MORE ON GENERALIZED REPETITION THRESHOLDS

FRANCESCA FIORENZI AND PASCAL OCHEM

Abstract. The *repetition threshold* introduced by Dejean and Brandenburg is the smallest real number α such that there exists an infinite word over a k-letter alphabet that avoids β-powers for all $\beta > \alpha$. Ilie, Ochem, and Shallit generalized this concept to include the lengths of the avoided words. We give a lower and an upper bound on this generalized repetition threshold.

1. Introduction

The study of repetitions in words has been one of the main topics in combinatorics on words. Thue [12] showed the existence of an infinite square-free word on three letters, that is without concatenated occurrences of the same non empty factor. This fact was actually implicated by the existence of an infinite binary overlap-free word (i.e. without factors of the form $uvuvu$ with u non empty).

A natural extension of this problem takes into account repetitions of a fractional order, where the order of a non empty finite word is the ratio between its length and its period. This notion has been introduced by Dejean [6] and Brandenburg [1]. Dejean proved that every sufficiently long word over a three-letter alphabet contains a 7/4-power, and this bound is the best possible. The least real number $\alpha > 1$ such that there exists an infinite word on k letters avoiding β-powers for all $\beta > \alpha$ is called the repetition threshold on k letters. Thus Thue's result implies that the repetition threshold on two letters is 2, while Dejean's result means that the repetition threshold on three letters is 7/4.

Dejean observed that for $k \geq 5$, the repetition threshold is not smaller than $\frac{k}{k-1}$, while for $k = 4$ it is not smaller than 7/5. She conjectured that these are the actual values of the repetition thresholds. This conjecture has been proved true for $k = 4$ by Pansiot [10]. For $k \geq 5$, the conjecture has been solved thanks to the contribution of many authors: Moulin-Ollagnier [9], Currie and Mohammad-Noori [3], Rao [11], Currie and Rampersad [4, 5], and Carpi [2].

In [7] the authors generalize the repetition threshold of Dejean to handle avoidance of all sufficiently large fractional powers. They define a real number, depending on the size k of the alphabet and on the length ℓ of a repetition. This number naturally extends the classical notion of repetition threshold. Moreover, in [7] its value has been calculated in some particular cases and general lower and upper bounds have been given. In this paper we improve these bounds by studying the asymptotics of the generalized repetition threshold.

2. Definitions

Let $\alpha > 1$ be a rational number, and let $\ell \geq 1$ be an integer. A word w is an (α, ℓ)-*repetition* if $w = (uv)^n u$, where $|uv| = \ell$ and $\alpha = \frac{|w|}{\ell}$. In this case α is called *order* of the repetition and ℓ its *length* or *period*. Notice that in our definitions, order and length of a repetition are not univocally defined. For instance, the word $aabcaabcaa$ on the alphabet $\{a, b, c\}$ is a $(\frac{5}{2}, 4)$-repetition, a $(\frac{5}{4}, 8)$-repetition and a $(\frac{10}{9}, 9)$-repetition. An *α-power* is an (α, ℓ)-repetition for some ℓ. We say a word *(α, ℓ)-free* if it contains no factor that is an (α', ℓ')-repetition with $\alpha' \geq \alpha$ and $\ell' \geq \ell$. We say a word *(α^+, ℓ)-free* if it is (α', ℓ)-free for all $\alpha' > \alpha$. Finally, a word is *α-free* if it does not contain α'-powers with $\alpha' \geq \alpha$ and is *α^+-free* if it is α'-free for all $\alpha' > \alpha$.

Let Σ_k denote the k-letter alphabet $\{0, 1, \ldots, k-1\}$. For integers $k \geq 2$ and $\ell \geq 1$, the *generalized repetition threshold* $R(k, \ell)$ is defined as the smallest real number α such that there exists an infinite (α^+, ℓ)-free word over Σ_k. Actually, there always exists an infinite $(R(k, \ell)^+, \ell)$-free word over Σ_k. Nevertheless, as pointed out in [7], there is no known instance of a $(R(k, \ell), \ell)$-free infinite word. Finally, notice that by definition, $R(k+1, \ell) \leq R(k, \ell)$ and $R(k, \ell+1) \leq R(k, \ell)$.

The finiteness of $R(k, \ell)$ is due to the existence of an infinite binary overlap-free word. Ilie *et al.* [7] also obtained a lower bound on $R(k, \ell)$, namely

$$1 + \frac{\ell}{k^\ell} \leq R(k, \ell) \leq 2.$$

The aim of the paper is to improve the above inequalities.

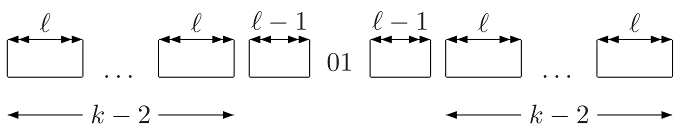
The case $\ell = 1$ corresponds to the classical repetition threshold and the values of $R(k, 1)$ are now all determined. Moreover, the proof of our upper bound explicitly uses the fact that $R(k, 1) = \frac{k}{k-1}$ for $k \geq 5$.

3. Lower Bound

A natural way of obtaining a bound of the form $R(k, \ell) \geq \alpha$ is to show that sufficiently long words over Σ_k contain a repetition uvu with length $|uv| \geq \ell$ and order at least α. In [7], the proof of $R(k, \ell) \geq 1 + \frac{\ell}{k^\ell}$ focused on repetitions such that $|u| = \ell$ in order to imply that $|uv| \geq \ell$. The next result mainly uses repetitions such that $|u| = 1$, and marginally such that $|u| = \ell$.

Theorem 1. *For $k \geq 3$ and $\ell \geq 2$, we have $R(k, \ell) \geq 1 + \frac{1}{(k-1)\ell}$.*

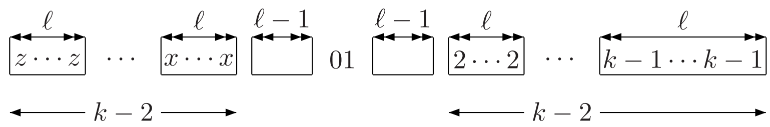
Proof. Consider a word over Σ_k containing a block w of the following form



Let w_1 be the the suffix of w starting at 01, we have that $|w_1| = (k-1)\ell + 1$. Suppose that in w_1 there is a repetition of length $\geq \ell$. It is easy to see that in a word of length n, the minimum exponent of a repetition is $\frac{n}{n-1}$. Hence, the repetition in w_1 has order $\geq 1 + \frac{1}{(k-1)\ell}$, and the theorem is proved.

Suppose now that in w_1 there is no repetition of length $\geq \ell$. This implies that the maximal distance between to occurrences of the same letter is $\ell - 1$ and in particular each letter of the alphabet appears at most ℓ times in w_1. Notice that, because of the distance argument, the $(k-2)\ell$-block at the end of w_1, do not contain letters 0 and 1. By the pigeonhole principle we have that each letter of the alphabet $\Sigma_k \setminus \{0, 1\}$ appears exactly ℓ times in this $(k-2)\ell$-block. Without loss of generality, we can suppose that these letters appear in the increasing order.

Let w_0 be the prefix of w ending at 01. By the same argument we have that w_0 either contains a repetition of length $\geq \ell$ and order $\geq 1 + \frac{1}{(k-1)\ell}$, or each letter of the alphabet $\Sigma_k \setminus \{0,1\}$ appears exactly ℓ times in the $(k-2)\ell$-block prefixing w_0.



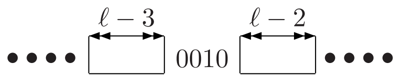
Consider the suffix of w of length $(k+1)\ell$ starting at the ℓ-block $x \cdots x$. This word contains a repetition of length $\geq 3\ell$ and order $\geq 1 + \frac{1}{k} \geq 1 + \frac{1}{(k-1)\ell}$.

Hence, in order to prove our statement, we can exclude the case of a sufficiently long word containing a factor xy with $x \neq y$. This implies that the only remaining cases are those of words $0^\omega, 1^\omega, \ldots, (k-1)^\omega$, which obviously contain repetitions of arbitrarily great order and length. □

Theorem 2. $R(2,\ell) \geq 1 + \frac{2}{\ell+2}$.

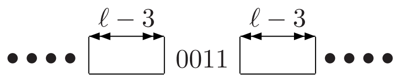
Proof. Suppose for the sake of contradiction that $R(2,\ell) < 1 + \frac{2}{\ell+2}$. That is, there exists an infinite binary word w with no repetition of length $\geq \ell$ and order $\geq 1+\frac{2}{\ell+2}$. In particular, repetitions uvu such that $|u| = 2$ and $\ell \leq |uv| \leq \ell + 2$ are forbidden in w. Moreover, we can assume without loss of generality that w is *recurrent*, that is, every finite factor of w appears infinitely many times in w.

First, we check that the factor 0010 is forbidden in w.



By previous considerations about the distances, we have that the blocks $\bullet\bullet\bullet\bullet$ on the left and on the right of the factor 0010 must be 1111. This creates a repetition of length $2\ell + 3$ and order $1 + \frac{4}{2\ell+3} > 1 + \frac{2}{\ell+2}$. Since 0010 is forbidden, the factors 0100, 1101, and 1011 are also forbidden by symmetry.

Now, we check that the factor 0011 is forbidden in w.



The block $\bullet\bullet\bullet\bullet$ on the left of the factor 0011 must be $\bullet 10 \bullet$, then $110\bullet$, and finally 1100 since 1101 is forbidden. The block $\bullet\bullet\bullet\bullet$ on the right of the factor 0011 must be $\bullet 10 \bullet$, then $\bullet 100$, and finally 1100 since 0100 is forbidden. This creates a repetition of length $2\ell + 2$ and order $1 + \frac{2}{\ell+1} > 1 + \frac{2}{\ell+2}$.

The factor 001 is forbidden because 0010 and 0011 are forbidden. By symmetry, 100, 110, and 011 are also forbidden. Thus, the only remaining possibilities for w are the words 0^ω, 1^ω, and $(01)^\omega$, which obviously contain repetitions of arbitrarily great order and length. □

For $k \geq 3$, the lower bound $1 + \frac{1}{(k-1)\ell}$ is at least as good as the previous one, namely $1 + \frac{\ell}{k^\ell}$. The same holds when $k = 2$ with the lower bound $1 + \frac{2}{\ell+2}$.

Theorem 2 is certainly not optimal, since numerical evidences suggest that $R(2, \ell) = 1 + \frac{1}{\ell/3+1}$ for $\ell \geq 6$, $\ell = 0 \pmod 3$. We also mention that Kolpakov and Rao [8] have proved that $R(3, \ell) \geq 1 + \frac{1}{\ell}$, which would be a tight lower bound for the conjecture in [7] that $R(3, \ell) = 1 + \frac{1}{\ell}$ for $\ell \geq 2$.

4. Upper bound

In this section we use the fact that Dejean's conjecture is proved for $N \geq 5$. We describe a morphism from an N-letter alphabet to a k-letter one that transforms an infinite $\left(\frac{N}{N-1}\right)^+$-free word into a word in which the sufficiently large repetitions are of order not much bigger than $\frac{N}{N-1}$.

Let $S_{k,t}$ be the set of words of length t over Σ_k of the form $0^e w$ where $e \geq 2$, $|w| \geq 1$, the first and the last letter of w are different from 0, and w does not contain 00 as a factor. For example, we have that $S_{2,5} = \{00001, 00011, 00101, 00111\}$. Let $N = |S_{k,t}|$ and let h be a t-uniform morphism $h : \Sigma_N^* \to \Sigma_k^*$ such that the set of h-images of letters in Σ_N is $S_{k,t}$. Now, if $N \geq 5$, we consider the h-image of some infinite $\left(\frac{N}{N-1}\right)^+$-free word over Σ_N.

A uniform morphism $m : \mathcal{A}^* \to \mathcal{B}^*$ is said to be *synchronizing* if for any $a, b, c \in \mathcal{A}$ and $s, r \in \mathcal{B}^*$, $m(ab) = rm(c)s$ implies that either $r = \varepsilon$ and $a = c$ or $s = \varepsilon$ and $b = c$.

Remark. A synchronizing morphism m is always injective (actually it is injective on the set $\mathcal{A}$ of monoid generators). Moreover, if it is t-uniform, then for each factor u of a word in $m(\mathcal{A}^*)$ such that $|u| \geq 2t - 1$, there exists a unique factorization $u = xm(u')y$ where $u' \in \mathcal{A}^*$ and $0 \leq |x|, |y| < t$.

Lemma 3. *The t-uniform morphism $h : \Sigma_N^* \to \Sigma_k^*$ defined above is synchronizing.*

Proof. Suppose that the h is not synchronizing. Then there exist $a, b, c \in \Sigma_N$ and $s, r \in \Sigma_k^*$ such that $m(ab) = rm(c)s = w[1, \dots, 2t]$ with $0 < |r| < t$. We obtain a contradiction for every possible value of $|r|$:

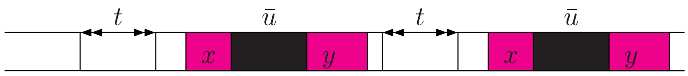
- if $|r| = 1$ or $|r| = 2$, then the letter $w[t + |r|]$ is 0 in $m(ab)$ and is not 0 in $rm(c)s$,
- if $|r| = t - 1$ or $|r| = t - 2$, then the letter $w[t]$ is not 0 in $m(ab)$ and is 0 in $rm(c)s$,
- if $2 < |r| < t - 2$, then $m(c)$ contains the factor $w[t, \dots, t + 2]$. In $m(ab)$, this factor is of the form $x00$ with $x \neq 0$, whereas factors of this form do not exist in $m(c)$.

□

In order to get the mentioned repetition-freeness property in $h(\Sigma_N^*)$, we use the following lemma. As it will appear clear from the context, for any real number $\ell \geq 1$, we write (α^+, ℓ)-free to mean $(\alpha^+, \lceil \ell \rceil)$-free and hence $R(k, \ell)$ to mean $R(k, \lceil \ell \rceil)$.

Lemma 4. *Let $\alpha, \beta \in \mathbb{R}$, $1 < \alpha < \beta < 2$. Let $h : \mathcal{A}^* \to \mathcal{B}^*$ be a synchronizing t-uniform morphism. If $w \in \mathcal{A}^*$ is α^+-free, then $h(w)$ is $(\beta^+, \frac{2t-2}{\beta-\alpha})$-free.*

Proof. Let uvu be a β'-repetition in $h(w)$ with $\beta' > \beta$. Suppose $|u| \geq 2t-1$. Hence u contains an h-image and it can be uniquely written as $xh(u')y$ with $0 \leq |x|, |y| < t$. Thus this h-image $\bar{u} = h(u')$ appears at the same position of u in uvu.



The factor $\bar{v} = yvx$ is an h-image. We have that $\bar{u}\bar{v}\bar{u}$ is the h-image of a repetition in w and hence $\frac{|\bar{u}\bar{v}\bar{u}|}{|\bar{u}\bar{v}|} \leq \alpha$. Moreover, $\beta' = \frac{|\bar{u}\bar{v}\bar{u}|}{|\bar{u}\bar{v}|} + \frac{|x|+|y|}{|uv|}$ and $\beta' > \beta$ implies that $\frac{|x|+|y|}{|uv|} > \beta - \alpha$. Hence

$$|uv| < \frac{|xy|}{\beta - \alpha} \leq \frac{2t-2}{\beta - \alpha}.$$

Suppose now that $|u| \leq 2t-2$. Hence $\beta' > \beta$ implies that $\frac{|u|}{|uv|} > \beta - 1$. Thus $\frac{|uv|}{|u|} < \frac{1}{\beta - 1}$ and

$$|uv| < \frac{|u|}{\beta - 1} \leq \frac{2t-2}{\beta - 1} < \frac{2t-2}{\beta - \alpha}.$$

□

Recall that, by definition, if there exists a (β^+, ℓ)-free infinite word over Σ_k then $R(k, \ell) \leq \beta$.

Corollary 5. *Let $h : \Sigma_N^* \to \Sigma_k^*$ be a synchronizing t-uniform morphism as above. If $\beta \in \mathbb{R}$, $\frac{N}{N-1} < \beta < 2$, and $N \geq 5$, then $R\left(k, \frac{2t-2}{\beta - \frac{N}{N-1}}\right) \leq \beta$.*

We now compute $N = N_{k,t}$. Consider the prefixes of length $(t-1)$ of the words in $S_{k,t}$:

- $N_{k,t-1}$ of them are such that the last letter is not 0,
- $N_{k,t-2}$ of them are such that the last letter is 0 and the penultimate letter is not 0,
- one them is the word 0^{t-1}.

Each prefix can be extended by one of the $(k-1)$ letters distinct from 0 to get a word in $S_{k,t}$, so $N_{k,t}$ satisfies the recurrence relation

$$N_{k,1} = N_{k,2} = 0, N_{k,t} = (k-1)(N_{k,t-1} + N_{k,t-2} + 1).$$

Solving this relation, we obtain that

$$N_{k,t} = \frac{k-1}{(2k-3)\sqrt{(k-1)(k+3)}} \left(\lambda^t - \mu^t - (k-2)(\lambda^{t-1} - \mu^{t-1})\right) - \frac{k-1}{2k-3},$$

where $\lambda = \frac{(k-1)+\sqrt{(k-1)(k+3)}}{2}$ and $\mu = \frac{(k-1)-\sqrt{(k-1)(k+3)}}{2}$.

We thus have

$$N_{k,t} = C_k \lambda^{t-1} - O(1), \text{ where } C_k = \frac{(k-1)(\sqrt{(k-1)(k+3)} - k + 3)}{2(2k-3)\sqrt{(k-1)(k+3)}}.$$

Theorem 6. *If k is fixed and ℓ tends to infinity, then $R(k, \ell) \leq 1 + \frac{2 \ln \ell}{\ell \ln \lambda} + O\left(\frac{1}{\ell}\right)$.*

Proof. Let us fix $t = \lfloor \frac{\ln \ell}{\ln \lambda} \rceil + 1$ and $\beta = 1 + \frac{2t-2}{\ell} + \frac{1}{N-1}$. For ℓ sufficiently large, we have $t \geq 6$ which ensures that $N \geq 5$, and we also have $\beta < 2$. We can thus use Corollary 5, which gives

$$R(k, \ell) = R\left(k, \frac{2t-2}{\beta - \frac{N}{N-1}}\right) \leq \beta = 1 + \frac{2t-2}{\ell} + \frac{1}{N-1}.$$

Since

$$\frac{2t-2}{\ell} = \frac{2\lfloor \frac{\ln \ell}{\ln \lambda} \rceil}{\ell} = \frac{2 \ln \ell}{\ell \ln \lambda} + O\left(\frac{1}{\ell}\right)$$

and

$$\frac{1}{N-1} = \frac{1}{C_k \lambda^{t-1} - O(1)} = \frac{1}{C_k \lambda^{\lfloor \frac{\ln \ell}{\ln \lambda} \rceil} - O(1)} = O\left(\frac{1}{\ell}\right),$$

the result follows. □

5. An example

Let us illustrate our results with a concrete example: $k = 8$ and $\ell = 100$. Theorem 1 gives $R(8, 100) \geq 1 + \frac{1}{(8-1)\times 100} = 1.00142857\ldots$. For the upper bound, we have to decide which morphism h will be used, or equivalently to choose the value of the parameter t. For a given t, we can compute $N = N_{k,t}$ and then the bound $\beta = 1 + \frac{2t-2}{\ell} + \frac{1}{N-1}$. So, we have to choose t so that β is minimized. The choice of $t = \lfloor \frac{\ln \ell}{\ln \lambda} \rceil + 1$ in Theorem 6 is well-suited to get such an asymptotic result, but for a given pair (k, ℓ) like this example, it is better to make a specific case study.

- If $t = 3$, then $N = 7$ and $\beta = \frac{181}{150} = 1.20666666\ldots$.
- If $t = 4$, then $N = 56$ and $\beta = \frac{593}{550} = 1.07818181\ldots$.
- If $t = 5$, then $N = 448$ and $\beta = \frac{12094}{11175} = 1.08223713\ldots$.

Since β gets bigger if $t > 5$, the minimum is reached at $t = 4$, whereas $\lfloor \frac{\ln \ell}{\ln \lambda} \rceil + 1 = 3$. We thus obtain $R(8, 100) \leq \frac{593}{550} \leq 1.078182$.

6. Conclusion

For k fixed and ℓ tending to infinity, we know now in particular that the asymptotics of the generalized repetition threshold $R(k, \ell)$ is between $1 + \Omega(1/\ell)$ and $1 + O(\ln \ell / \ell)$. New ideas are needed to settle this and other questions about $R(k, \ell)$, such as good estimates for $R(k, 2)$ or $R(k, k)$. The case $1.001428 \leq R(8, 100) \leq 1.078182$ suggests that there is still room for improvement.

References

[1] F.-J. Brandenburg. Uniformly growing k-th power-free homomorphisms, *Theor. Comput. Sci.* **23** (1983), 69–82.
[2] A. Carpi. On Dejean's conjecture over large alphabets, *Theoret. Comput. Sci.* **35** (2007), 137–151.
[3] J.D. Currie and M. Mohammad-Noori. Dejean's conjecture and sturmian words, *Eur. J. Combin.* **28(3)** (2007), 876–890.
[4] J.D. Currie and N. Rampersad. Dejean's conjecture holds for $n \geq 27$, http://arxiv.org/abs/0901.3188
[5] J.D. Currie and N. Rampersad. Dejean's conjecture holds for $n \geq 30$, *Theoret. Comput. Sci.* To appear.
[6] F. Dejean. Sur un théorème de Thue, *J. Combin. Theory. Ser. A* **13** (1972), 90–99.
[7] L. Ilie, P. Ochem, and J.O. Shallit. A generalization of repetition threshold, *Theoret. Comput. Sci.* **345** (2005), 359–369.

[8] R. Kolpakov and M. Rao. Personal communication.

[9] 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.

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

[11] M. Rao. Last cases of Dejean's conjecture, *WORDS 2009*.

[12] A. Thue. Über unendliche Zeichenreihen, Christiania Vidensk.-Selsk. Skrifter. I. Mat. Nat. Kl, 7 (1906) 1–22. Rreprinted in: T. Nagell (Ed.), Selected Mathematical Papers of Axel Thue, Universitetsforlaget, Oslo, 1977, pp. 139–158.

Francesca Fiorenzi, Laboratoire de Recherche en Informatique, UMR 8623, Université Paris-Sud, 91405 Orsay, France

E-mail address: `fiorenzi@lri.fr`

Pascal Ochem, Laboratoire de Recherche en Informatique, UMR 8623, Université Paris-Sud, 91405 Orsay, France & Laboratoire J.-V.Poncelet, Moscow

E-mail address: `ochem@lri.fr`