Counting maximal-exponent factors in words

Golnaz Badkobeh[1], Maxime Crochemore[2,3], and Robert Mercaş[2,4⋆]

[1] Department of Computer Science, University of Warwick, UK
[2] Department of Informatics, King's College London, UK
[3] Université Paris-Est, France
[4] Department of Computer Science, Kiel University, Germany
Golnaz.Badkobeh@gmail.com, Maxime.Crochemore@kcl.ac.uk,
RobertMercas@gmail.com

Abstract. This article shows tight upper and lower bounds on the number of occurrences of maximal-exponent factors occurring in a word.

Keywords: combinatorics on words, word exponent, maximal-exponent factor

1 Introduction

The topic of repeating segments in words is one of major interest in combinatorics on words. The topic has been studied for more than a century by many authors after the seminal work of [20] who described infinite words containing no consecutive occurrences of the same factor.

Beyond the theoretical aspect of questions related to redundancies in words, repetitions, also called repeats in the following, are often the base for string modelling adapted to compression coding. They play an important role in run-length compression and in Ziv–Lempel compression, e.g., [5]. Moreover, repetitions receive considerable attention in connection with the analysis of genetic sequences. Their occurrences are called tandem repeats, satellites or SRS and should accept some notion of approximation. The existence of some palindromic repeats is crucial for the prediction of the secondary structure of RNA molecules influencing their biological functions, see [6].

Repetitions are composed of consecutive occurrences of the same factor. Their occurrences have been extended to runs, maximal periodic factors, by Iliopoulos et al. [15] and their number has been shown to be less than the word length n by Bannai et al. [4, 3] (see also [10]) and even further less than $22n/23$ by Fischer et al. [12].

In this article we consider factors that repeat non consecutively in a given word of length n. They are of the form uvu where u is their longest border (factor occurring both at the beginning and end of the word). Their exponent,

⋆ Supported by the P.R.I.M.E. programme of DAAD co-funded by BMBF and EU's 7th Framework Programme (grant 605728) and the Newton International Fellowship with funds from the Royal Society and the British Academy.

defined as the ratio of their length over their smallest period length, that is, $|uvu|/|uv|$, is smaller than 2. The number of occurrences of these factors may be quadratic with respect to the word length even if they are restricted to non extensible occurrences. This is why we focus on factors having the maximal exponent among all factors occurring in a square-free word. They are called maximal-exponent factors, MEFs in short, and thus have all the same exponent.

The first attempt to count the number of occurrences of MEFs is done in [2]. In there, authors restrict themselves to considering square-free words, and prove that this number is upper bounded by $2.25n$. They also give the example of a word containing $0.66n$ such factors. The reason for restricting the question to only square-free words, words that contain no factor with an exponent at least 2, comes from the question related to the maximum number of runs in a word. If the word contains squares, the maximal exponent of factors is at least 2 and MEF occurrences become runs whose largest number is known to be less than the word length (see [3, 10, 12]).

The concept of α-gapped repeats provides another way to circumvent the quadratic number of repeat occurrences. They are factors of the form uvu where $|uv| \leq \alpha|u|$ for some real $\alpha > 1$ such that u cannot be extended to the right or to the left, without breaking the repeat. Note that allowing the two occurrences of u relates to counting runs and the condition implies that the exponent of α-gapped repeats is at least $1 + 1/\alpha$. After a more restrictive notion of fix-gapped repeat in [14, 17], locating and counting α-gapped repeats was studied first in [7], then more deeply in [16] and in [11, 19]. Eventually, algorithms to locate α-gapped repeats optimally in time $O(\alpha n)$ are described in [9, 13]. The optimality is based on the tight upper bound $O(\alpha n)$ on their occurrences number.

In this article we improve both the upper and the lower bounds on the number of MEF occurrences provided in [2, 1]. While the rest of this section contains preliminaries, the following two sections establish the tools that are to be used. In Section 4 we upper bound by $1.8n$ the number of occurrences of maximal exponent in a length n word, and in Section 5 we give examples of words with an asymptotic number of MEF occurrences of $5n/6$.

Preliminaries. An alphabet is any set, the members of which are called letters. A word or a string is a sequence of letters drawn from an alphabet. The length of a string w is denoted by $|w|$, and represents the number of occurrences of letters in w. Hence $|\texttt{abaca}| = 5$. The empty word ε is a string of length 0 that is considered to be a word over every alphabet.

A word y is a factor (substring) of the word w if the latter can be factorised as $w = xyz$ for two words x and z. Furthermore, y is a prefix of w if x is empty and a suffix of w if z is empty. A factor that is a suffix and also a prefix of w is called a border of w. The mid-position of an occurrence of a factor y whose first letter is at position i on w is defined by $i + \lfloor |y|/2 \rfloor - 1$.

A positive natural number p is a period of y if $y[i] = y[i+p]$ for all i for which the equation is meaningful. Let us denote by $p(y)$ the smallest period length of a word y. The exponent of y, denoted by $e(y)$, is defined as $\frac{|y|}{p(y)}$. The maximal-exponent factors, MEFs for short, are factors of w whose exponents are maximal

amongst all exponents of other factors of w. Note that for any two MEFs uvu and $u'v'u'$ of the same word, the following properties hold:

$$\frac{|u|}{|uv|} = \frac{|u'|}{|u'v'|} \text{ and } \frac{|v|}{|u|} = \frac{|v'|}{|u'|}.$$

For the previous example `abaca`, `aba` and `aca` represent MEFs of the word with $u = \mathtt{a}$ and $v = \mathtt{b}$ or $v = \mathtt{c}$, respectively, where the exponent of the MEF is 1.5. In this work, we investigate the number of occurrences of all maximal-exponent factors in a fixed square-free word w of length n, thus assuming that the minimal period of every such factor is longer than its border.

2 Partitions of the maximal-exponent factors

We begin this section with a recollection of the results from [1], directly related to our topic of investigation. Later on, we build on these results and techniques and give our improved bound.

Lemma 1 ([1]). *Consider two occurrences of MEFs with the same border length b starting at respective i and j positions in the word. Then, $|j - i| > b$.*

Following Lemma 1, counting the occurrences of MEFs by grouping them with respect to their border lengths, will lead to an initial part of the harmonic series, a quantity that is not linear with respect to the length of w. Therefore, in order to obtain a linear upper bound on the number of occurrences of MEFs the authors introduced in [1] the notion of δ-MEFs, for a positive real number δ, as follows. A MEF uvu is a δ-MEF if its border length $b = |u| = |uvu| - p(uvu)$ satisfies $2\delta < b \leq 4\delta$. Then any MEF is a δ-MEF for some $\delta \in \Delta$, where $\Delta = \{1/4, 1/2, 1, 2, 2^2, 2^3, \ldots\}$. This is not a new technique and it has been previously applied to count runs in words, e.g., [18, 8].

Lemma 2 ([1]). *Let uvu and $u'v'u'$ be occurrences of δ-MEFs in w whose left borders mid-positions are at respective positions i and j on w. Then, $|j - i| \geq \delta$.*

Exploiting this lemma gives the following upper bound for the number of occurrences of MEFs.

$$\left(\sum_{b=1}^{b=k} \frac{n}{b+1}\right) + \frac{1}{k}\left(2 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right)n = \left(\sum_{b=1}^{b=k} \frac{n}{b+1}\right) + \frac{4n}{k} \qquad (1)$$

The direct consequence of the previous lemma is that if uvu and $u'v'u'$ are two δ-MEFs, then u cannot contain u'.

Next, we study in more detail the positioning of an overlap between two consecutive occurrences of MEFs. We first observe that, for a given word, there exists a unique rational number q such that for every MEF uvu, $|v| = q|u|$. In particular, if the exponent is greater than 1.5, then $q < 1$ and $q \geq 1$ otherwise.

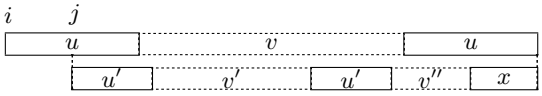


Fig. 1. Two MEFs occur at the distance $j-i$, where $j-i+|u'v'|+|x| \leq |uvu|$, where x is a suffix of u and a prefix of u'.

Lemma 3. *Consider two MEF occurrences with borders u and u', starting respectively at positions i and j, where $i < j$ and $|u'| < |u| < 2|u'|$. Then $j-i \geq |u'|$.*

Proof. It is straightforward to show that $|u| < 2|u'|$ is a necessary condition. For example *abcdebcfghabcde* is a word with maximal-exponent of 1.5 and it also contains a MEF *bcdebc* whose border length is 2. However the distance between the starting positions of these two MEFs is only 1, which is less than $2 = |ab|$.

Let us assume that $j-i < |u'|$ and consider a MEF uvu with $|v| = q|u|$ starting at position i and another MEF $u'v'u'$ with $|v'| = q|u'|$ starting at position $j > i$, where q is a rational number. Furthermore, as a direct consequence of Lemma 2 we know that $j - i + |u'| \geq |u|$.

Denote by x the overlap between u and u', and let $u = yx$ and $u' = xz$ for non-empty words y and z. Since $|u| > |u'|$ thus $|y| > |z|$, there are two cases to consider. In the first case, we consider that $j - i + |u'v'| + |x| > |uvu|$, while in the second case, the contrary.

The first case is not possible because $|y|+(q+1)|u'|-(q+2)|u| \geq q|x|$, which leads to the conclusion that $(q+1)(|u|-|z|) \leq 0$. This is a contradiction since $|u| > |u'|$ according to the assumption.

The second case is depicted in Figure 1. The factor between the occurrence of x in the second u' and that of x in the second u is denoted by zv'' and its length must be at least $q|x|$, because otherwise there will be a factor of greater exponent than of the current MEF. Thus $(q+1)|u| + |y| - |y| - (q+1)|u'| - |x| \geq q|x|$. This leads to $j - i = |y| \geq |xz| = |u'|$, which is a contradiction. □

Now, we investigate the minimum distance between the starting positions of two MEFs of which the leftmost one has a smaller border.

Lemma 4. *Consider two MEFs with borders u and u', starting respectively at positions j and i, where $i < j$ and $|u'| < |u|$. Then $j - i \geq 2|u'| - |u|$.*

Proof. As depicted in Figure 2, consider a MEF uvu with $|v| = q|u|$ starting at position j and a MEF $u'v'u'$ with $|v'| = q|u'|$ starting at position i, where $i < j$

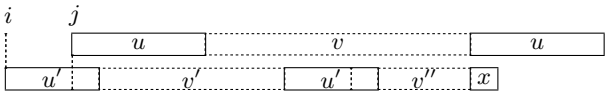


Fig. 2. Two MEFs occur at distance $j-i$, where x is a suffix of u' and prefix of u. If $j-i$ is small then the exponent of $xv''x$ is higher than maximal exponent.

and q is a rational number. One can simply eliminate other possible ending positions of these two MEFs, since $|u'v'u'| < |uvu|$ and $i < j$.

Denote by x the overlap between u and u', and let $u' = yx$ and $u = xz$ for non-empty y and z. Since $|u| > |u'|$ thus $|y| < |z|$, to avoid having an exponent greater than that of the MEF, it is necessary that $v'' \geq q|x|$. Hence the following holds

$$|x| + q|u'| + |u'| + q|x| \leq (q+1)|u|,$$
$$|u'| + |x| \leq |u|,$$

and substituting $|u'| - (j-i)$ for $|x|$ leads to our result $j - i \geq 2|u'| - |u|$. □

The following is the result of merging the last two lemmas.

Lemma 5. *Let $S = [r, .., s]$ be an interval of integers such that $r > \frac{2s}{3}$ and w be a word. Then within every $r+1$ positions of w, there are at most two MEFs whose border lengths are in S.*

Proof. Consider three consecutive MEFs, uvu, $u'v'u'$ and $u''v''u''$ starting at positions i, j and k, respectively, such that $|u|, |u'|, |u''| \in S$. Following Lemmas 3 and 4, there are four cases to consider, depending on the relations between $|u|$ and $|u'|$, and between $|u'|$ and $|u''|$. Observe that, at no point can two consecutive MEFs have identical border lengths because this would contradict Lemma 1.

Assume $|u'| < |u|$, then following Lemma 3, it must be that $j - i \geq |u'|$. Now, if $|u''| < |u'|$, following the same lemma leads to $k - j \geq |u''|$. Adding these gives $k - i \geq |u'| + |u''|$. Since $|u'|, |u''| \in S$ and $3r > 2s$ it can be concluded that $k - i > s + 1$. On the other hand, if $|u'| < |u''|$, then $k - j \geq 2|u'| - |u|''$. Adding now to this the quantity $j - i$, gives the following

$$k - i \geq 3|u'| - |u''| > 2s - s = s, \tag{2}$$

and the conclusion follows in this case.

Now assume that $|u'| > |u|$. Then following Lemma 4, $j - i \geq 2|u| - |u'|$. The conclusion for both cases here, derives in a manner similar to Equation 2. □

Next, following the idea from [1], we introduce the notion of γ-MEFs, for a positive real number γ: a MEF uvu is a γ-MEF if its border length $b = |u|$ satisfies $2\gamma \leq b < 3\gamma$. Then any MEF is a γ-MEF for some $\gamma \in \Gamma$ where $\Gamma = \{\frac{1}{2}, \frac{1}{2} \cdot (\frac{3}{2}), \frac{1}{2} \cdot (\frac{3}{2})^2, \ldots\}$.

Corollary 6. *Let uvu, $u'v'u'$ and $u''v''u''$ be three consecutive γ-MEFs starting at positions i, j and k, respectively, on some word w. Then* $\max\{k-i, j-i\} > 3\gamma$.

Proof. This is a direct consequence of Lemmas 3, 4, and 5, by considering all of the possible four cases. □

We are now ready to improve on the result of Equation 1

Theorem 7. *There are less than $4n/b$ occurrences of MEFs with maximum length border at least b in a length n word.*

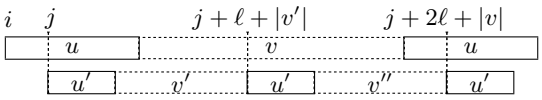


Fig. 3. Two MEFs uvu and $u'v'u'$ with the left occurrence of u' entirely contained in the left occurrence of u. If u is twice longer than u', then another MEF $u'v''u'$ appears.

Proof. We apply Corollary 6 for values of $\gamma \in \Gamma_b$, where $\Gamma_b = \{\frac{b}{2}, \frac{b}{2}\cdot\frac{3}{2}, \frac{b}{2}(\frac{3}{2})^2, \ldots\}$. This will cover all possible MEFs with border length at least b. Hence, we obtain the following upper bound:

$$\sum_{\gamma\in\Gamma_b} \frac{2n}{3\gamma} = \frac{4n}{3b}\sum_i \left(\frac{2}{3}\right)^i = \frac{4n}{b} \tag{3}$$

for the number of occurrences of MEFs with border length at least b. □

As a direct consequence of Theorem 7, one can count the number of MEFs with border length at least b for any positive b. We choose $b = 8$ in this paper because of the way we structured the counting of all MEFs.

Corollary 8. *There are less than $n/2$ occurrences of MEFs with border length at least 8 in a word of length n.*

3 MEFs with double border lengths

In this section, we look at the positioning of overlaps between two MEFs, one of which has border length twice of the other. First, we make an observation regarding the case where the border of the smaller MEF is entirely contained within the border of the bigger MEF.

Lemma 9. *If two MEFs uvu and $u'v'u'$ start at positions i and j, respectively, for which $|u| = 2|u'| = 2\ell$ and $i \le j \le i + \ell$, then the factor starting at position $j + |v'| + l$ and ending at $j + |v| + 3l$ is also a MEF with border length ℓ.*

Proof. The situation is depicted in Figure 3. Because $|u| = 2|u'| = 2\ell$ the following relations hold: $|v| = 2|v'| = 2m$ and $|uvu| = 2|u'v'u'|$.

It is straightforward to show that u' is a proper factor of u because $i \le j \le i + \ell$, therefore u' also occurs in the right occurrence of u. Hence, u' occurs at position $j + \ell + |v'|$ and also at $j + 2\ell + |v|$. Let the factor starting at $j + 2\ell + |v'|$ and ending at $j + 2\ell + |v|$ be denoted by v'', then $|v''| = |v| - |v'| = |v'|$. Thus $u'v''u'$ is also a MEF with border length ℓ. □

The above essentially says that if a MEF has its left border totally included in the left border of a MEF which is twice as long, then another MEF of the same size as the former one and the same borders, will have its right border totally included in the right border of the longer one. This fact, combined with the fact that within the border of a MEF we cannot have more than two starting positions of MEFs of half its length, leads us to the following result:

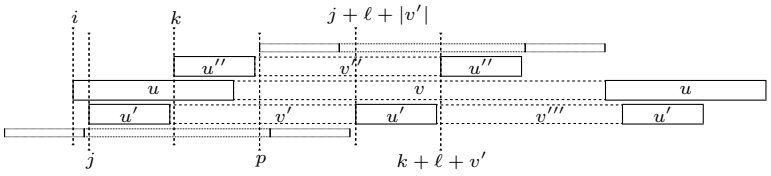


Fig. 4. The MEF with border u is double in length compared to the others, and while u' and u'' start in u, the top and bottom MEFs that are dashed here, start outside u

Lemma 10. *If a MEF of border length 2ℓ starts at position i on w, and two MEFs of border ℓ start at positions j and k, for $i \leq j, k \leq i + 2\ell$, then either:*

- *there exists a factor of length $3\ell + 2$ containing the starting position of only one MEF with border length ℓ, or*
- *there exist two factors of total length $5\ell + 5$ containing the starting positions of at most two MEFs with border lengths ℓ, or*
- *the first MEF with border length 2ℓ following uvu starts at position $i + 2\ell + 3$ or after, and there exist two factors of total length $5\ell + 1$ containing the starting positions of at most two MEFs with border lengths ℓ.*

Proof. Let us denote the MEF of border 2ℓ starting at position i by uvu, the MEF of border length ℓ starting at position j by $u'v'u'$, and the MEF of border length ℓ starting at position k by $u''v''u''$. Due to the fact that $|u| = 2|u'| = 2|u''| = 2\ell$ the following relations hold: $|v| = 2|v'| = 2|v''|$ and $|uvu| = 2|u'v'u'| = 2|u''v''u''|$. Furthermore, following Lemma 1, assuming that $j < k$ we have $i \leq j < i + \ell$ and $j + \ell + 1 < k < i + 2\ell$.

Since $i \leq j < i + \ell$, following Lemma 9, there is another MEF with border of length ℓ at position $j + \ell + |v'|$, which we shall denote by $u'v'''u'$.

First, assume that there exists a MEF of border length ℓ ending somewhere between positions $i-1$ and $j-2$. This is the situation represented at the bottom of Figure 4 by the thinner/dashed border MEF. However, as a consequence of this fact, there is no MEF with border length ℓ ending between positions $i-1+\ell+|v'|$ and position $j+\ell+|v'|$, as otherwise the symbol on position $i-1$ will appear again on position $i + \ell + |v| - 1$, which leads to a contradiction regarding the maximal exponent of our word (we have a factor with border au and length $2\ell + |v| + 1$, where a is the respective symbol). Furthermore, having any MEF of border length ℓ starting anywhere between positions $j + \ell + |v'| + 1$ and $i + 3\ell + |v'|$ will once more lead to a contradiction, either following similar arguments regarding the letter on position $i + 2\ell$, or by overlapping with $u'v'''v'$. Hence, in this case we conclude that there is only one MEF with border length ℓ starting between positions $i - 1 + |v'|$ and $i + 3\ell + |v'|$.

Now, assume the contrary where no MEF of border length ℓ ends between positions $i - 1$ and $j - 2$. Then it is straightforward that between positions $i - \ell - 1$ and $k - 1$ there exists the starting position of only one MEF of border length ℓ. Furthermore, if there is a MEF with border length ℓ starting between position $i + 3\ell + |v'|$ and position $k + 2\ell + |v'| + 1$, then there cannot be any

MEF with border length 2ℓ starting after uvu, before position $k+\ell$, because otherwise it results in a contradiction regarding the maximality of the exponent. But then either there exists also only one MEF with border length ℓ starting between positions $j+|v'|-1$ and $k+2\ell+|v'|+1$, or the first occurrence of a MEF with a border of length 2ℓ following uvu starts after position $k+\ell$. From the former (depicted on the top of Figure 4 by the thinner/dashed border MEF) it immediately follows that there exist two factors of total length

$$(k-1-i+\ell+1)+(k+2\ell+1-j+1) \geq 2k+3\ell-2j+2 > 5\ell+4,$$

which contain the starting positions of at most two MEFs with border length ℓ.

The only case left for analysis is when the first occurrence of a MEF with a border of length 2ℓ following uvu starts after position $k+\ell$, and there is a MEF with border length ℓ starting at some position p with $i+3\ell+|v'| \leq p \leq k+2\ell+|v'|+1$. However, observe that this, in conjunction with Lemma 9, imposes the fact that there are no MEFs of border length ℓ starting between positions $k+1$ and $p-|v'|+1$. Hence, we conclude that between position $i-\ell-1$ and position $i+3\ell+1$ there are only two MEFs of border length ℓ starting. Since between position $j+\ell+|v'|+1$ and position $i+3\ell+|v'|-1$, we do not have the start position of any MEF with border length ℓ, we conclude that once again there exist two factors of total length

$$(4\ell+2)+(i+3\ell-1-j-\ell-1) \geq 6\ell+i-j > 5\ell+1,$$

that contain the starting positions of at most two MEFs with border length ℓ. □

The direct outcome of the previous result is as follows. If there are two MEFs with overlapping left borders such that one has twice the border length of the other, then the possible total number of occurrences of MEFs will be reduced by one. This circumstance is further analysed in the following lemma where a more particular case is investigated.

Lemma 11. *Consider two MEFs uvu and $u'v'u'$ starting at positions $i+1$ and i, respectively, for which $|u| = 2|u'| = 2\ell$. Then, there exists a factor of length $3\ell+2$ within which only one MEF of border length ℓ starts.*

Proof. Obviously, we again have $|v| = 2|v'|$ and $|uvu| = 2|u'v'u'|$.

A first observation is that there is no MEF of border ℓ starting between positions $i+1$ and $i+\ell$ since otherwise the border of this MEF would overlap or be right next to u', which according to Lemma 1, is not possible. Therefore, following Lemma 9 there is no MEF of border ℓ starting between positions $i+\ell+|v'|+1$ and $i+2\ell+|v'|$.

Since $u'v'u'$ starts at position i, the first letter of u', call it a, is also present at position $i+\ell+|v'|$. Now, if there exists another MEF of border length ℓ starting anywhere between position $i+|v'|$ and position $i+\ell+|v'|$, then it will follow that at position $i+\ell+|v|$ there is a letter a. However, this will lead to a contradiction regarding the maximal exponent of the word (we have a factor with border au and length $2\ell+|v|+1$). This together with the fact that there

is no MEF with border of length ℓ starting between positions $i+\ell+|v'|+1$ and $i+2\ell+|v'|$ renders the desired result. □

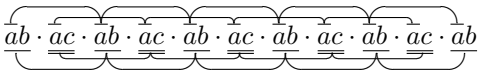
Now collating the previous results leads to a bound on the number of occurrences of MEFs whose border lengths are twice of each other.

Proposition 12. *There are at most $2n/(2\ell+1)$ MEFs with border lengths ℓ and 2ℓ in a word of length n.*

Proof. According to Lemmas 10 and 11 whenever a MEF with border length ℓ overlaps but is not fully contained in a MEF with border of length 2ℓ, there exist two factors whose total length is $2\ell+1$ within which no MEF of border ℓ starts. Hence, whenever a MEF of border 2ℓ overlaps more than one MEF of border ℓ, the maximum number of MEFs of border length ℓ decreases by 1. Furthermore, in some situations, this can also enforce a reduction in the number of MEFs with a border length 2ℓ. As a consequence of Lemma 1 there are at most $n/(2\ell+1)$ MEFs of border length 2ℓ, thus our result follows. □

Following Proposition 12 there are at most $2n/3$ MEFs with border lengths 1 and 2, $2n/5$ MEFs with borders 2 and 4, $2n/7$ MEFs with borders 3 and 6, and so on. In addition, the following example shows that these bounds are tight.

Example 13. Consider words $u = ab, v = ac$ and alphabet $\Sigma = \{a_1, b_1, a_2, b_2, \ldots\}$ for which $a, b, c \notin \Sigma$. Let $S_1 = ua_1vb_1$ and $S_i = S_{i-1}ua_ivb_i$ for $i \geq 2$.



This sequence is a prefix of S_∞, where all symbols from Σ have been replaced by dots, for simplicity. It is not difficult to observe that in every factor of length 3, there is an occurrence of a causing a factor of exponent $4/3$. Furthermore, because there are four letters between every two occurrences of u and every two occurrences of v, within every factor of length three there is a starting position of a factor of length 8 whose exponent is $4/3$. Since $a_i \neq a_j$ and $b_i \neq b_j$ for any $i \neq j$, and $a_i \neq b_j$ for any $i, j \geq 1$, we conclude that in fact $4/3$ is the maximum exponent of the sequence S_∞. Therefore, every length n prefix of S_∞ contains at most $2n/3$ factors of exponent $4/3$ and this bound is reached for certain values of n. □

Following the same strategy the subsequent lemma refines the number of maximal-exponents factors whose border lengths are exponentially increasing.

Lemma 14. *Every word of length n contains, for any positive integer $\ell \leq \frac{n}{2}$, at most $\frac{n^2}{\ell n+2\ell}$ MEFs with the border length of the form $\ell\, 2^i$, for any $0 \leq i \leq \log(\frac{n}{\ell})$.*

Proof. As seen from the previous proofs, if there is a MEF with border length $\ell\, 2^{i-1}$, it cannot overlap more than 2^{i-2} MEFs of border length $\ell\, 2^j$ with $j < i-1$, without reducing the total possible number of MEFs. Thus a MEF of border

length $\ell\, 2^i$ will contain a MEF of border length $\ell\, 2^{i-1}$ and the 2^{i-2} MEFs of smaller border length included in it, plus as many as $\ell 2^{i-2}$ that are not included in the MEF of border length $\ell\, 2^{i-1}$ but are fully contained in the one of border length $\ell\, 2^i$. Since $\ell\, 2^i \leq \frac{n}{2}$ we have $i \leq \log(\frac{n}{2\ell})$. This combined with Lemma 1 completes the proof. □

By choosing $\ell = 1$, it is immediate from Lemma 14 that the number of MEFs whose border length have the form 2^i is bounded by $\frac{n^2}{n+2}$. We conclude this section with a further refinement on the number of MEFs with border lengths that are small and exponentially increasing.

Lemma 15. *Every word of length n contains at most $4n/5$ MEFs with the border length in the set $\{1, 2, 4\}$.*

Proof. Following Proposition 12 we know that there are at most $2n/5$ MEFs of border lengths 2 and 4. Since we want to maximise the total number of MEFs and the number of MEFs with border of length 1 is dependent on the number of MEFs of border of length 2 (that is relative to the overlaps between them according to the above results), we conclude that we cannot have more than $2n/5$ such MEFs (one inside every MEF with a border of length 2 and another one inside the MEF with border length 4, not adjacent to the border 2 one corresponding to it). Hence the total number of MEFs with border length an element of $\{1, 2, 4\}$ is bounded by $\frac{2}{5}n + \frac{2}{5}n = \frac{4}{5}n$. □

4 Upper bounds on the number of MEF occurrences

This section makes the final stride towards improving the upper bound on the number of MEFs. This upper bound is further improved in the case of words with a maximal exponent greater than 1.5. In addition, an optimal upper bound is presented for a specific class of words with a maximal exponent of 1.5 and the length of the MEFs not divisible by their border length.

Lemma 16. *There are at most $13n/10$ occurrences of MEFs whose border length is at most 7 in a word of length n.*

Proof. According to Lemma 5, MEFs with border length at most 7 can be partitioned into three groups: MEFs with border length in $S_1 = \{1, 2, 4\}$ or in $S_2 = \{3\}$, or in $S_3 = \{5, 6, 7\}$. There are at most $4n/5$ occurrences of MEFs with border lengths 1, 2, or 4 according to Lemma 15. There are at most $n/4$ occurrences of MEFs with border length 3 by Lemma 1. Finally, there are at most $n/4$ occurrences of MEFs with border lengths 5, 6 or 7 by Lemma 5. Adding all these together leads to the stated result. □

The following theorem is a direct consequence of Lemma 16 and Corollary 8.

Theorem 17. *There exist at most $1.8n$ number of occurrences of MEFs in a word of length n.*

Although the upper bound in Theorem 17 is true in general, this bound can be further improved when special cases of MEFs are considered. The cases are distinguished by the value of the maximal exponent.

Remark 18. If the exponent of the MEFs is greater than 1.5, then for every MEF uvu, $|u| > |v|$.

This observation implies that no MEF with a border of length 1 exists in this case. Furthermore, given the fact that for two different lengths MEFs, uvu and $u'v'u'$, the following hold

$$\frac{|u|}{|uv|} = \frac{|u'|}{|u'v'|} \text{ and } \mid |u| - |u'| \mid \geq 2.$$

The rationale for the latter observation is that the lengths of the borders are always integers that increase proportionally with the exponent. Hence, the following result is implied:

Lemma 19. *There are at most n occurrences of maximal-exponent factors in a word of length n, whenever the maximal exponent is greater than* 1.5.

Proof. First, according to the previous remark, there exists no MEF of border length 1. Furthermore, the lengths of every two different lengths MEFs, are proportional to some q and differ by at least 2. The counting can be split into two parts: counting MEFs with border length at most 7 and the remaining MEFs. For the first case a simple arithmetic argument can show that having MEFs of border lengths $2, 4$ or 6 will maximise the total count of MEFs complying with these conditions. The following is an upper bound on number of MEF occurrences whose border length is at most 7. The calculation of this upper bound is realised by grouping the MEFs with borders of length $\{2, 4\}$ and counting separately those with border length 6.

$$\frac{2n}{5} + \frac{n}{6+1} = \frac{19n}{35}$$

It is straightforward that according to Corollary 8 there are at most $n/2$ occurrences of MEFs whose border length is at least 8, of which at most half comply with the constraint on the minimum difference between the border length which is 2. Therefore, there are at most $n/4$ such occurrences of MEFs. Finally, connecting these two cases yields the following upper bound:

$$\frac{19n}{35} + \frac{n}{4} < n$$

This concludes the proof. □

The previous result can be further strengthened by also looking at MEFs of a smaller exponent but having a further restriction. The following result sums up the result of the above lemma and this particular class of MEFs.

Theorem 20. *Every length n word contains at most n occurrences of MEFs whenever the length of these factors is not a multiple of their longest border.*

Proof. Just as for Lemma 19, the condition implies that there exists no MEF of border 1. Furthermore, once more we make the simple observation that the difference between the lengths of different length MEFs is at least 2. Thus the result follows in a manner similar to that in Lemma 19. □

Observe however that for the case where MEFs have exponent at most 1.5 and their length is a multiple of the border's length, i.e., if uvu is a MEF, then there exists a integer $q > 0$ such that $|v| = q|u|$ and $q > 0$, the best upper bound remains that of Theorem 17. Nevertheless, we can assume that for this class of MEFs, the smallest border has length at most 3, as otherwise, we can simply drop the first three terms of the left-most sum in Equation 1, which gives us a result similar to that of Theorem 20, i.e., in this case $b = 4$ would be the first value we consider, and thus we would have, again, at most n occurrences of maximal-exponent factors.

5 Lower bounds on the number of MEF occurrences

Finally, we end this work with an example of a construction that generates a word that has a ratio of 5/6 of MEF occurrences relative to its length, with the maximal exponent 10/9. This improves the result presented in [1, Section 6.2].

In the following we consider the fixed alphabet

$$\Sigma = \{\mathtt{a}_1, \mathtt{a}_2, \mathtt{a}_3, \mathtt{b}_1, \mathtt{b}_2, \mathtt{b}_3, \mathtt{b}_4, \mathtt{c}_1, \mathtt{c}_2, \mathtt{c}_3, \mathtt{c}_4, \mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_3, \mathtt{d}_4, \mathtt{e}_1, \mathtt{e}_2, \mathtt{e}_3, \mathtt{e}_4\},$$

and the infinite alphabet

$$\Sigma_\infty = \{\mathtt{f}_{1,1}, \mathtt{f}_{2,1}, \ldots, \mathtt{f}_{8,1}, \mathtt{f}_{1,2}, \mathtt{f}_{2,2}, \ldots\}.$$

We define the following sequence for $i > 0$:

$$\begin{aligned}
u_{(1,i)} &= \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_1\ \mathtt{a}_2\ \mathtt{d}_1\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_1\ \mathtt{f}_{1,i}\\
u_{(2,i)} &= \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_2\ \mathtt{a}_2\ \mathtt{d}_2\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_1\ \mathtt{f}_{2,i}\\
u_{(3,i)} &= \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_3\ \mathtt{a}_2\ \mathtt{d}_3\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_2\ \mathtt{f}_{3,i}\\
u_{(4,i)} &= \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_4\ \mathtt{a}_2\ \mathtt{d}_4\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_2\ \mathtt{f}_{4,i}\\
u_{(5,i)} &= \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_1\ \mathtt{a}_2\ \mathtt{d}_5\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_3\ \mathtt{f}_{5,i}\\
u_{(6,i)} &= \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_2\ \mathtt{a}_2\ \mathtt{d}_6\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_3\ \mathtt{f}_{6,i}\\
u_{(7,i)} &= \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_3\ \mathtt{a}_2\ \mathtt{d}_7\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_4\ \mathtt{f}_{7,i}\\
u_{(8,i)} &= \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_4\ \mathtt{a}_2\ \mathtt{d}_8\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_4\ \mathtt{f}_{8,i}
\end{aligned}$$

and the infinite word $\Omega = \prod_{i=1}^{\infty} \left(\prod_{j=1}^{8} u_{(j,i)} \right)$.

Proposition 21. *The ratio between the length of the prefixes of Ω and the number of occurrences of its maximal-exponent factors they contain tends to 5/6.*

Proof. Since each factor $u_{(j,i)}$ is identified by its last letter $\mathtt{f}_{j,i}$, it follows that no factor of length greater than 8 repeats in Ω. Hence we focus on short factors.

Let x be a generic factor $\prod_{j=1}^{8} u_{(j,i)}$ for some i. From the construction of Ω we observe that every letter $\mathtt{a}_k$, $0 < k < 4$, is repeated at every 9 positions leading to 8 MEF occurrences in Ω and 7 in x. These are the symbols, which occur most often. Next, every factor of length 2 ending with $\mathtt{b}_k$, $0 < k < 5$, is repeated at every 18 positions producing 4 MEF occurrences in Ω and 3 in x. In the same manner, factors of length 4 with $\mathtt{c}_k$, $0 < k < 5$, at their third position repeat at every 36 positions and produce 2 MEF occurrences in Ω and 1 in x each, while factors of length 8 containing $\mathtt{d}_k$, $0 < k < 9$, do not repeat in x but produce 1 MEF occurrence in Ω each. Finally, every two consecutive symbols $\mathtt{e}_k$, $0 < k < 5$, separated by 9 positions yield a total of 4 MEF occurrences in x.

The above repeats are the only ones producing MEFs since factors of lengths 3, 5, 6 and 7 repeat as factors of lengths 4 or 8 producing a smaller exponent.

Therefore, overall there are $3 \times 8 + 4 \times 4 + 4 \times 2 + 8 \times 1 + 4 = 60$ MEF occurrences whose starting positions are on x. If x appears at the end of a prefix of Ω only $3 \times 7 + 4 \times 3 + 4 \times 1 + 4 = 41$ of its positions are starting positions of MEF occurrences in the prefix.

Eventually, a prefix $\prod_{i=1}^{m} \left(\prod_{j=1}^{8} u_{(j,i)} \right)$ of Ω has length $72m$ and contains $60(m-1) + 41$ MEF occurrences. When m tends to infinity, we get an average of $60/72 = 5/6$ MEF occurrences per position as stated. □

Note that the maximal exponent of factors in Ω is $10/9$ and that its construction can be extended to whatever exponent of the form $(2^\ell + 2)/(2^\ell + 1)$, in a similar fashion. It is also our belief that this construction can be generalised as to generate, for any integer ℓ, an infinite word Ω_ℓ in which every MEF has a border length of the form 2^i, $i \leq \ell$, and whose asymptotic number of MEF occurrences per position grows very closely to 1 with ℓ.

Finally, observe that letters $\mathtt{f}_{j,i}$ occurring in Ω can be drawn from an 11-letter alphabet disjoint from Σ. To do so, it suffices to replace the infinite subsequence of $\mathtt{f}_{j,i}$ by an infinite sequence whose maximal exponent of factors is $11/10$, Dejean's repetitive threshold of the alphabet. No MEFs considered in the previous proof will be affected.

References

1. G. Badkobeh and M. Crochemore. Computing maximal-exponent factors in an overlap-free word. *Journal of Computer and System Sciences*, 2015. to appear.
2. G. Badkobeh, M. Crochemore, and C. Toopsuwan. Computing the maximal-exponent repeats of an overlap-free string in linear time. In *19th SPIRE*, volume 7608 of *Lecture Notes in Computer Science*, pages 61–72, 2012.
3. H. Bannai, T. I, S. Inenaga, Y. Nakashima, M. Takeda, and K. Tsuruta. The "runs" theorem. *CoRR*, abs/1406.0263, 2014.
4. H. Bannai, T. I, S. Inenaga, Y. Nakashima, M. Takeda, and K. Tsuruta. A new characterization of maximal repetitions by Lyndon trees. In *26th SODA*, pages 562–571. SIAM, 2015.

5. T. C. Bell, J. G. Clearly, and I. H. Witten. *Text Compression*. Prentice Hall Inc., New Jersey, 1990.
6. H.-J. Böckenhauer and D. Bongartz. *Algorithmic Aspects of Bioinformatics*. Springer, Berlin, 2007.
7. G. S. Brodal, R. B. Lyngsø, C. N. S. Pedersen, and J. Stoye. Finding maximal pairs with bounded gap. In *10th CPM*, volume 1645 of *Lecture Notes in Computer Science*, pages 134–149, 1999.
8. M. Crochemore, L. Ilie, and L. Tinta. The "runs" conjecture. *Theoretical Computer Science*, 412(27):2931–2941, 2011.
9. M. Crochemore, R. Kolpakov, and G. Kucherov. Optimal searching of gapped repeats in a word. *CoRR*, abs/1509.01221, 2015.
10. M. Crochemore and R. Mercaş. Fewer runs than word length. *CoRR*, abs/1412.4646, 2014.
11. M. Dumitran and F. Manea. Longest gapped repeats and palindromes. In *40th MFCS*, volume 9234 of *Lecture Notes in Computer Science*, pages 205–217, 2015.
12. J. Fischer, Š. Holub, T. I, and M. Lewenstein. Beyond the runs theorem. In *22nd SPIRE*, volume 9309 of *Lecture Notes in Computer Science*, pages 277–286, 2015.
13. P. Gawrychowski, T. I, S. Inenaga, D. Köppl, and F. Manea. Efficiently finding all maximal α-gapped repeats. *CoRR*, abs/1509.09237, 2015.
14. D. Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.
15. C. S. Iliopoulos, D. Moore, and W. F. Smyth. A characterization of the squares in a Fibonacci string. *Theoretical Computer Science*, 172(1–2):281–291, 1997.
16. R. Kolpakov, M. Podolskiy, M. Posypkin, and N. Khrapov. Searching of gapped repeats and subrepetitions in a word. In *25th CPM*, volume 8486 of *Lecture Notes in Computer Science*, pages 212–221, 2014.
17. R. M. Kolpakov and G. Kucherov. Finding repeats with fixed gap. In *7th SPIRE*, pages 162–168, 2000.
18. W. Rytter. The number of runs in a string. *Information and Computation*, 205(9):1459–1469, 2007.
19. Y. Tanimura, Y. Fujishige, T. I, S. Inenaga, H. Bannai, and M. Takeda. A faster algorithm for computing maximal α-gapped repeats in a string. In *22nd SPIRE*, volume 9309 of *Lecture Notes in Computer Science*, pages 124–136, 2015.
20. A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. I. Mat. Nat. Kl. Christiana*, 7:1–22, 1906.

The paper presents new upper and lower bounds for the
number of maximal-exponent factors contained in a word.
The results are significant improvements on the previously
known bounds (obtained by the first two authors of this
paper in some previous publications); as such, the results
are relevant to the community interested in combinatorics
and algorithms on words. The techniques used in this
paper are not surprising, but I think that their usage is
both nice and clever. In conclusion I think that this paper
should be accepted to TCS.

The current presentation of the paper is unfortunately not
so good. Below I attach a (long) list of improvement suggestions.

Introduction:
p.1, line -22: Repeating sequence might be a major research
topic in combinatorics on words. I do not see how they can
be a question. Probably you wanted to say that their study
is "an important question"

Answer: We have rephrased the paragraph, as suggested by the referee.

p.1, line -22: The second sentence starting on this line: who is "it"?

Answer: We have fixed this.

p.1, line -17: what is the difference between repetitions and repeats?
The first sentence of this paragraph does not make much sense.

Answer: We clarified that they represent the same thing.

p.1, line -15: replace "i.e." by "e.g.". "i.e." means "id est"
(that is) while "e.g." means "exempli gratia" (for example
or see for example). I guess you want the second one.

Answer: Done.

p.1, line -15: the part of the paragraph starting with "Moreover.."
until the end of the paragraph. Now I am totally confused: are
repetitions sometimes called repeats? Please rewrite the
beginning of the paper more carefully.

Answer: They mean the same thing. We previously clarified that.

p.1, line -9: maybe define runs (at least intuitively): maximal periodic factors. Also, define the notion of border that you use few lines below.

Answer: Done.

p.1, line -2: replace "according" by "with respect to"

Answer: Done.

p.2, line 3,4: is the first attempt done in two papers?

Answer: We changed this citing only the conference version of the paper, where this notion was introduced.

p.2, paragraph 2: indeed, if you consider words of length n that also contain squares, you get n to be an upper bound on the number of MEFs. How tight is this bound? Can you produce examples that have a number of MEFs close to n, even if they have exponent greater than 2?

Answer: Following the model of [2] we focus here only on factors of a lower exponent. However, the question posed by the referee might deserve further investigation.

p.2, line 16: fixe-gapped should be fix-gapped. In general in this paragraph: upper bounding the length of the gap (by a constant or with respect to the length of the arm of the repeat) seems to be a good idea towards upper bounding the number of gapped repeats. See, for instance, the reference:
Brodal, G.S., Lyngs, R.B., Pedersen, C.N.S., Stoye, J.: Finding maximal pairs with bounded gap. In: Proc. 10th Annual Symposium on Combinatorial Pattern Matching. Volume 1645 of LNCS., Springer (1999) 134-149

Answer: We fixed the notation. The second part of the referee's suggestion seems strongly related to the previous comment. Bounding the gap by a constant would not guarantee a MEF (according to our definition). However, bounding it with respect to the length of the arm is precisely what we did here.

p.2, paragraph about α-gapped repeats: say that

you are talking here about *maximal* repeats (i.e., repeats
whose arms cannot be extended simultaneously); otherwise
the O(\alpha n) bound does not hold.

Answer: Done

p.2, line 18: comma after "Eventually"

Answer: Done

p.2: why not make the preliminaries a standalone section?

Answer: We consider these to be too short as to constitute
a proper section by themselves.

Preliminaries:
p.2, line -3: maybe explain again (as in the end of page 1) who
is u and who is v. It is a good occasion to state that you are only
looking into square free words, so the longest border is shorter
than the period, such a decomposition of a MEF can be defined,
etc. I find it funny that you give an example for the length of a
word but you do not give an example for, e.g., the period or
exponent of a word, which seem notions harder to grasp to me.

Answer: We ended the section with an example. Furthermore,
we added an explanation regarding the relation between the
period and the border for the factors of square-free words.

Section 2:
p.3, line 9: I do not understand the first sentence of this
paragraph. Please rephrase.

Answer: We rephrased the sentence.

p.3: the style in which you cite [1] in the two lemmas is not uniform.
Please use the same way of citing.

Answer: Done

p.3, line -20: replace "i.e." with "e.g.".

Answer: Done

p.3, line -10: generally for every two words u and v there exist a q
such that |v|=q|u|, namely q=|v|/|u|. Remove this superfluous

remark and rewrite this paragraph.

Answer: We rephrased the sentence in order to express the fact
that for every word such a q is unique.

p.3, line -3: delete semicolon after 1.

Answer: Done

p.4, Lemma 4: I think it is a bit confusing to use now i as the
starting point of u? (so far it was the starting point of u) and j
as the starting point of u (so far it was the starting point of u?).

Answer: We kept i and j as the starting positions of the two
strings in the order they occur (i comes before j, alphabetically).

p.4, line -4: "..greater than *that* of.."

Answer: Done.

p.5, line 2: maybe rewrite 3r>2s as $r>\frac{2s}{3}$.

Answer: Done.

p.5, line 8: I would remove the comma before and

Answer: We prefer to keep the comma as to separate the
two conditions.

p.5, line 14: which is the first inequality?

Answer: We rephrased the sentence.

p.5, line -5: do you really mean enumerate or rather count?

Answer: We only count them.

Section 3:
p.6, proof of Lemma 9: explain briefly why the 2nd u' cannot
overlap the 3rd u?

Answer: We never claimed such a case is not possible. If ℓ
is much greater than m than this case is also possible. However,
this does not affect our result.

p.7, Fig. 4: this figure is very confusing. Why are some words thinner and do not have names? Maybe you should draw two figures replacing this one.

Answer: The top and bottom dashed words are the ones that start outside and are not that important in our argument. Hence we preferred to thin them as for the reader not to get confused.

p.7, line -15: do you mean only one MEF in general? or only one MEF with a certain border?

Answer: only one MEF of border length ℓ. We revised this.

p.8, line 1: comma after "Lemma 9", otherwise the subject and predicate of that sentence are separated by a comma, which is wrong.

Answer: Done

p.8, line 6: maybe add "and" before "we conclude"

Answer: no "and" is necessary here.

p.8, line 10: I would rewrite the sentence starting with "If there are two MEFs.." as "If there are two MEFs, with overlapping left borders, such that one has twice the border length of the other.."

Answer: Done

p.8, line -4: delete "then" from the end of the line.

Answer: Done

p.9, Example 13: the sentence starting with "Furthermore, because.. , thus.." sounds a bit weird to me. Maybe rephrase.

Answer: removed "thus"

p.9, Example 13: the connection to the counting of MEFs with certain border lengths should be clarified. I only see how you count MEFs of certain exponent.

Answer: The exponent in the example is fixed to 4/3. We obviously count all factors having such an exponent (which is maximal among all factors of the word - proved in the first part). The counting of MEFs of different border lengths is also done in the first part, where

we look at all MEFs with borders 1 and 2.

Section 4:
p.10, line 11: "the final stride towards *proving*". Further, I think that you should discuss about improving this upper bound only after you?ve shown it. At this point, improving an unknown bound know makes no sense to me.

Answer: We have changed our statement. However, note that an upper bound exists, following [1].

p.10, line -12: writing both "although" and "however" in that sentence is a bit too much.

Answer: Removed "however"

p.10, Remark 18: sometimes MEFs with |u|>|v| are called long armed repeats (see the works of Kolpakov and Kucherov)

Answer: No point in introducing such a new notion at this stage.

p.11, line 4: You really use a lot "Furthermore" in this paper. Maybe use other words instead sometimes.

Answer: We tried to fix this situation.

p.11, line 7: a simple arithmetic what? argument?

Answer: Added argument

p.11, line 10: I would rather say "calculation of this upper bound"

Answer: Done

p.11, line 16: replace "strain" by "constraint"

Answer: Done

p.11, line -18: Remove "However" from the beginning to that phrase.

Answer: Done

p.11, Theorem 20: I find this restriction highly artificial (compared, e.g., to the restriction in Lemma 19). I am not even sure that it is worth mentioning in this paper. Same for the

following paragraph.

Answer: We do not see why such a restriction is artificial.
It basically considers all cases where $|v|>|u|$, but not a
multiple. This therefore leaves only one case to be further
investigated, when $|v|$ is a multiple of $|u|$.

Section 5:
p.12, Prop. 21: I do not really like saying "contain asymptotically".
Maybe say that the ratio between the length of the prefix and its
MEFs tends to 5/6, or something like that (like you say in your proof).

Answer: Changed

p.13, the last two paragraph: I do not understand why you do not
formalise this intuitions or arguments. They seem right and the
current "hand-waving" presentation is not satisfactory.

Answer: Formalising these two paragraphs would imply another
few pages of proofs which would provide only a slight improvement
for the presented lower bound. We consider that these are not as
important.

References:
Ref. [4]: I guess Lyndon should be written with capital L.

Answer: Done

Ref. [20]: "Zeichenreihen" should be written with capital Z.
Substantives are always capitalised in German.

Answer: Done