Clusters of repetition roots forming prefix chains

Szilárd Zsolt Fazekas[1⋆] and Robert Mercaş[2]

[1] Akita University, Graduate School of Engineering Science, Japan
szilard.fazekas@ie.akita-u.ac.jp
[2] Loughborough University, Department of Computer Science, UK
R.G.Mercas@lboro.ac.uk

Abstract. We investigate lower bounds on the size of clusters (sets of starting positions of occurrences) of common prefixes shared by repetition roots. Such lower bounds in terms of the constituent roots in the sets provide upper bounds on the number of distinct repetitions. In the case of distinct square roots which are totally ordered by the prefix relation it has been shown that there must be more occurrences of the common prefix than the number of roots. Here we develop the theory further by presenting the tools to extend the bounds to exponents higher than 2 and we show that they are optimal in the sense that any sequence of cluster sizes satisfying the lower bounds can be realized. We also take the next step towards the bounds on arbitrary (only partially prefix-ordered) sets of roots by proving a lower bound on unbordered prefixes shared by two overlapping prefix chains of roots.

1 Introduction

Repetitions in words are one of the most studied topic in word combinatorics [17], partly due to their various applications in string matching [5], molecular biology [11], or text compression [19]. The most basic repetition is xx, where x is a non-empty string. Such strings are also called, due to the form $xx = x^2$, *squares*.

A string is said to be square-free or repetition-free if it contains no squares. Combinatorics on words arguably started with the work of Thue [21, 22] who showed that there exist square-free strings over ternary alphabets and cube-free ones over two letters. Over two letters, trivially every string of length at least 4 contains a square and it has also been shown that any sufficiently long binary string must contain at least three *distinct squares* [9].

A string of length n can have $\Theta(n^2)$ squares (just take a unary sequence). If the root x of each square xx must be primitive (not a repetition), one can still have at most $\Theta(n \log n)$ squares [5]. When the roots of the squares must be distinct, then the maximal number becomes linear in the length of the string. Fraenkel and Simpson proved [10] that the maximum number of distinct squares in a string is not more than twice the length of the string and they conjectured that the bound can be significantly improved:

⋆ This Work Was Supported By JSPS KAKENHI Grant Number JP19K11815.

Conjecture 1. The number of distinct squares in a length n word is less than n.

They also constructed lower bounds which asymptotically match the conjectured upper bound except for a sublinear term. We will use another simple lower bound construction by Jonoska, Manea and Seki [15] as our starting point for discussing optimality later on.

There have been several developments in the last 25 years on the topic. Some alternative and simple proofs of the $2n$ upper bound were found [13, 12], after which the bound was improved to $2n-\Theta(\log n)$ [14]. Deza, Franek and Thierry [6] proved the best (peer-reviewed) bound as of now, $11n/6$, by a deep investigation of left aligned last occurrences of distinct squares. There was a claim of further improvement to $3n/2$ very recently[20], but it has not appeared in peer-reviewed publication to the best of our knowledge.

Regarding exponents larger than 2 it was shown [3] that for fixed integers $\ell > 2$, there can be no more than $\frac{n}{\ell-2}$ powers of exponent ℓ in a word of length n. For cubes, that is, $\ell = 3$ the bound was improved to $4n/5$ [4]. The study of repetitions of higher fixed exponents was inspired by the importance of counting runs, i.e., repetitions whose exponent is at least 2 and which cannot be extended in either direction without increasing the period. The bound on this number was conjectured to be less than the word's length [16] (not much after Fraenkel and Simpson's square conjecture was published) and recently proved to be so by a very elegant and simple argument [1].

There were other developments relevant to the question even though they did not necessarily improve upper bounds. By using square density increasing mappings it was shown that binary strings can achieve maximum density if the conjectured upper bound holds [18]. In the case of partial words (strings with holes) tight upper bounds have been proved depending on the number of holes [2]. Another recent paper [8] proposed a framework to integrate existing results and facilitate new ones in the analysis of distinguished positions of squares.

§§ **Our contribution.** Finally, the basis of our current work proposed another angle of attack using clusters of repetition roots [7]. The techniques used there extract global properties of occurrences of repetitions in a word from local ones and we continue that line of investigation. We group the repetitions by the partial order imposed on their roots by the prefix ordering. All repetitions whose roots share a common prefix are in one group and our aim is to show that there are 'many' occurrences of this common prefix forced by the occurrences of the repetitions. We are working toward proving the conjecture on the lower bound on the number of those prefixes which would imply Fraenkel and Simpson's. We will introduce notation and the line of attack in the next section. In Section 3 we generalize the lower bound technique used recently for prefix chains of squares, to the case of higher exponents. More specifically, we show that if two ℓ-powers are aligned at the end of their second or further root occurrence, then the shorter root must be non-primitive. In our pervious work this was used to assign unique positions to primitively rooted squares, followed by a different assignment procedure for non-primitively rooted ones, so it forms the basis of

lower bound results for prefix chains of repetition roots. Afterwards we discuss the optimality of the bounds obtained for squares. As opposed to the other bounds mentioned in the introduction, ours are tight in the sense that for each sequence of cluster sizes satisfying the bounds we can find a word and repetition roots in it which have those exact cluster sizes. We present a simple construction to achieve those bounds. We also show that a counting argument of similar flavor can be applied to runs whose suffixes of length equal to the run's period form a prefix chain. In Section 4 we develop the technique further by designating special occurrences of a shared unbordered prefix of roots in two overlapping prefix chains. The main result in that section is a counterpart of the theorem in Section 3: alignment of repetitions at their suitably defined anchor means that the shorter one is non-primitive. The challenge is that the anchor has to be defined differently in the case of root sets which are not linearly ordered by the prefix relation. We present a solution in the case when such a set is the union of two prefix chains with minimal elements that are unbordered.

2 Preliminaries

A *word* or *string* is a concatenation of letters from a *finite alphabet* Σ. The *empty word* ε is the word of length 0. For a word $w = xyz$, we call x a *prefix* (denoted by $x \leq_p w$, or $x <_p w$ if $x \neq w$) and z a *suffix* of w, while each of x, y, z are called *factors* of w. The word y is an inside factor of w if neither x nor z are empty. A factor is *proper* if it is non-empty and not equal to w. If $x = z$, then x is also a *border* of w. If two words u and v are not comparable by the prefix relation, we write $u <>_p v$. The longest common prefix of two words $u = xau'$ and $v = xbv'$ is $\mathbf{lcp}(u, v) = x$ if either au' or bv' is empty or otherwise $a \neq b$.

We call p a *period* of w if the letters repeat every p positions apart in w. The *minimal period* is given by the smallest such p. By $|w|_x$ we denote the number of times x occurs as a factor of w (including overlaps).

A *repetition* represents consecutive concatenations of the same word. An ℓ-*power* (ℓ-*repetition*) represents ℓ such repetitions of the same factor. If a word is not a repetition, then it is called *primitive*. Moreover, if $w = u^\ell$ is an ℓ-repetition we say that u is a *root* of w, and call u *the primitive root* of w when u is primitive.

For a word u and a prefix u' of u, all words $u^\ell u'$ with integer exponent $\ell \geq 0$ have period $|u|$. A word can have multiple periods, e.g., *ababa* has periods 2 and 4, since $ababa = (ab)^2 a = (abab)^1 a$. While repetitions are defined in terms of integer powers, rational powers are also possible. Namely, $u = t^k$ for some rational k, if $|u| = k|t|$ and $|t|$ is a period of u. For instance, the word *abcabca* is a fractional power of *abc* since $abcabca = (abc)^{\frac{7}{3}}$. A *run* is given by the positions in the word that contain a maximal repetitive factor with period at most half as long as the length of the factor (a repetition is maximal, if taking a previous or following position changes the period). In other words, a run is a factor that has an exponent at least 2, and which cannot be extended to either left or right. Finally, by t^ω we denote the infinite word consisting in consecutive repetitions of t.

We also recall the following well-known results about primitivity of words and multiple preiods.

Lemma 1. *[17] A word w is primitive if and only if it occurs only twice in ww.*

Theorem 1 (Fine and Wilf). *[17] If a word w has periods p and q and $|w| \geq p + q - \gcd(p, q)$, then $\gcd(p, q)$ is also a period of w.*

2.1 Clusters of repetition roots

In this subsection we introduce clusters of repetition roots and explain the conjecture which is the final goal of our study.

When wanting to count all distinct ℓ-powers for a fixed ℓ, we denote by $\mathbf{clust}_w(u)$, for each factor u^ℓ of w, the set that contains the starting position of all suffixes having u as a prefix. We will call this set the *cluster* of u. Clearly, if an ℓ-repetition u^ℓ is a factor of a word, then the cluster of u is of size at least ℓ. As every word, and therefore every suffix starting with v also has u as prefix when $u <_p v$, the next observation is straightforward.

Observation 1 *For any two factors u and v of a word w, we have $u \leq_p v \Leftrightarrow \mathbf{clust}_w(v) \subseteq \mathbf{clust}_w(u) \Leftrightarrow \mathbf{clust}_w(u) \cap \mathbf{clust}_w(v) \neq \emptyset$ and $|u| \leq |v|$.*

In this paper we attempt to get closer to the following conjecture, which, if true, would give a general upper bound for integer exponent distinct repetitions:

Conjecture 2. [7] For any word w, any positive integer $\ell > 1$, and any set of words $S = \{u_1, u_2, \ldots, u_n\}$ such that, for all $i \in \{1, \ldots, n\}$, u_i^ℓ is a factor of w and $u_1 \leq_p u_i$, we have $|S| < \frac{1}{\ell - 1}|w|_{u_1}$.

In the paper proposing the conjecture, it was proved for the case where $\ell = 2$ and $u_1 \leq_p \cdots \leq_p u_n$, that is, S is a set of roots of distinct squares, totally ordered by the prefix relation. Such a collection of square roots is called a *(prefix) chain* and with that, the result can be stated as

Theorem 2. *[7] For a word w and a prefix chain $S = \{u_1, u_2, \ldots, u_n\}$ of square roots of w, with $u_i \leq_p u_{i+1}$ for all $i \in \{1, \ldots, n-1\}$, we have $|S| < |w|_{u_1}$.*

In the next section we generalize the results necessary to prove Conjecture 2 for prefix chains of roots in the case of repetitions of arbitrary exponents. Due to the page limit we do not present the reassignment procedure, which allocates distinct positions to the non-primitively rooted repetitions. Compared to the results in [10, 14, 15, 6], the bound in Theorem 2 is different because it is in a sense optimal, as we will argue at the end of Section 3. Furthermore, while the bounds on distinct repetitions would be direct corollaries of Conjecture 2, the converse does not hold.

3 Single chains

In this section we show that the non-primitivity conditions on the roots of colliding powers used to prove Conjecture 2 in the special case of single chains of square roots, are valid for arbitrary exponent K. These are conceptually simple proofs following the argument of their counterparts for squares (Lemma 5 and Corollary 2 in [7]). Afterwards we discuss the optimality of the bound w.r.t. the existence of words $w, u_1, \ldots, u_n$ for every possibility of cluster sizes satisfying the bound. Finally we show that prefix chains of square roots at the end of runs can help find alternative techniques for counting maximal repetitions, too.

For a prefix $x \leq_p u$ and natural number $\ell \in \{2, \ldots, K\}$, we say that the (ℓ, x)-representative ((ℓ, x)-rep) of u^K is the longest prefix of u^ℓ which ends in x. Note that this x-rep is of length at least $(\ell-1)|u| + |x|$. Formally, the (ℓ, x)-rep of u^K is $u^{\ell-1}u'x \leq_p u^K$ such that for all y we have that $u^{\ell-1}yx \leq_p u^\ell$ implies $|y| \leq |u'|$.

Let w be a word which contains u^K as a factor. For the leftmost occurrence in w of the (ℓ, x)-rep $u^{\ell-1}u'x$ of the K-power u^K, let u_s be its starting position and $u_m = u_s + (\ell-1)|u|$.

We define the (ℓ, x)-*anchor* of u^K in w as the starting position of the rightmost occurrence of x in the first occurrence of the (ℓ, x)-rep of the power u^K in w. This (ℓ, x)-anchor is denoted by $\mathbf{\Psi}_w(u^\ell, x)$. If the (ℓ, x)-rep of u^K is $u^{\ell-1}u'x$, then $\mathbf{\Psi}_w(u^\ell, x) = u_s + (\ell-1)|u| + |u'|$.

For example, in the word w below

a	b	a	a	b	c	a	b	a	a	b	a	a	b	c	a	b	a	a	b	a	a	b	a
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

we have the cube (3-power) $u = (aba)^3$ starting at position 16. The $(2, \boldsymbol{a})$-rep of u^3 is $abaab\boldsymbol{a} = u^2$, first occurring at 7, so $\mathbf{\Psi}_w(u^2, a) = 7 + |abaab| = 12$. The $(2, \boldsymbol{ab})$-rep of u^3 is $aba\boldsymbol{ab}$, first occurring at 1, therefore $\mathbf{\Psi}_w(u^2, ab) = 1 + |aba| = 4$. The $(3, \boldsymbol{ab})$-rep of u^3 is $(aba)^2\boldsymbol{ab}$ whose only occurrence is at 16, meaning that $\mathbf{\Psi}_w(u^3, ab) = 16 + |(aba)^2| = 22$.

While the (ℓ, x) anchors are not exactly at the right edge of the repetitions u^ℓ, as we will see, when two repetitions are aligned by their anchors it has a similar consequence as if they were aligned at their right edge: the shorter one is non-primitive. We show that this is true for all pairs of coinciding anchors.

Lemma 2. *Let w be an arbitrary word with two K-powers u^K, v^K such that $u <_p v$, and let x be a common prefix of u and v. If there are $\ell, \ell' \in \{2, \ldots, K\}$ such that $\mathbf{\Psi}_w(u^\ell, x) = \mathbf{\Psi}_w(v^{\ell'}, x)$, then $u = t^k$ for some primitive word t with $|t| < |x|$ and $k \geq 2$. Moreover, $tu'x \leq_p v$, where $u'x$ is the longest prefix of u bordered by x.*

Proof. Assume $\mathbf{\Psi}_w(u^\ell, x) = \mathbf{\Psi}_w(v^{\ell'}, x)$. We distinguish three cases based on the relative positions of u_s, u_m and v_m, and will derive contradictions in all of them, except in the last case, when u is non-primitive with its root shorter than

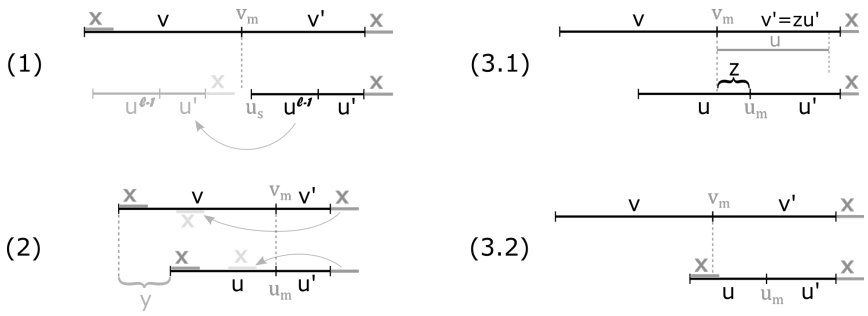

Fig. 1. The cases analyzed in Lemma 2.

x. Note that $v_m \leq u_m$ always holds, since $u \leq_p v$ implies $\mathbf{\Psi}_w(u^\ell, x) - u_m \leq \mathbf{\Psi}_w(v^{\ell'}, x) - v_m$. In what follows, let the (ℓ, x)-rep of u^K be $u^{\ell-1}u'x$.
(1) $v_m \leq u_s$, see Fig. 1(1). In this case the (ℓ, x)-rep of u^K is a factor of v, therefore it also occurs at $u_s - |v|$, a contradiction.
(2) $v_m = u_m$, see Fig 1(2). This means that u is a suffix of v and since $|v| > |u|$, we have $v = yu$, for some non-empty word y. From $\mathbf{\Psi}_w(u^\ell, x) = \mathbf{\Psi}_w(v^{\ell'}, x)$, we get that the x-rep of v^K is $v^{\ell'-1}u'x$. However, $yu'x \leq_p v$ which means that the rightmost x occurrence in v is at least $|yu'|$ positions from its start, so

$$\mathbf{\Psi}_w(v^{\ell'}, x) \geq v_s + (\ell' - 1)|v| + |yu'| > v_s + (\ell' - 1)|v| + |u'| = \mathbf{\Psi}_w(v^{\ell'}, x),$$

a contradiction.
(3) $u_s < v_m < u_m$. Let the (ℓ', x)-rep of v^K be $v^{\ell'-1}zu'x$, where z is the non-empty word starting at v_m and ending at $u_m - 1$. Both $zu'x$ and u are prefixes of v, so they are prefixes of each other. If $zu'x \leq_p u$, then

$$\mathbf{\Psi}_w(u^\ell, x) \geq u_s + (\ell - 1)|u| + |zu'| > u_s + (\ell - 1)|u| + |u'| = \mathbf{\Psi}_w(u^\ell, x),$$

which is a contradiction. The only remaining possibility is if $u \leq_p zu'x$. Then, there is an occurrence of u at v_m and by Lemma 1 this means that u is not primitive. (see Fig. 1(3.1)).

Now let $u = t^k$, with t primitive and $k \geq 2$. If $|uu'| \geq |v|$ then a conjugate of v is a prefix of uu', because uu' is a factor of v^2. From here, v has period $|t|$ and the fact that t^k is its prefix and t is its suffix means that $v = t^m$ for some $m > k$. This, in turn, means that u^ℓ and hence the (ℓ, x)-rep of u^K occurs at position v_s, so the occurrence at u_s is not the leftmost, another contradiction.

We are left with the case $|uu'| < |v|$. We have an occurrence of x at $u_m - |u|$. If that x finishes before position v_m, that is, $v_m - |x| \geq u_m - |u|$, then there should be an occurrence of x located $|v|$ positions further to the right in v^K. That would give $\mathbf{\Psi}_w(v^{\ell'}, x) \geq u_m - |u| + |v| > u_m - |u| + |uu'| = \mathbf{\Psi}_w(u^\ell, x)$, contradicting $\mathbf{\Psi}_w(v^{\ell'}, x) = \mathbf{\Psi}_w(u^\ell, x)$. Hence, we get that $v_m - (u_m - |u|) < |x|$, which means $|t| < |x|$. (see Fig. 1(3.2)).

As x is a prefix of $u = t^k$, it has the form $x = t^r t'$ for some $r < k$ and $t' \leq_p t$. The longest prefix of $u = t^k$ bordered by x is $t^{k-1}t' = u'x$. As $u_m > v_m$, we get that $tt^{k-1}t' = tu'x \leq_p v$. □

Corollary 1. *Let $u_1^K, \ldots, u_n^K$ and $v_1^K, \ldots, v_n^K$ be powers in some word w with their roots all from the same chain and let x be a common prefix of those roots, such that for all $i \in \{1, \ldots, n\}$ there are $\ell_i, \ell'_i \in \{2, \ldots, K\}$ with $\mathbf{\Psi}_w(u_i^{\ell_i}, x) = \mathbf{\Psi}_w(v_i^{\ell'_i}, x)$. Then, there exists some primitive word t shorter than x, such that $u_i = t^{k_i}$ with $k_i \geq 2$, for all $i \in \{1, \ldots, n\}$.*

Proof. From Lemma 2, whenever the (ℓ_i, x)-anchor of u_i^K and the (ℓ'_i, x)-anchor of v_i^K coincide, there is some primitive t_i with $|t_i| < |x|$ such that $u_i = t_i^{k_i}$ with $k_i \geq 2$ and $t_i x$ is a prefix of v_i. Given that the roots of these powers form a prefix chain, we get that the words $t_i x$ also form a prefix chain, that is, for all $i, j \in \{1, \ldots, n\}$ either $t_i x \leq_p t_j x$ or $t_j x \leq_p t_i x$. Furthermore, since x is a common prefix of all the powers, we have $x \leq_p t_i x$, so x has period $|t_i|$, and therefore, trivially, so does $t_i x$. For any pair t_i, t_j, with $|t_i| \leq |t_j|$, we know that $t_i x \leq_p t_j x$, so $t_i x$ also has period $|t_j|$. Since $|t_i x| > |t_i| + |t_j| > |t_i| + |t_j| - \gcd(|t_i|, |t_j|)$, we can apply Theorem 1 and get that t_i and t_j have a common primitive root t. We already know that t_i and t_j are primitive, so $t_i = t_j = t$. □

Not surprisingly, the same anchor assignment procedure does not produce the desired conclusion if we apply it to powers whose roots are not linearly ordered by the prefix relation. The reason is that what we exploit in the proofs above is that aligning the right edge of powers whose roots are prefixes of each other results in (at least the shorter one of) them being non-primitive. However, right-aligning powers which merely share some prefix does not provide the same strict conclusion.

An alternative way of anchoring which might work for powers in two prefix chains with an overlapping part is to assign the symmetric difference of the chains by their longest common prefix, and anchoring the intersection by the shortest root as before. Further on we show a scheme which works in a special case when the shortest root is unbordered. Before moving on to that, however, we first discuss the sharpness of the bounds implied by our conjecture.

3.1 Optimality.

Consider a chain of square roots $u_1 <_p \cdots <_p u_n$ as before. From Theorem 2 we already know that $|\mathbf{clust}(u_i)| \geq n - i + 2$, for all $i \in \{1, \ldots, n\}$, and trivially, $|\mathbf{clust}(u_{i-1})| \geq |\mathbf{clust}(u_i)|$, but it is natural to ask whether the bounds are optimal, that is, whether all possible combinations of cluster sizes satisfying those conditions can actually be realized in some string w. Using the lower bound construction in [15], we can easily illustrate the extremal cases. We are only interested in the situations where $|\mathbf{clust}_w(u_1)| = n + 1$ because we can trivially add further occurrences of all roots at the end of w to accomodate the other cases. When $|\mathbf{clust}_w(u_i)| = n - i + 2$, that is, the topmost cluster has size 2

and then each subsequent cluster is one larger than the previous, take $u_i = ab^{i-1}$ and the word $w = u_1u_2 \cdots u_nu_n$. The case $|\mathbf{clust}_w(u_1)| = |\mathbf{clust}_w(u_n)| = n+1$ is realized by the roots $u_i = a^{n-1}ba^{i-1}$ and again a word of the form $u_1u_2 \cdots u_nu_n$.

From this starting point we can develop an algorithm to realize any combination of cluster sizes. The idea is to start from the case when all clusters are equal and then reduce the relevant clusters by adding further as to the end of their roots. We start out with $u_i = a^{n-1}ba^{i-1}$ as before and the word in which we realize the clusters will be the concatenation of all the u_i. At this point all clusters are equal to $n+1$ and we set $r_i = i-1$ for all $i \in \{1, \ldots, n\}$. We will refine iteratively the values r_i and in the end will set $u_i = a^{n-1}ba^{r_i}$. To remove the occurrence immediately preceding the k-th b from the clusters of each root u_i with $i \geq j$ we add $r_k + n - r_j$ many as to each such r_i. After updating the r_i in question, we keep repeating the removal as many times as necessary. To see whether the construction is correct, note that increasing r_i does not affect the cluster of any of the u_j with $j < i$. By adding $r_k + n - r_j$ to r_j we get that the unary a-tail of u_j is of length $r_k + n$ which is more than the distance between the k-th and $(k+1)$-th b in the word, removing all occurrences of u_j (and hence all longer roots, as well) starting before the k-th b.

For example, let the clusters of $u_1, \ldots, u_6$ be of length $7, 7, 5, 5, 3, 3$, respectively. This means $n = 6$, so initially we set $u_1 = a^5b$, $u_2 = a^5ba$, $u_3 = a^5ba^2$, $u_4 = a^5ba^3$, $u_5 = a^5ba^4$ and $u_6 = a^5ba^5$ and $r_i = i - 1$. First we need to remove the first occurrence of the clusters of $u_3, \ldots, u_6$, so we get $k = 1$ and $j = 3$. This means adding $r_k + n - r_j = 0 + 6 - 2 = 4$ to each r_i with $i \geq 3$. Now the r_i values are $0, 1, 6, 7, 8, 9$. Next we need to remove the occurrences preceding the second b from the same cluster so $k = 2$ and $j = 3$, and hence we need to add $r_k + n - r_j = 1 + 6 - 6 = 1$ to each of those r values, resulting in $0, 1, 7, 8, 9, 10$. Removing the next two occurrences, $k = 3$ and 4, respectively, from the clusters of u_5 and u_6 is by first adding $7 + 6 - 9 = 4$ to them and then $8 + 6 - 13 = 1$, respectively. The end result is $0, 1, 7, 8, 14, 15$, so the clusters are realized by the occurrences of $u_1 = a^5b$, $u_2 = a^5ba$, $u_3 = a^5ba^7$, $u_4 = a^5ba^8$, $u_5 = a^5ba^{14}$ and $u_6 = a^5ba^{15}$ in the word $u_1 \cdots u_6u_6$.

This construction is not optimal in the sense that in most cases there exist much shorter words w and $u_1, \ldots, u_n$ which have a chain of clusters satisfying the same conditions. We expect that investigating the shortest words which realize a combination of cluster sizes could lead to improvements in both lower and upper bounds on distinct repetitions.

3.2 Single chains of run ending squares

A related direction for expanding the theory of clusters is to find a proof of the upper bound on runs in terms of clusters. We present a brief argument for a simple bound for runs whose "ending squares" form a prefix chain. We cannot readily apply the technique used for distinct squares, because here multiple occurrences of a repetition have to be taken into account.

Consider a run $(a_1 \cdots a_n)^{\frac{k}{n}}$ in a word w, where $a_i \in \Sigma$, $k \geq 2n$, and $a_1 \cdots a_n$ is primitive. Let this run begin at some position i in w. The run ending square is

the square starting at position $i+k-2n$ and ending at $i+k-1$. For example, if $w = aababaa$ and we consider the run $(ab)^{\frac{5}{2}}$ starting at $i = 2$ in w, then the run ending square is $baba$, which starts at $i+k-2n = 3$ and ends at $i+k-1 = 6$.

Each run has a run ending square, so an upper bound on their number is implicitly an upper bound on the number of runs. The crucial property of run ending squares uu is that the letter following uu in the word is different from the first letter of u. Consider roots of run ending squares $u <_p v \in \Sigma^*$, with a being their first letter. Although uu may occur followed by a, but in those cases it is not the suffix of a run with period $|u|$. An occurrence of uu in w is a run ending square if it is followed by some $b \neq a$ or if it is a suffix of w. Let the run ending occurrences of u^2 start at positions $i_1 < \cdots < i_k$. This means that $\{i_1, i_1+|u|, \ldots, i_k, i_k+|u|\} \subseteq \mathbf{clust}(u)$. However, for all $j \in \{1, \ldots, k-1\}$ we have $w[i_j+|u|] \neq w[i_j+2\cdot|u|]$, and $w[i_k+|u|] \neq w[i_k+2\cdot|u|]$ or $i_k+2\cdot|u| = |w|+1$. From here, for each $j \in \{1, \ldots, k\}$, at least one of the two positions i_j and $i_j+|u|$ is not in $\mathbf{clust}(v)$, so $|\mathbf{clust}(u)|-|\mathbf{clust}(v)| \geq k$. Applying this argument to consecutive roots in a prefix chain $u_1 <_p \cdots <_p u_n$, we get that $\mathbf{clust}(u_i)$ is larger than the number of all runs with run ending square u_j^2, $j \geq i$. However, similarly to the case of distinct powers, this argument does not extend easily to overlapping chains of run ending squares, so one either has to define roots differently for a run or figure out how to treat the case of run ending squares u^2, v^2, w^2 where u is a common prefix of v and w, but the latter two are incomparable.

4 Two overlapping chains

Using the anchor positions seen before one can prove the hypothesis for single chains in the general case. As a first extension of the bounds to multiple chains, we will prove a special case when two overlapping chains share an unbordered prefix, in terms of whose occurrences we can upper-bound the number of distinct roots in the two chains combined. Here we will use a type of argument relying on the fact that the prefixes in question are unbordered. First we look at some simple bounds for single chains which, although already obsolete because of Theorem 2, serve as simple demonstrations of the benefits afforded by considering unbordered prefixes.

We will need the following simple lemma establishing restrictions on the relative positions of the rightmost occurrences of two squares whose roots have the same cluster.

Lemma 3. *[7] Let $u^2 \neq v^2$ be two squares in some word w with $u \leq_p v$ and $\mathbf{clust}_w(u) = \mathbf{clust}_w(v)$. If their corresponding rightmost occurrences start at positions u_s and v_s, respectively, then $|u_s - v_s| \geq |u|$.*

We call S a *grounded chain* if the shortest u which is the root of a square occurring in w and is a common prefix of all elements of S, is also in S. For some $u_i, u_j \in S$, we call u_i^2 *covered* by u_j if $u_i <_p u_j \leq_p u_i^2$ or $u_i^2 \leq_p u_j$. The *shortest square root* of a grounded chain S is denoted by $\mathbf{ssr}(S)$ and represents the shortest element in S. Note that $\mathbf{ssr}(S)$ is not bordered. If it were, say $\mathbf{ssr}(S) =$

pqp, for some $p \neq \varepsilon$, then $\mathbf{ssr}(S)^2 = pqppqp$, contains p^2, so $p \in S$, which contradicts $\mathbf{ssr}(S)$ being the shortest element in S. Finally, for two different square roots x and u with $x <_p u$ denote $\mathbf{diff}(x, u) = |\mathbf{clust}(x)| - |\mathbf{clust}(u)|$.

Lemma 4. *For a grounded chain S, let m be the number of covered squares with roots in S and let $x = \mathbf{ssr}(S)$. Then, $|u_n|_x \geq m$.*

Proof. If a square u_i^2 is covered by some u_j, then x occurs at position $|u_i|$ in u_j and in all u_k, with $k > j$, because $u_j \leq_p u_k$. In fact, x occurs at position $|u_i|$ even in u_ℓ, for $i < \ell < j$, as $u_i \leq_p u_\ell \leq_p u_j$. □

Lemma 5. *For a grounded chain of square roots $S = \{u_1, \ldots, u_n\}$, with $x = \mathbf{ssr}(S)$, we have $\mathbf{diff}(x, u_i) + |u_i|_x < \mathbf{diff}(x, u_{i+1}) + |u_{i+1}|_x$.*

Proof. Since $u_i \leq_p u_{i+1}$, we have $|u_i|_x \leq |u_{i+1}|_x$, while $\mathbf{clust}(u_{i+1}) \subseteq \mathbf{clust}(u_i)$ gives $\mathbf{diff}(x, u_i) \leq \mathbf{diff}(x, u_{i+1})$, so both terms in the sum are non-decreasing. Moreover, at least one of them increases in each step: if u_{i+1} covers u_i^2, then $|u_i|_x < |u_{i+1}|_x$, while if it does not, then $|\mathbf{clust}(u_{i+1})| < |\mathbf{clust}(u_i)|$). □

Corollary 2. *For a grounded chain of square roots $S = \{u_1, \ldots, u_n\}$ with $x = \mathbf{ssr}(S)$ we have $n < \frac{3|\mathbf{clust}(x)|}{2} - 1$.*

Proof. Since $u_1 = x$ we have $\mathbf{diff}(x, x) = 0$ and $|u_1|_x = 1$. By Lemma 5, for each $i \in \{1, \ldots, n\}$, the sum $\mathbf{diff}(u_i, x) + |u_i|_x$ is strictly increasing, so $n < \mathbf{diff}(x, u_n) + |u_n|_x$. Since $|\mathbf{clust}(u_n)| \geq 2$, we have $\mathbf{diff}(u_n, x) = |\mathbf{clust}(x)| - |\mathbf{clust}(u_n)| \leq |\mathbf{clust}(x)| - 2$. Also, since u_n^2 occurs, the size of $\mathbf{clust}(x)$ is at least $2|u_n|_x$, that is, $|u_n|_x \leq \frac{\mathbf{clust}(x)}{2}$. Adding the two gives us the statement. □

By the above we have that, for a chain S, the number of clusters is bounded by $3n/2$, where $n = |\mathbf{clust}(\mathbf{ssr}(S))|$. Using Lemma 3 we can further refine this.

Proposition 1. *For a grounded chain of square roots $S = \{u_1, \ldots, u_n\}$ with $x = \mathbf{ssr}(S)$ we have $n < 4|\mathbf{clust}(x)|/3$.*

Proof. Let us look at the topmost level where two clusters are equal, that is, suppose $\mathbf{clust}(u) = \mathbf{clust}(v)$ for $u <_p v$ and for all y with $v <_p y$ there exists no z with $\mathbf{clust}(y) = \mathbf{clust}(z)$.

Since $\mathbf{clust}(u) = \mathbf{clust}(v)$, by Lemma 3 we have that $|\mathbf{clust}(u)| \geq 3$ and there are at least three non-overlapping occurrences of u. From here, if $x = \mathbf{ssr}(S)$, we get that $|u|_x \leq \frac{|\mathbf{clust}(x)|}{3}$, but the consecutive clusters above v are never equal, hence the number of clusters is at most $|\mathbf{clust}(x)|$ plus the number of times when two consecutive cluster are equal. The latter is at most $|u|_x$, hence we get that the number of clusters is at most $\frac{4 \cdot |\mathbf{clust}(x)|}{3}$. □

As the main focus of this section we present an adaptation of the technique we used for the upper bound on single chains, for showing that the combined size of two overlapping prefix chains of roots cannot be larger than the number of occurrences of their common prefix, *when that prefix is unbordered*. The latter

qualification is an important one, even though we believe that this is a promising direction towards the full solution of the conjecture. The requirement that the prefix is unbordered not only means that we cannot deduce our conjecture for arbitrary base clusters, but also that we cannot generalize the result to multiple overlaps between multiple chains in a straightforward manner. This, in turn, means that a piece of the puzzle is still missing for the proof of Conjecture 2.

For easy referencing we will denote by different letters the roots which are in the shared part of the two chains and the differing parts of the chains, respectively. Let $X = \{x_1 <_p \cdots <_p x_k\}$ be the common part. The chains $U = \{u_1 <_p \cdots <_p u_m\}$ and $V = \{v_1 <_p \cdots <_p v_n\}$ are the differing parts, so we have $u_1 <>_p v_1$, and of course, as the x_i are the common part, $x_k <_p u_1$ and $x_k <_p v_1$. Since the result in this section does not yield a full proof of the conjecture yet anyway, we will only treat the case of squares instead of general K-powers, to simplify the exposition.

First we show a slightly stronger version of Lemma 1, where we do not necessarily need the whole word to occur three times in its square to imply its non-primitivity.

Lemma 6. *Let $t_1, \ldots, t_n$ with $n \geq 2$ be arbitrary words and let x be any unbordered word such that $|t_i|_x = 0$, for all $i \in \{1, \ldots, n\}$. Let P_i denote the product $xt_1xt_2 \cdots xt_i$. If*

$$|P_n^2|_{P_{n-1}x} > 2$$

then P_n is non-primitive.

Proof. Since x is unbordered and is not contained in t_i, we can reformulate the statement into an equivalent one over the alphabet containing the letters t_i, $i \in \{1, \ldots, n\}$ as follows: $|(t_1 \cdots t_n)^2|_{t_1 \cdots t_{n-1}} > 2$ implies that the word $t_1 \cdots t_n$ of length n is non-primitive. Since $P_{i-1}x$ occurs at least 3 times in P_i^2, we get that $t_1 \cdots t_{n-1}$ occurs at least three times in $(t_1 \cdots t_n)^2$. Let the second occurrence of $t_1 \cdots t_{n-1}$ start at the ith letter (with $i > 1$) of the square $(t_1 \cdots t_n)^2$. This means that $t_1 \cdots t_n$ has period $i - 1$ and therefore $r = t_1 \cdots t_{i-1} t_1 \cdots t_{n-1}$ also has period $i - 1$. At the same time r is the prefix of the square $(t_1 \cdots t_n)^2$, so it also has period n, moreover, its length is $n - 1 + (i - 1) = n + (i - 1) - 1$. From here by the Fine and Wilf theorem r has period $\gcd(i - 1, n)$ and we get that $t_1 \cdots t_n$ is not primitive which implies the statement. □

To describe the assignment of positions to squares we need some definitions first. For a given x, the *ℓ-level x-prefix* ((ℓ, x)-prefix) of a word z is a word z' such that

- $z' <_p z$, and
- $|z'|_x = \ell$.

Further, the *ℓ-level x-representative* ((ℓ, x)-rep) of a square z^2 is the longest prefix of z^2 bordered by the (ℓ, x)-prefix of z^2. The assignment will differ for u_i and v_j depending on the number of occurrences of x in them. Let us partition the roots in U based on whether they have more x's than $\mathbf{lcp}(u_1, v_1)$ or not, so

$U = U_= \cup U_>$ with $U_= = \{u \in U \mid |u|_x = |\mathbf{lcp}(u_1, v_1)|_x\}$ and $U_> = U \setminus U_=$. We partition V similarly into $V_=$ and $V_>$. Now we are ready to describe the assignment of anchors as follows:

- For all x_i^2 and *also for all* u_i^2, v_j^2 with $u_i \in U_=$ and $v_j \in V_=$, we set the x-rep as in the single chain case, i.e., the longest prefix of the square ending in x and start of the last x in the leftmost x-rep as the anchor.
- For the other u_i and v_j, we set the start of the last x in their leftmost occurring (ℓ, x)-rep as the anchor, where $\ell = |\mathbf{lcp}(u_1, v_1)|_x + 1$.

Lemma 7. *Let $u, v \in X \cup U \cup V$ be two distinct square roots. If the anchors of u^2 and v^2 coincide then the shorter between u and v is non-primitive.*

Proof. We have to check what happens when squares collide for each pairing of $X, U_=, U_>, V_=$ and $V_>$. These potentially 25 pairings reduce to 15 as the order does not matter, and can be treated in 7 groups, as we will see below. Like before, the starting position of the x-rep of an arbitrary square z^2 will be denoted by z_s and we set $z_m = z_s + |z|$.

1. $u \in U_>$ and $v \in V_>$: impossible, because in the x-rep of u^2 at the ℓth x before the anchor we have the (ℓ, x)-prefix starting, whereas in the x-rep of v^2 we would have the (ℓ, x)-prefix of v at the same position, but the two are incomparable by the prefix relation as they are longer than $\mathbf{lcp}(u_1, v_1)$.
2. $u \in U_>$ and $v \in U_>$ (analogous to pairing $(V_>, V_>)$): possible; the (ℓ, x)-prefix of u and v are the same, say y. In this case we can apply Lemma 2 as the y-anchors of u and v coincide, giving the non-primitivity of the shorter between the two with a primitive root of length less than u_1.
3. $u \in U_>$ and $v \in V_=$ (analogous to the pairing $(V_>, U_=)$): possible; in this case we have $u_s < v_s$. If $u_m \leq v_s$, then the x-rep of v^2 occurs earlier, a contradiction. If $u_s < v_s < u_m$, then we can apply Lemma 6 and we get that v is non-primitive.
4. $u \in U_>$ and $v \in U_=$ (analogous to the pairings $(V_>, V_=)$, $(U_>, X)$, $(V_>, X)$): possible; here the anchor of u is defined as the last x occurrence in a copy of its (ℓ, x)-prefix $u'x$, whereas the anchor of v is the last occurrence of x in its x-rep $vv'x$. As u contains more occurrences of x than v does, which has exactly as many as the $\mathbf{lcp}$ of u_1 and v_1, we get that $v'x <_p u'x$. Now we can apply Lemma 6 and conclude that v is non-primitive.
5. $u \in U_=$ and $v \in V_=$: impossible because the fact that $|u|_x = |v|_x$ implies $u_m = v_m$ which, in turn, also means $u_s = v_s$. However, at u_s we have an occurrence of u_1 and at v_s an occurrence of v_1, which are incomparable.
6. $u \in U_=$ and $v \in U_=$ (analogous to $(V_=, V_=)$): impossible, by an argument similar to the previous point. Since u and v have the same number of x occurrences, we get $u_m = v_m$ and then $u_s = v_s$ which implies $u = v$.
7. $u \in U_=$ and $v \in X$ (analogous to pairings $(V_=, X)$, (X, X)): possible; this is again a case where Lemma 2 applies as the anchors are all x-anchors, giving non-primitivity of v with root shorter than x.

All 15 cases have been listed above and all are either impossible or result in the non-primitivity of the shorter root. □

We obtained that the collision of anchors results in non-primitive shorter root. A reallocation of the non-primitively rooted squares is likely possible following the logic used for squares ([7] proof of Theorem 2). However, it is probably more technically involved in this overlapping case, and since we do not know how to generalize Lemma 7 to more chains with complex overlapping structure, it seems of limited use at the moment and we decided not to pursue it here due to the space restrictions. However, some manner of separately anchoring the chains based on their **lcp** with neighboring incomparable chains seems a promising way towards a final solution, so we expect that the analysis above will prove useful.

References

1. Bannai, H., I, T., Inenaga, S., Nakashima, Y., Takeda, M., Tsuruta, K.: The "runs" theorem. SIAM J. Comput. **46**(5), 1501–1514 (2017)
2. Blanchet-Sadri, F., Mercaş, R., Scott, G.: Counting distinct squares in partial words. Acta Cybern. **19**(2), 465–477 (2009)
3. Crochemore, M., Fazekas, S., Iliopoulos, C., Jayasekera, I.: Number of occurrences of powers in strings. Internat. J. Found. Comput. Sci. **21**(4), 535–547 (2010)
4. Crochemore, M., Iliopoulos, C., Kubica, M., Radoszewski, J., Rytter, W., Waleń, T.: The maximal number of cubic runs in a word. J. Comput. System Sci. **78**(6), 1828–1836 (2012)
5. Crochemore, M., Rytter, W.: Squares, cubes, and time-space efficient string searching. Algorithmica **13**(5), 405–425 (1995)
6. Deza, A., Franek, F., Thierry, A.: How many double squares can a string contain? Discrete Appl. Math. **180**, 52–69 (2015)
7. Fazekas, S.Z., Mercaş, R.: Clusters of repetition roots: Single chains. In: Proc. 47th SOSFSEM. LNCS, vol. 12607, pp. 400–409. Springer (2021)
8. Fazekas, S.Z., Seki, S.: Square network on a word. Theoretical Computer Science **894**, 121–134 (2021)
9. Fraenkel, A., Simpson, J.: How many squares must a binary sequence contain? Electron. J. Combin. **2**, #R2 (1995)
10. Fraenkel, A., Simpson, J.: How many squares can a string contain? J. Combin. Theory Ser. A **82**(1), 112–120 (1998)
11. Gusfield, D.: Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology. Cambridge University Press (1997)
12. Hickerson, D.: Less than $2n$ distinct squares in a word of length n (2003), communicated by Dan Gusfield
13. Ilie, L.: A simple proof that a word of length n has at most $2n$ distinct squares. J. Combin. Theory Ser. A **112**(1), 163–164 (2005)
14. Ilie, L.: A note on the number of squares in a word. Theoret. Comput. Sci. **380**(3), 373–376 (2007)
15. Jonoska, N., Manea, F., Seki, S.: A stronger square conjecture on binary words. In: Proc. 40th SOFSEM. LNCS, vol. 8327, pp. 339–350 (2014)
16. Kolpakov, R., Kucherov, G.: Finding maximal repetitions in a word in linear time. In: Proc. 40th FOCS. pp. 596–604. IEEE Computer Society Press (1999)

17. Lothaire, M.: Combinatorics on Words. Cambridge University Press (1997)
18. Manea, F., Seki, S.: Square-density increasing mappings. In: Manea, F., Nowotka, D. (eds.) Combinatorics on Words. pp. 160–169. Springer International Publishing, Cham (2015)
19. Storer, J.A.: Data Compression: Methods and Theory. Comp. Sci. Press, Inc. (1988)
20. Thierry, A.: A proof that a word of length n has less than 1.5n distinct squares (2020), https://arxiv.org/abs/2001.02996
21. Thue, A.: Über unendliche Zeichenreihen. Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl. **7** (1906)
22. Thue, A.: Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. Kra. Vidensk. Selsk. Skrifter. I Mat. Nat. Kl. **1** (1912)