the i-letter alphabet $\{0, 1, \ldots, i-1\}$. We consider here the frequency of the letter 0. Let $n(v)$ denote the number of occurrences of 0 in the finite word v. So the letter frequency in v is $\frac{n(v)}{|v|}$. We say that the letter frequency in the infinite word w is q if for every $\epsilon > 0$, there exists an integer n_ϵ such that for every finite factor v of w of length at least n_ϵ, we have $\left|\frac{n(v)}{|v|} - q\right| < \epsilon$.

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 [9] proved that $\alpha(4) = \frac{7}{5}$ and Moulin-Ollagnier [6] proved that Dejean's conjecture holds for $5 \leq k \leq 11$. Recently, Currie and Mohammad-Noori [3] also

[1] LIF, CNRS
Université de Provence, CMI, 39 rue Joliot-Curie, 13453 Marseille, France
jeremie.chalopin@lif.univ-mrs.fr

[2] LRI, CNRS
Université Paris-Sud 11, Bât 490, 91405 Orsay Cedex, France
ochem@lri.fr

proved the cases $12 \leq k \leq 14$, and Carpi [1] settled the cases $k \geq 33$. For more information, see [2].

In a previous paper, we proposed the following conjecture which implies Dejean's conjecture.

Conjecture 1.1. [7]

(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 1.1 are best possible. For $\left(\frac{5}{4}^+\right)$-free words over Σ_5, we obtained $f_{\max} < \frac{103}{440} = 0.23409090\dots < \frac{1}{4}$ [7]. That is why Conjecture 1.1.2 is stated with $k \geq 6$.

In this paper, we prove the first case of each part of Conjecture 1.1:

Theorem 1.2.

(1) *There exists an infinite* $\left(\frac{5}{4}^+\right)$*-free word over* Σ_5 *with letter frequency* $\frac{1}{6}$.

(2) *There exists an infinite* $\left(\frac{6}{5}^+\right)$*-free word over* Σ_6 *with letter frequency* $\frac{1}{5}$.

The C++ sources of the programs and the morphisms used in this paper are available at: `http://www.lri.fr/~ochem/morphisms/`.

2. Structure and encoding

In the following, a *k-word* will denote a $\left(\frac{k}{k-1}^+\right)$-free word w over Σ_k, for $k \geq 5$. We easily check that in a k-word, the distance between two consecutive occurrences of the same letter is either $k-1$, k, or $k+1$. This implies that if there exists an infinite k-word with letter frequency $\frac{1}{k+1}$ (resp. $\frac{1}{k-1}$), then there exists an infinite k-word such that the distance between consecutive occurrences of 0 is always $(k+1)$ (resp. $(k-1)$). Notice that 0's cannot be regularly spaced if the letter frequency is $\frac{1}{k}$. Such k-words in which 0's are regularly spaced are the catenation of factors of size $k+1$ (resp. $k-1$) of the form $0\pi_1\dots\pi_{k-1}\pi_1$ (resp. $0\pi_1\dots\pi_{k-2}$), where π is a permutation of the elements $[1,\dots,k-1]$. Let Π_k denote the set of permutations of $[1,\dots,k-1]$. The k-word w can thus be encoded by the word $p \in \Pi_k^*$ consisting in the catenation of the permutations that correspond to the factors of size $k+1$ (resp. $k-1$) in w.

Let $p = p_0p_1p_2\dots$ be the code of w and we suppose that p_0 is the identity. We now encode p by the word $c = c_0c_1c_2\dots \in \Pi_k^*$ such that $p_{i+1} = c_i(p_i)$. Notice that whereas any permutation in Π_k may appear in p, only a small subset $S \subset \Pi_k$ of permutation can be used as letters in c. This is because the latter permutations rule the transition between two consecutive factors w_i and w_{i+1} in w, and then w_iw_{i+1} has to be $\left(\frac{k}{k-1}^+\right)$-free.

In the following, we will call *coding word* a word over S^* consisting of the transition permutations for a k-word. Moreover, if the transition corresponding to a coding word c is the identity of Π_k, we will say that c is an identity.

Remark 2.1. If a coding word c is an identity, then every conjugate (cyclic shift) of c is also an identity.

3. Proof of main result

Let us consider the possible transitions for 5-words with letter frequency $\frac{1}{6}$. There are exactly two of them:

- 012341024312 corresponds to the transition permutation 2431 (noted 0).
- 012341032143 corresponds to the transition permutation 3214 (noted 1).

There are also exactly two possible transitions for 6-words with letter frequency $\frac{1}{5}$:

- 0123405132 corresponds to the transition permutation 51324 (noted 0).
- 0123405213 corresponds to the transition permutation 52134 (noted 1).

In both cases, we have $|S| = 2$ and we construct a suitable infinite code c as the fixed point of the following binary endomorphisms:

- For 5-words with letter frequency $\frac{1}{6}$:
  $0 \mapsto 010010010100101001001001010100101001001001010010101001001001010$
  $1 \mapsto 100101001001010100101001001010100101001001001010010010010100101$

- For 6-words with letter frequency $\frac{1}{5}$:
  $0 \mapsto 001001010011100011010001 0$
  $1 \mapsto 1000100111000100110100011$

These morphisms m satisfy the following properties:

1. m is q-uniform, that is, for all $i \in \Sigma_2$, we have $|m(i)| = q$.
2. m is *synchronizing*, which means that for any $a, b, c \in \Sigma_2$ and $s, r \in \Sigma_2^*$, if $m(ab) = rm(c)s$, then either $r = \varepsilon$ and $a = c$ or $s = \varepsilon$ and $b = c$.
3. for all $i \in \Sigma_2$, we have $m(i) = ifi$ and the factor if is an identity (thus fi is also an identity by Remark 2.1).

Let Φ denote the decoding function. In the case of 6-words, we thus have $\Phi(0) = 0123405132$, $\Phi(1) = 0123405213$ and $\Phi(c = m^\omega(0)) = w$. We have checked that for every factor x of c of size at most $2kq$, $\Phi(x)$ is $\left(\frac{k}{k-1}^+\right)$-free.

Let f be a smallest repetition in w of exponent strictly greater than $\frac{k}{k-1}$. This repetition in w implies that there is a repetition $r = is$ in c whose prefix i is an identity. Since $|s| \geq 2q$, s contains a full m-image. So $|i|$ and $|s|$ are multiples of q because m is synchronizing. By property 3 and Remark 2.1, we can assume without loss of generality that $|i|$ starts at the beginning of an m-image. Then our repetition is of the form $r = is = m(i')m(s') = m(i's') = m(r')$. By property 3, r' is a repetition whose prefix i' is an identity, and thus the factor $\Phi(r')$ is a repetition that appears in w. The exponent of $\Phi(r')$ is $\frac{(k\pm1)(|r'|+1)}{(k\pm1)|i'|}$, the exponent of f is less than $\frac{(k\pm1)(|r|+3)}{(k\pm1)|i|} = \frac{(k\pm1)(q|r'|+3)}{(k\pm1)q|i'|}$, and we have $\frac{(k\pm1)(|r'|+1)}{(k\pm1)|i'|} \geq \frac{(k\pm1)(q|r'|+3)}{(k\pm1)q|i'|}$ if $q \geq 3$. This is a contradiction because the exponent of $\Phi(r')$ is greater than the exponent of f and $\Phi(r')$ is strictly smaller than f.

4. Concluding remarks

Theorem 1.2 shows the existence of two types of infinite words, but does not prove that there exist exponentially many such words (which is probably true). On the other hand, the growth rate of these words is significantly smaller than those of $\left(\frac{k}{k-1}^+\right)$-free words. For example, the growth rate of 5-words is about 1.159 [8], whereas 1.048 is a rough upper bound on the growth rate of 5-words with letter frequency $\frac{1}{6}$.

Other cases of Conjecture 1.1 might be harder to settle. For 6-words with letter frequency $\frac{1}{7}$, we have $|S| = 3$, and it is impossible to construct an infinite code using only two of these three transition permutations. We have not been able to find a $\Sigma_3^* \to \Sigma_3^*$ morphism with suitable properties for them.

References

[1] A. Carpi. On Dejeans conjecture over large alphabets, *Theoret. Comput. Sci.* **385** (2007), 137–151.

[2] 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.

[3] M. Mohammad-Noori and J.D. Currie. Dejean's conjecture and Sturmian words, *Europ. J. Combinatorics* **28(3)** (2007), 876–890.

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

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

[6] 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.

[7] P. Ochem. Letter frequency in infinite repetition-free words, *Theoret. Comput. Sci.* 380 (2007), 388–392.

[8] P. Ochem and T. Reix. Upper bound on the number of ternary square-free words, In *Workshop on Words and Automata (WOWA'06)*, St. Petersburg, Russia, June 7 2006.

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

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

[11] 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).

Communicated by (The editor will be set by the publisher).