On maximal repetitions of arbitrary exponent

Roman Kolpakov[*]   Gregory Kucherov[†]   Pascal Ochem[‡]

June 19, 2018

Abstract

The first two authors have shown [KK99, KK00] that the sum the exponent (and thus the number) of maximal repetitions of exponent at least 2 (also called runs) is linear in the length of the word. The exponent 2 in the definition of a run may seem arbitrary. In this paper, we consider maximal repetitions of exponent strictly greater than 1.

Keywords: theory of computation, combinatorial problems, repetitions, periodicities

1 Introduction

Repetitions (periodicities) are fundamental concepts in word combinatorics [Lot83, CK97, KK05]. Recall that each word w is characterized by the *minimal period* $p(w)$ and by the exponent $e(w)$ which is the ratio $\frac{p(w)}{|w|}$. A great deal of work in word combinatorics has been devoted to the study of words that do not contain subwords of a given exponent [CK97]. Another research direction, of more algorithmic nature, is the efficient identification of all subwords of a given exponent in a word [KK05], which raises the combinatorial question of the possible number of such subwords.

In [KK99, KK00], the first two authors considered the notion of *maximal repetitions* of a word, which are subword occurrences that cannot be extended outwards without changing their minimal period. They proved that the number of maximal repetitions of exponent at least 2 is linearly bounded in the length of word. It has been conjectured that this number is actually smaller than the word length. It has been also proved that not only the number of maximal repetitions of exponent 2 or more is linearly bounded, but the sum of exponents of these repetitions is linearly bounded too. The linear bound on the number of repetitions, in turn, allowed them to prove that all such maximal repetitions can be found in linear time. More recently, other researchers attempted to improve these results by finding a simpler proof of the linear bound implying a smaller multiplicative constant. The last current achievement in this direction is presented in [CIT08].

---

[*]Moscow University, Russia, `foroman@mail.ru`

[†]CNRS (LIFL, Lille and J.-V.Poncelet Lab., Moscow) and INRIA Lille - Nord Europe, France, `Gregory.Kucherov@lifl.fr`

[‡]CNRS (LRI, Orsay and J.-V.Poncelet Lab., Moscow) France, `Pascal.Ochem@lri.fr`

A big question that remained open in this development concerns the lower bound of 2 on the exponent of considered repetitions. While this bound is intuitively natural (as it requires some subword to be consecutively repeated at least twice), it has no formal justification. Moreover, word combinatorics provides many separation results when the "right" bound on the exponent is not an "intuitive" number. For example, the famous Dejean's result states that the exponents that can be avoided on a ternary alphabet are exponents greater than $\frac{7}{4}$ [Dej72]. As another example, there are exponentially many binary words avoiding exponents greater than $\frac{7}{3}$, while there are only polynomially many of them avoiding smaller exponents [KS04].

In this paper, we completely lift the lower bound on the exponent and focus on the maximal repetitions of *any* exponent greater than 1. Note that repetitions with exponent between 1 and 2 are subwords of the form uvu that can be viewed as *non-consecutive repetitions*. Therefore, in this paper we consider both consecutive (periodicities) and non-consecutive repetitions. To the best of our knowledge, the number of repetitions of exponent smaller than 2 has not been studied.

Instead of directly counting the repetitions or the sum of their exponents, we consider the *sum of exponents decremented by* 1. The main idea is that repetitions with exponents close to 1 (i.e. subwords uvu with $|v| \gg |u|$) contribute to the sum with an amount close to 0. We prove that this sum is upper-bounded by $n \ln(n)$ (Theorem 1) which immediately implies that the number of maximal repetitions of *any* exponent greater than $1 + \varepsilon$ is bounded by $\frac{1}{\varepsilon} n \ln(n)$. On the other hand, the number of *all* maximal repetitions can be quadratic (Theorem 5). We also obtain that the lower bound for the sum is $\frac{|w|}{k} - 1$, where k is the alphabet size, and we characterize the word achieving this lower bound (Theorem 6). Finally, we study this sum for the words containing only repetitions with a period bounded by a constant.

While the "whole picture" of the count of the number of maximal repetitions with exponent smaller than 2 is still incomplete, we believe that our results represent the first step in this direction.

2 Definitions

Recall that for any word w, the (minimal) *period*, denoted $p(w)$ is the minimal natural p such that $w[i] = w[i + p]$ whenever positions i and $i + p$ both exist in w. The *exponent* of w is defined as $e(w) = \frac{|w|}{p(w)}$ ($|w|$ is the length of w). A *root* of w is any subword of w of length $p(w)$. The prefix (resp. suffix) root of w is the prefix (resp. suffix) of w of length $p(w)$.

Given w, a *maximal repetition* in w is a subword $w[i..j]$ such that $p(w[i..j]) > p(w[i-1..j])$ (provided that $i \neq 1$) and $p(w[i..j]) > p(w[i..j+1])$ (provided that $j \neq |w|$). Informally, "maximality" means that the subword is extended outwards as much as possible so long as its period is preserved.

In this paper, we will be interested in *maximal repetitions* of any exponent greater than 1. The set of these subwords of w will be denoted $\mathcal{M}(w)$.

Note that any two occurrences of the same letter in w define a maximal rep-

etition with a period that is a divisor of the distance between these occurrences. In this case, we will speak about a maximal repetition *defined by* a letter match.

3   Sum of decremented exponents

For a word w, we will be interested in the sum of exponents of all maximal repetitions, decremented by 1:

$$\sum_{r \in \mathcal{M}(w)} (e(r) - 1). \tag{1}$$

This quantity can be viewed as the difference between the sum of exponents of all maximal repetitions and the number of these repetitions.

Theorem 1. *For every word w of length n, we have $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \leq n \ln(n)$.*

Proof. For each maximal repetition r with period p, we distribute the value $e(r) - 1 = \frac{|r| - p}{p}$ over $(|r| - p)$ pairs of matching letters $(w[i], w[i + p])$, $w[i] = w[i + p]$ within the repetition. Each such pair contributes to the sum with weight $\frac{1}{p}$. Consider two positions i and j, $1 \leq i < j \leq n$, in w. If $w[i] = w[j]$, then this match participates in some repetition, but it is counted only if the period of this repetition is $j - i$, in which case it contributes to the sum with the amount $\frac{1}{j-i}$. We thus have $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \leq \sum_{1 \leq i < j \leq n} \frac{1}{j-i} = \sum_{i=1}^{n-1} \frac{n-i}{i} = n \sum_{i=1}^{n-1} \frac{1}{i} - (n - 1) \leq n \ln(n)$ for $n > 2$. $\quad\square$

If we count only maximal repetitions of period at most p, then the following bound holds.

Corollary 2. *For every word w of length n, we have $\sum_{r \in \mathcal{M}(w), p(r) \leq p} (e(r) - 1) \leq n(\ln(p) + 1)$.*

Proof. If only repetitions of period at most p are considered, then, according to the proof of Theorem 1, the sum is bounded as follows. $\sum_{r \in \mathcal{M}(w), p(r) \leq p} (e(r) - 1) \leq \sum_{1 \leq i < j \leq \min\{i+p, n\}} \frac{1}{j-i} \leq n(\ln(p) + 1)$. $\quad\square$

Complementarily, if we count only maximal repetitions of period at least p, then we get

Corollary 3. *For every word w of length n, we have $\sum_{r \in \mathcal{M}(w), p(r) \geq p} (e(r) - 1) \leq n \ln(n/p)$.*

Proof. Similar to Corollary 2. $\quad\square$

Assume now that we focus only on maximal repetitions of exponent $(1 + \varepsilon)$ or more, and we want to count their number. Theorem 1 immediately provides a nontrivial upper bound.

Corollary 4. *For every word w of length n and every $\varepsilon > 0$, the number of maximal repetitions of exponent at least $(1 + \varepsilon)$ in w is at most $\frac{1}{\varepsilon} n \ln(n)$.*

Proof. Consider the sum of Theorem 1. Each repetition contributes at least ε to it and therefore the number of those is at most $\frac{n \ln(n)}{\varepsilon}$. □

Similarly, Corollaries 2 and 3 imply respective upper bounds $\frac{1}{\varepsilon} n \ln(p)$ and $\frac{1}{\varepsilon} n \ln(n/p)$ on the number of maximal repetitions of exponent at least $(1 + \varepsilon)$ and of period respectively at most p and at least p.

The following Theorem shows that the upper bound of Theorem 1 is asymptotically tight within a factor of 8 and that the number of *all* repetitions of arbitrary exponent can be quadratic (to be contrasted with Corollary 4).

Theorem 5. *Let $w = (0011)^{n/4}$. Then*

(i) $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \geq \frac{1}{8} n \ln(n)$.

(ii) *the number of all maximal repetitions of w is $\Theta(n^2)$,*

Proof. (i) The whole word w is an obvious repetition of period 4, its contribution to the sum is $(n/4 - 1)$. Any other repetition can be specified by a match between two 0's or two 1's that occur at a distance other than a multiple of 4.

Consider a repetition r in which letter 0 at some position m, $m \equiv 1 \pmod 4$, matches letter 0 at a position $\ell > m$, $\ell \equiv 2 \pmod 4$. This match corresponds to end letters of the repetition, as $w[m - 1] = 1$ (if $m \neq 1$) while $w[\ell - 1] = 0$, and $w[\ell + 1] = 1$ (if $\ell \neq n$) while $w[m + 1] = 0$.

Furthermore, this repetition has period $\ell - m = |r| - 1$ and this period is minimal, as word $w[m..\ell - 1]$ contains one more 0 than 1's and therefore the number of 0's and the number of 1's in $w[m..\ell - 1]$ are mutually prime, which shows that $w[m..\ell - 1]$ is primitive (i.e. not an integer power of some other word).

Therefore, any two such positions m and ℓ define a repetition that contributes $1/(\ell - m)$ to the sum. In total, all such repetitions contribute $\sum_{i=1}^{n/4} (n/4 - i + 1)/(4i - 3) \geq \frac{1}{32} n \ln(n)$.

There are three other symmetric cases: one corresponds to another way of matching two 0's and the other two correspond to matching two 1's. The four cases together yield $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \geq (\frac{n}{4} - 1) + 4\frac{1}{32} n \ln(n) \geq \frac{1}{8} n \ln(n)$.

(ii) is obvious from the above, as the number of pairs of 0's and pairs of 1's defining repetitions is quadratic. □

We now focus on the lower bound for sum (1). In the rest of the paper, we assume that we have a k-letter alphabet $A_k = \{a_1, a_2, \ldots, a_k\}$.

Theorem 6. *For all $w \in (A_k)^*$, $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \geq \frac{n}{k} - 1$ and the equality holds if and only if $w = (a_1 a_2 \ldots a_k)^{\frac{n}{k}}$ (modulo a permutation of alphabet letters).*

Proof. Given a word $w \in (A_k)^*$, consider all occurrences of a letter $a_i \in A_k$ in w, and let $d_1^i, d_2^i, \ldots, d_{\ell_i}^i$ be the distances between all consecutive occurrences of a_i in w. Consider the sum

$$\sum_{a_i \in A_k} \sum_{j=1}^{\ell_i} \frac{1}{d_j^i}. \tag{2}$$

4

Observe that $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \geq \sum_{a_i \in A_k} \sum_{j=1}^{\ell_i} \frac{1}{d_j^i}$ since two consecutive occurrences of a_i necessarily participate in a repetition with period equal to the distance between these occurrences, and then contribute to sum (1) (see proof of Theorem 1).

Therefore, if we construct a word that minimizes sum (2) and for which $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) = \sum_{a_i \in A_k} \sum_{j=1}^{\ell_i} \frac{1}{d_j^i}$, this will prove that this word also minimizes sum (1). Our goal is to prove that this minimum is reached if and only if for any letter a_i, all $d_j^i = k$, i.e. on words of the form $w = (a_1 a_2 \dots a_k)^{\frac{n}{k}}$ (modulo a permutation of alphabet letters). Clearly, for such words, sum (1) and sum (2) are both equal to $\frac{n}{k} - 1$.

By contradiction, consider a word w that does not have the form $(a_1 a_2 \dots a_k)^{\frac{n}{k}}$ and assume that it minimizes sum (2). Then there exists a pair of positions $m_\ell < m_r$ such that $w[m_\ell] = w[m_r]$ and $m_r - m_\ell < k$. Among all such pairs, consider the one with minimal m_r.

Show that for any position m, $k < m < m_r$, we must have $w[m] = w[m-k]$. This is because letter $w[m]$ cannot repeat on the left at a distance smaller than k, as this would contradict the definition of m_r. On the other hand, the closest occurrence of $w[m]$ to the left cannot be at a distance larger than k either. Indeed, if $w[m] = w[m']$ for some $m' < m$ and $m - m' > k$ and there is no occurrence of $w[m]$ in $w[m'+1..m-1]$, then subword $w[m'+1..m'+k]$ is composed of $k-1$ letters and has length k, and therefore contains a letter repeated at a distance at most $k-1$. This contradicts again the definition of m_r.

By the above, we can assume that $w[1..m_r - 1] = (a_1..a_k)^q a_1..a_i$ (up to a permutation of alphabet letters), $q \geq 1$, and $w[m_r] = a_j$ for some $j \neq i'$ where $i' = i+1$ if $i < k$ and $i' = 1$ if $i = k$. Consider the closest position of $a_{i'}$ to the right of m_r, that we denote m'. (If such a position does not exist, the proof below will trivially apply.)

We modify w by simultaneously

- replacing all occurrences of a_j at positions $\geq m_r$ by $a_{i'}$, and

- replacing all occurrences of $a_{i'}$ at positions $\geq m'$ by a_j.

We show that this modification makes sum (2) smaller.

The only distances between consecutive occurrences of letters that will be affected by the modification of w are the distance $m_r - m_\ell$ between the corresponding occurrences of a_j and the distance $m' - (m_r - k)$ between the occurrences of $a_{i'}$. The new distances become respectively k (between occurrences m_r and $m_r - k$ of $a_{i'}$) and $m' - m_l$ (between corresponding occurrences of a_j). We show that

$$\frac{1}{m_r - m_\ell} + \frac{1}{m' - (m_r - k)} > \frac{1}{k} + \frac{1}{m' - m_l}.$$

This will show that sum (2) becomes smaller after the modification. For this, we show that

$$\frac{1}{m_r - m_\ell} - \frac{1}{k} > \frac{1}{m' - m_l} + \frac{1}{m' - (m_r - k)},$$

5

or

$$\frac{k - m_r + m_\ell}{(m_r - m_\ell)k} > \frac{k - m_r + m_\ell}{(m' - m_l)(m' - (m_r - k))}.$$

The numerators of both sides are equal. In denominator, we have $m' - m_l > m_r - m_\ell$ and $m' - (m_r - k) > k$, which proves the inequality.

We obtained a contradiction with the assumption that w minimizes sum (2). This shows that a word that minimizes sum (2) must have the form $w = (a_1 a_2 \ldots a_k)^{\frac{n}{k}}$ (modulo a permutation of alphabet letters). On this word, sum (1) and sum (2) are both equal $\frac{n}{k} - 1$. This proves that w also minimizes sum (1). □

4   Words with repetitions of bounded period

In this section, we study sum (1) in the case when all repetitions in w are of period at most p. Recall that k is the alphabet size.

Theorem 7. *Let the period of all repetitions of a word w ($|w| = n$) be bounded by p. Then $\sum_{r \in \mathcal{M}(w)} (e(r) - 1) \leq n + 3kp(\ln(p) + 1)$.*

The proof will use the Fine and Wilf's theorem (see e.g. [Lot83]) asserting that if w have (not necessarily minimal) periods p_1 and p_2 and $|w| \geq p_1 + p_2 - gcd(p_1, p_2)$, then w has also the period $gcd(p_1, p_2)$. This implies, in particular, that two different repetitions with minimal periods p_1 and p_2 cannot intersect on $(p_1 + p_2)$ letters or more.

Proof. Consider a word w such that the period of any repetition in w is bounded by p.

Assume that for some letter a, two occurrences of a are located at a distance $3p$ or more. Consider a repetition r defined by the match of these two occurrences of a. We will show that r has a very particular form, namely

(a) all letters within a root of r are different,

(b) any letter of r does not occur outside r.

Figure 1: Proof of condition (a) of Theorem 7

First observe that since the period of r cannot exceed p, then the two occurrences of a are separated by at least three periods $p(r)$. To prove (a), assume that there is another occurrence of a in the suffix root of r (cf Figure 1). Then, there is a repetition r' formed by matching this occurrence of a with the left

6

occurrence of a. These two occurrences are separated by $3p - p(r) \geq 2p$ letters. Consider $p(r')$. Since $p(r') \leq p$, there are at least $2p(r')$ letters between these two occurrences of a. This means that repetitions r and r' intersect by length at least $2 \cdot \max\{p(r), p(r')\}$ and by Fine and Wilf's theorem, r and r' must coincide. This contradiction proves that a cannot have another occurrence within a root of r. More generally, the same argument shows that any letter occurs in a root only once.

Condition (b) is proved by a similar argument. Assume that some letter b of r occurs outside r, for instance to the right of r. Then consider the match of this occurrence of b with the leftmost occurrence of b inside r. This match defines a repetition r'. Similar to part (a), r and r' intersect by length at least $2 \cdot \max\{p(r), p(r')\}$ and therefore must coincide by Fine and Wilf's theorem. This contradicts to the assumption that of an occurrence of b outside r and proves (b).

Now, we split all repetitions into two disjoint classes: repetitions verifying conditions (a) and (b) and the others, called respectively repetitions of type 1 and repetitions of type 2. By condition (b), for any word w, repetitions of type 1 and type 2 in w are non-intersecting. Furthermore, conditions (a) and (b) insure that two distinct repetitions of type 1 cannot intersect. Therefore, all repetitions of type 1 together cannot contribute more than n to the sum.

On the other hand, repetitions of type 2 cannot take more than $3kp$ letters altogether in w, as each letter cannot occur more than $3p$ times as this would lead to a repetition of type 1 by the above reasoning. Therefore, by Corollary 2, sum (1) for repetitions of type 2 is bounded by $3kp(\ln(p) + 1)$. This gives the final bound $n + 3kp(\ln(p) + 1)$. $\qquad\square$

Notice that the bound in Theorem 7 is optimal in some sense, since sum (1) is $n-1$ for the word a^n and $\Theta(kp \ln(p))$ for the word $(a_1 a_1 a_2 a_2)^{p/4} (a_3 a_3 a_4 a_4)^{p/4} \ldots$, according to Theorem 5.

5 Concluding remarks

Many questions related to the combinatorics of repetitions of arbitrary exponent remain unanswered. A major such question is the precise bound on the number of such repetitions. Corollary 4 provides an $O(n \log n)$ bound for the exponents at least $(1 + \varepsilon)$, for any fixed $\varepsilon > 0$. It would be of great interest to refine this bound, possibly depending on ε. It is not excluded that, possibly starting from some $\varepsilon > 0$, or even for any fixed $\varepsilon > 0$, the number of all repetitions of exponent at least $(1 + \varepsilon)$ is $O(n)$. This is a challenging question, that seems, however, difficult to solve, as it would generalize the result of [KK99, KK00] on the linear number of runs.

References

[CIT08] M. Crochemore, L. Ilie, and L. Tinta. Towards a solution to the "runs" conjecture. In Paolo Ferragina and Gad M. Landau, editors, *Combina-*

torial Pattern Matching, 19th Annual Symposium, CPM 2008, Pisa, Italy, June 18-20, 2008, Proceedings, volume 5029 of *Lecture Notes in Computer Science*, pages 290–302. Springer, 2008.

[CK97] C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook on Formal Languages*, volume I, pages 329–438. Springer Verlag, Berlin-Heidelberg-New York, 1997.

[Dej72] F. Dejean. Sur un théorème de Thue. *J. Combinatorial Th. (A)*, 13:90–99, 1972.

[KK99] R. Kolpakov and G. Kucherov. On maximal repetitions in words. In *Proceedings of the 12-th International Symposium on Fundamentals of Computation Theory, 1999, Iasi (Romania)*, Lecture Notes in Computer Science, pages 374 – 385. Springer Verlag, August 30 - September 3 1999.

[KK00] R. Kolpakov and G. Kucherov. On maximal repetitions in words. *Journal of Discrete Algorithms*, 1(1):159–186, 2000.

[KK05] R. Kolpakov and G. Kucherov. Identification of periodic structures in words. In J. Berstel and D. Perrin, editors, *Applied combinatorics on words*, volume Encyclopedia of Mathematics and its Applications, vol. 104 of *Lothaire books*, chapter 8, pages 430–477. Cambridge University Press, 2005.

[KS04] J. Karhumäki and J. Shallit. Polynomial versus exponential growth in repetition-free binary words. *Journal of Combinatorial Theory, Series A*, 105:335–347, 2004.

[Lot83] M. Lothaire. *Combinatorics on Words*, volume 17 of *Encyclopedia of Mathematics and Its Applications*. Addison Wesley, 1983.