On the sum of exponents of maximal repetitions in a word

Roman Kolpakov, Gregory Kucherov

On the sum of exponents of maximal repetitions in a word *

Roman Kolpakov

French-Russian Institute
for Informatics and Applied Mathematics,
Moscow University,
119899 Moscow, Russia
e-mail: roman@vertex.inria.msu.ru

Gregory Kucherov

INRIA-Lorraine/LORIA,
615, rue du Jardin Botanique,
54602 Villers-lès-Nancy,
France
e-mail: kucherov@loria.fr

Abstract

This paper continues the study presented in [KK98], where it was proved that the number of maximal repetitions in a word is linearly-bounded in the word length. Here we strengthen this result and prove that the sum of exponents of maximal repetitions is linearly-bounded too.

Similarly to [KK98], we first estimate the sum of exponents of maximal repetitions in Fibonacci words. Then we prove that the sum of exponents of all maximal repetitions in general words is linearly-bounded. Finally, some algorithmic applications of this results are discussed.

1 Introduction

This paper continues the study of maximal repetitions in words, presented in [KK98]. Given a word, a *repetition* is a subword such that its minimal period is no larger than a half of the subword length. The *exponent* of a repetition is the ratio of its length to its minimal period. A *maximal repetition*[1] is a repetition which cannot be extended in the word to the right or to the left without increasing the minimal period. The set of maximal repetitions encodes, in a compact way, the whole repetitive structure of the word. The concept of maximal repetition, in comparison with other definitions of a repetition in a word, has been discussed in [KK98].

A central result of [KK98] is that the number of maximal repetitions in words of length n over an arbitrary alphabet is linear in n. Here we reinforce this result, proving that not only their number is linear, but the sum of their exponents is linearly-bounded too.

Applying the same approach as in [KK98], we first estimate the sum of exponents of maximal repetitions in Fibonacci words, that have a rich combinatorial structure and are often used as test objects in word combinatorics and string matching [Cro81, IMS97]. In particular, Fibonacci words are known to have "many" repetitions all of which have a short exponent (smaller than $2 + \varphi$, φ is a golden ratio [MP92]). It is known, for example, that Fibonacci words contain $\Theta(n \log n)$ primitively-rooted squares (repetitions of exponent 2). In [KK98], we obtained an exact formula for the number of maximal repetitions in Fibonacci words. Specifically, n-th Fibonacci word f_n contains $2|f_{n-2}| - 3$ maximal repetitions. Here we estimate approximately the sum of maximal

*Part of this work has been done during the first author's visit of LORIA/INRIA-Lorraine supported by a grant from the French Ministry of Public Education and Research. The first author has been in part supported by the Russian Foundation of Fundamental Research, under grant 96–01–01068, and by the Russian Federal Programme "Integration", under grant 473. The work has been done within a joint project of the French-Russian A.M.Liapunov Institut of Applied Mathematics and Informatics at Moscow University

[1] called *run* in [IMS97], *maximal periodicity* in [Mai89], and *m-repetition* in [KK98]

repetitions in Fibonacci words and show that this quantity is asymptotically between $1.922 \cdot |f_n|$ and $1.926 \cdot |f_n|$. This suggest that the sum of maximal repetitions might be linear for general words too.

We then confirm this conjecture by showing that the proof of the main result of [KK98] can be modified in order to prove this stronger result. The result brings a better insight into combinatorial properties of repetitions in words. For example, it explains a trade-off between the number of repetitions in a word and their exponents. Intuitively, a word cannot contain "many" repetitions of big exponent. As another consequence, the result confirms the conjecture made in [SG98a] about the linear number of so-called "branching tandem repeats" in a word.

In the final part, we discuss some algorithmic applications of this result. A main algorithmic application of the result of [KK98] has been that the linearity of the number of maximal repetitions allowed us to infer an algorithm which finds them all in linear time. Having the set of all maximal repetitions allows then to extract all repetitions of any other type, typically in time $O(n+T)$, where T is the output size. For example, this implies that all branching tandem repeats can be found in linear time, as their number is linear. As an example of another application, we show that one can compute, in linear time, the number of (primitively- or non-primitively-rooted) repetitions of a given exponent k, starting at each position of the word. For non-primitively-rooted repetitions, a linear solution to this problem is implied by the result of this paper.

2 Definitions and basic results

Consider a word $w = a_1 \dots a_n$. Any word $a_i \dots a_j$ for $i \leq j$, which we denote $w[i..j]$, is a *subword* of w. A position in w is an integer number between 0 and n. Each position π in w defines a decomposition $w = w_1 w_2$ where $|w_1| = \pi$. The position of letter a_i in w is $i-1$. If $v = w[i..j]$, we denote $initpos(v) = i-1$ and $endpos(v) = j$. We say that subword $v = w[i..j]$ *crosses* a position π in w, if $i \leq \pi < j$.

If w is a subword of u^n for some natural n, $|u|$ is called a *period* of w, and word u is a *root* of w. Clearly, p is a period of $w = a_1 \dots a_n$ iff $a_i = a_{i+p}$ whenever $1 \leq i, i+p \leq n$. Another equivalent definition is (see [Lot83]): p is a period of $w = a_1 \dots a_n$ iff $w[1..n-p] = w[p+1..n]$. The last definition shows that each word w has the minimal period that we will denote $p(w)$ and call often simply *the* period of w. The ratio $\frac{|w|}{p(w)}$ is called the *exponent* of w and denoted $e(w)$. Clearly, a root u of w such that $|u| = p(w)$, is *primitive*, that is u cannot be written as v^n for $n \geq 2$. Following [Lot83, Chapter 8], we call the roots u with $|u| = p(w)$ *cyclic roots*. We also call the cyclic root $w[1..p(w)]$ the *prefix cyclic root* of w, and the cyclic root $w[n-p(w)+1..n]$ the *suffix cyclic root* of w.

Consider $w = a_1 \dots a_n$. A *repetition* in w is any subword $r = w[i..j]$ with $e(r) \geq 2$. A *maximal repetition* in w, called for short a *maximal repetition*, is a repetition $r = w[i..j]$ such that

(i) if $i > 1$, then $p(w[i-1..j]) > p(w[i..j])$,

(ii) if $j < n$, then $p(w[i..j+1]) > p(w[i..j])$.

In other words, a maximal repetition is a repetition $r = w[i..j]$ such that no subword of w which contains r as a proper subword has the same minimal period as r. Note that any repetition in a word can be extended to a unique maximal repetition, that we will call the *corresponding* maximal repetition. For example, the repetition 1010 in word $w = 1011010110110$ corresponds to the maximal repetition 10101 obtained by one letter extension to the right.

A basic result about periods is the Fine and Wilf's theorem (see [Lot83]):

Theorem 1 (Fine and Wilf) *If w has periods p_1, p_2, and $|w| \geq p_1 + p_2 - gcd(p_1, p_2)$, then $gcd(p_1, p_2)$ is also a period of w.*

The following Lemma states some useful facts about maximal repetitions that will be used in the sequel.

Lemma 1 *(i) Two distinct maximal repetitions with the same period p cannot have an overlap of length greater than or equal to p,*

(ii) Two maximal repetitions with minimal periods p_1, p_2, $p_1 \neq p_2$, cannot have an overlap of length greater than or equal to $2\max\{p_1, p_2\}$.

Proof: Part (i) is easily proved by analyzing relative positions of two repetitions of period p and showing that if they intersect on at least p letters, at least one of them is not maximal. Part (ii) is a consequence of Fine and Wilf's theorem. If the intersection is at least $(p_1 + p_2 - gcd(p_1, p_2))$ long, then at least one of the cyclic roots of the two repetitions is not primitive, which is a contradiction. □

A repetition r is said to have a period in some subword of w if r overlaps with this subword on at least $p(r)$ letters. Also, we say that a repetition r has a period on the right (respectively on the left) of a position π with the meaning that $w[\pi + 1..\pi + p(r)]$ (respectively $w[\pi - p(r) + 1..\pi]$) is a subword of r.

The following reformulation of Lemma 1(i) will be often used in Section 4.

Corollary 1 *If two maximal repetitions of w have a period on the right (on the left) of the same position π, then either these maximal repetitions have different periods or they coincide.*

We will also use the following known Proposition which is also a consequence of Fine and Wilf theorem.

Proposition 1 *If u is a primitive word, then u cannot be an internal subword of uu (that is, a subword which is not a prefix or suffix).*

Proof: If u is an internal subword of uu, then $u = v_1v_2 = v_2v_1$ where both v_1 and v_2 are non-empty. This means that both v_1 and v_2 are roots of u, that is both $|v_1|$ and $|v_2|$ are periods of u. Since $|u| \geq |v_1| + |v_2| - gcd(|v_1|, |v_2|)$, $gcd(|v_1|, |v_2|)$ is also a period of u, which implies that u is an integer power. This contradicts the condition that u is primitive. □

We conclude this section with the following simple Lemma which will be also often used in Section 4. $\#S$ denotes the cardinality of a set S.

Lemma 2 *Let $S \subseteq \mathbb{R}$ be a set of real numbers, and each number of S belongs to interval $[a, b]$ (that is, $S \subseteq [a, b]$). Assume that there is $\Delta \in \mathbb{R}$ such that for every $x, y \in S$, $|x - y| \geq \Delta$. Then $\#S \leq (b - a)/\Delta + 1$.*

All logarithms are binary unless the base is indicated.

3 The sum of exponents of maximal repetitions in Fibonacci words

Fibonacci words are words over the binary alphabet $\{0, 1\}$ defined recursively by $f_0 = 0$, $f_1 = 1$, $f_n = f_{n-1}f_{n-2}$ for $n \geq 2$. The length of f_n, denoted F_n, is the n-th Fibonacci number. Fibonacci

words have numerous interesting combinatorial properties and often provide a good example to test conjectures and analyze algorithms on words (cf [IMS97]).

Let us recall some known facts about repetitions in Fibonacci words. As it was noted in Introduction, Fibonacci word f_n contains $\Theta(F_n \log F_n)$ squares. All squares are primitively-rooted as Fibonacci words don't contain repetitions of exponent 4. More precisely, Fibonacci words contain no repetition of exponent greater than $2+\varphi \approx 3.618$ (φ is the golden ratio) but do contain repetitions of exponent greater than $2+\varphi-\varepsilon$ for every $\varepsilon > 0$ [MP92]. In [FS99], the exact number of squares in Fibonacci words has been obtained, which is asymptotically $\frac{2}{5}(3-\varphi)nF_n + O(F_n) \approx 0.7962 \cdot F_n \log F_n + O(F_n)$

In [FS99], it has been shown that Fibonacci word f_n contains $2(F_{n-2}-1) = 2(2-\varphi)F_n + o(1)$ *distinct* (= syntactically different) squares. It has been also proved that the number of distinct squares in general words of length n is bounded by $2n$ (for an arbitrary alphabet). It is conjectured that this number is actually smaller than n, at least for the binary alphabet. Thus, in contrast to square occurrences, the maximal number of distinct squares is linear.

In [KK98], using the approach of [FS99], we showed that the number of maximal repetition in Fibonacci words is $2F_{n-2}-3$ (i.e. one less than the number of distinct squares). Thus, Fibonacci words contain a linear number (in the word length) of maximal repetitions. This result, together with the fact that Fibonacci words don't contain exponents greater than $(2+\varphi)$, implies that the sum of exponents of all maximal repetitions in f_n is no greater, asymptotically, than $2(3-\varphi)F_n \approx 2.764 \cdot F_n$. We now obtain a more precise estimate.

Denote $SR(n)$ the sum of exponents of all maximal repetitions in Fibonacci word f_n. We prove the following estimation for $SR(n)$.

Theorem 2 *$SR(n) = C \cdot F_n + o(1)$, where $1.922 \le C \le 1.926$.*

We follow the general proof scheme used in [FS99] for counting the number of squares occurrences. Consider the decomposition $f_n = f_{n-1}f_{n-2}$ and call the position between f_{n-1} and f_{n-2} the *boundary*. Clearly, maximal repetitions in f_n are divided into those which lie entirely in f_{n-1} or f_{n-2} and those which cross the boundary, that is overlap with f_{n-1} and with f_{n-2}. We call such repetitions *crossing* repetitions of f_n and its overlaps with f_{n-1} and f_{n-2} the *left part* and the *right part* respectively. Our goal is to obtain a recurrence relation

$$SR(n) = SR(n-1) + SR(n-2) + cx(n). \tag{1}$$

To compute $cx(n)$, we note that the left and right part of a crossing repetition cannot be both of exponent ≥ 2, since Fibonacci words don't have subwords of exponent 4. If the left (respectively right) part is of exponent ≥ 2, then the exponent of this part has been already counted in $SR(n-1)$ (respectively $SR(n-2)$) in (1), and only the exponent of the right (respectively left) part should be counted in $cx(n)$. If both the left and the right part is of exponent < 2 (we call such a repetition a *composed* repetition), the exponent of the whole repetition should be counted in $cx(n)$.

Similar to [KK98], we analyze now the crossing repetitions. Consider the representation

$$f_n = f_{n-1}|f_{n-2} = f_{n-2}f_{n-3}|f_{n-3}f_{n-4} = f_{n-2}[f_{n-3}|f_{n-4}]f_{n-5}f_{n-4} \tag{2}$$

where $|$ denotes the boundary, $n \ge 5$, and square brackets delimit the occurrence of f_{n-2} with the same boundary as for the whole word f_n (we call it the *central occurrence* of f_{n-2}). The goal is to express $cx(n)$ through $cx(n-2)$ by relating crossing repetitions of the whole word f_n with those of the central occurrence of f_{n-2}.

It is known that every repetition in Fibonacci words has the period F_k for some k (see [Séé85, FS99]). Since $F_{n-3} > F_{n-4} > 2F_{n-6}$, it follows from (2) that if a crossing maximal repetition of f_n

has a period F_k for $k \leq n-6$, then it is also a crossing repetition of the central occurrence of f_{n-2} and therefore is accounted for in $cx(n-2)$. It remains to examine crossing maximal repetitions of f_n with periods F_{n-2}, F_{n-3}, F_{n-4}, F_{n-5}. An analysis of these repetitions has been done in [KK98], and we use the result of it here without giving full details, that can be found in [KK98].

There is one crossing repetition in f_n for each of the periods F_{n-2}, F_{n-3}, F_{n-4}, F_{n-5}. The repetition with period F_{n-2} is composed (both its left and right part is of exponent < 2), its length can be shown to be $F_n - 2 = F_{n-1} + F_{n-2} - 2$, and the exponent $\frac{F_{n-1}+F_{n-2}-2}{F_{n-2}}$. The repetition with period F_{n-3} is composed too, it is of length $2F_{n-3} + F_{n-4} = F_{n-2} + F_{n-3}$, and of exponent $\frac{F_{n-2}+F_{n-3}}{F_{n-3}}$. As for the crossing repetition with period F_{n-4}, it extends a repetition already present in the central occurrence of f_{n-2}. Its right part is of exponent < 2, and is inside the central occurrence of f_{n-2}, therefore it is accounted for in $cx(n-2)$, and it does not have to be added. Its left part is of exponent ≥ 2, and does not have to be counted in $cx(n)$. However, a part of it which is in the central occurrence of f_{n-2} (namely $f_{n-4}f_{n-5}$), is of exponent < 2, and therefore has been counted in $cx(n-2)$. We than have to subtract $\frac{F_{n-4}+F_{n-5}}{F_{n-4}} = \frac{F_{n-3}}{F_{n-4}}$. Similarly, the crossing repetition with period F_{n-5} has its left part which is already counted in $cx(n-2)$, and its right part which should not be counted, but the part of it of exponent $\frac{F_{n-5}+F_{n-6}}{F_{n-5}} = \frac{F_{n-4}}{F_{n-5}}$ has been counted in $cx(n-2)$ and should be subtracted. Putting everything together, we obtain the equation

$$cx(n) = cx(n-2) + 2 - 2/F_{n-2} + F_{n-1}/F_{n-2} + F_{n-2}/F_{n-3} - F_{n-3}/F_{n-4} - F_{n-4}/F_{n-5}, \qquad (3)$$

for $n \geq 8$. Transforming further this expression, we obtain

$$cx(n) = n - 1 - 2(1/F_{n-2} + 1/F_{n-4} + ... + 1/F_4 + 1/F_2) + F_{n-1}/F_{n-2} + F_{n-2}/F_{n-3}$$

for even $n \geq 8$, and

$$cx(n) = n + 1/2 - 2(1/F_{n-2} + 1/F_{n-4} + ... + 1/F_3 + 1/F_1) + F_{n-1}/F_{n-2} + F_{n-2}/F_{n-3}$$

for odd $n \geq 9$. To join the cases, we rewrite (1) into

$$SR(n) = 2SR(n-2) + SR(n-3) + cx(n) + cx(n-1) = 2SR(n-2) +$$
$$SR(n-3) + 2n - 3/2 - 2(\sum_{j=1}^{n-2} 1/F_j) + F_{n-1}/F_{n-2} + 2F_{n-2}/F_{n-3} + F_{n-3}/F_{n-4}.$$

The following estimation can be obtained using some elementary consideration.

$$-2(\sum_{j=1}^{n-1} 1/F_j) + F_n/F_{n-1} + 2F_{n-1}/F_{n-2} + F_{n-2}/F_{n-3} < 2,$$

for $n \geq 8$. We omit the proof. Using this estimation, we get that for all $n \geq 9$,

$$SR(n) \leq 2SR(n-2) + SR(n-3) + 2n + 1/2.$$

Solving this recurrence with initial conditions $SR(6) = 15\frac{11}{30}, SR(7) = 29\frac{27}{40}, SR(8) = 53\frac{142}{195}$, we obtain that

$$SR(n) \leq \frac{33}{520}(-1)^{n+1} + \frac{1}{\sqrt{5}}(40\frac{47}{130} - 25\frac{281}{1560}\bar{\varphi})\varphi^{n-6} - \frac{1}{\sqrt{5}}(40\frac{47}{130} - 25\frac{281}{1560}\varphi)\bar{\varphi}^{n-6}$$
$$-n - \frac{15}{4} < \frac{1}{\sqrt{5}}(40\frac{47}{130} - 25\frac{281}{1560}\bar{\varphi})\varphi^{n-6} \approx 1.926 \cdot F_n,$$

using the fact that $F_n \approx \varphi^{n+1}/\sqrt{5}$.

The lower bound can be obtained as follows. A direct calculation gives the values $SR(23) = 1.922328 \cdot F_{23}, SR(24) = 1.922520 \cdot F_{24}$. Then using an obvious inequality $SR(n) \geq SR(n-1) + SR(n-2)$, we get $SR(n) \geq 1.922328 \cdot F_n$. Theorem 2 is proved.

4 The sum of exponents of maximal repetitions in a word

Theorem 2 provides a good indication that a similar linearity result might hold for general words. Here we prove that indeed, the sum of exponents of all maximal repetitions in a word is bounded by a linear function on the length of the word. Let $R(w)$ be the set of all maximal repetitions in word w (over an arbitrary alphabet), and let $Sexp(w) = \sum_{r \in R(w)} e(r)$, $Sexp(n) = \max_{|w|=n} Sexp(w)$.

We prove the following main result.

Theorem 3 *$Sexp(n) = O(n)$.*

The proof of Theorem 3 follows very closely the proof from [KK98] of the linearity of the number of maximal repetitions in a word. The additional argument which allows to apply the proof of [KK98] to show this stronger result, is the following simple Lemma.

Lemma 3 *Let $w = w'w''$, and let CR be the set of those repetitions of $R(w)$ which cross the frontier between w' and w''. Then*

$$Sexp(w) < Sexp(w') + Sexp(w'') + 4 \cdot \#CR.$$

Proof: For every repetition $r \in CR$, denote r' its intersection with w', and r'' its intersection with w''. It is easy to see that the difference $Sexp(w) - (Sexp(w') + Sexp(w''))$ is

$$\sum_{\substack{r \in CR \\ |r'| < 2p(r)}} \frac{|r'|}{p(r)} + \sum_{\substack{r \in CR \\ |r''| < 2p(r)}} \frac{|r''|}{p(r)} < 2 \cdot \#CR + 2 \cdot \#CR = 4 \cdot \#CR.$$

The Lemma follows. □

Corollary 2 *Let $w = w_1 \dots w_k$, and let CR_i be the set of repetitions of $R(w)$ crossing the frontier between w_i and w_{i+1}, $i = 1, \dots k-1$. Then*

$$Sexp(w) < \sum_{i=1}^{k} Sexp(w_i) + 4 \cdot \sum_{i=1}^{k-1} \#CR_i.$$

The rest of the proof follows almost exactly the proof of [KK98] (with some minor simplifications). To prove Theorem 3, we prove the following stronger statement.

Theorem 4 *There exist absolute positive constants C_1, C_2 such that*

$$Sexp(n) \leq C_1 n - C_2 \sqrt{n} \log n \tag{4}$$

We assume, without loss of generality, that C_1 is sufficiently bigger than C_2, say $C_1 \geq 2C_2$, so that the function $C_1 x - C_2 \sqrt{x} \log x$ is monotonically increasing for all $x \geq 1$. In the proof, we use induction over n. For technical reasons we assume that $n \geq 81$ ($\sqrt{n} \geq 9$). The base cases ($n < 81$) are trivially satisfied, as constant C_1 can be chosen as large as we need.

Recall that $e(w)$ and $p(w)$ denote respectively the exponent and the period of w.

Take any $n \geq 81$, and a word $w = a_1 \dots a_n$ of length n. We split the proof into two major cases depending on whether or not w contains a maximal repetition of "big" exponent. Formally, denote $\mathcal{BR}$ the set of words w containing a maximal repetition of exponent $\lceil \sqrt{n} \rceil$ or more.

Case 1 ($w \notin \mathcal{BR}$): Let $w \notin \mathcal{BR}$. Write $w = w_1 w_2$, where $|w_1| = \lceil \frac{n}{2} \rceil$, $|w_2| = \lfloor \frac{n}{2} \rfloor$. Then by Lemma 3 $Sexp(w) < Sexp(w_1) + Sexp(w_2) + 4 \cdot \#CR(w)$, where $CR(w)$ is the set of repetitions of $R(w)$ crossing the frontier between w_1 and w_2. By induction,

$$\begin{aligned} Sexp(w_1) + Sexp(w_2) &\leq Sexp(\lceil \tfrac{n}{2} \rceil) + Sexp(\lfloor \tfrac{n}{2} \rfloor) \\ &\leq C_1 n - C_2(\sqrt{\lceil \tfrac{n}{2} \rceil} \log \lceil \tfrac{n}{2} \rceil + \sqrt{\lfloor \tfrac{n}{2} \rfloor} \log \lfloor \tfrac{n}{2} \rfloor). \end{aligned}$$

A routine calculation shows that

$$\begin{aligned} \sqrt{\lceil \tfrac{n}{2} \rceil} \log \lceil \tfrac{n}{2} \rceil + \sqrt{\lfloor \tfrac{n}{2} \rfloor} \log \lfloor \tfrac{n}{2} \rfloor \geq \sqrt{\tfrac{n+1}{2}} \log \tfrac{n+1}{2} + \sqrt{\tfrac{n-1}{2}} \log \tfrac{n-1}{2} \geq \\ \sqrt{2n} \log \tfrac{n}{2} - \tfrac{\log \frac{n}{2}}{2n\sqrt{n}}. \end{aligned}$$

Therefore,

$$Sexp(w_1) + Sexp(w_2) \leq C_1 n - C_2 \left(\sqrt{2} - \frac{1}{2n^2} \right) \sqrt{n} \log \frac{n}{2}. \tag{5}$$

We now turn to estimating $\#CR(w)$. Our goal is to prove that $\#CR(w) = O(\sqrt{n} \log n)$.

We decompose $CR(w) = CRL(w) \cup CRR(w)$, where $CRL(w)$ (respectively $CRR(w)$) are those maximal repetitions of $CR(w)$ which have a period in w_1 (respectively in w_2). Note that $CRL(w)$ and $CRR(w)$ are generally not disjoint.

Case 1.1 ($CRL(w)$): Consider $CRL(w)$. Observe that by Corollary 1, no two distinct repetitions from $CRL(w)$ have the same period. We further split $CRL(w)$ into two non-intersecting subsets, $CRL1(w)$ and $CRL2(w)$, where $CRL1(w)$ (respectively $CRL2(w)$) consists of those repetitions $r \in CRL(w)$ which have $p(r)/2$ or more (respectively, less than $p(r)/2$) letters in w_2.

Case 1.1.1 ($CRL1(w)$): Consider $r_1, r_2 \in CRL1(w)$ with periods $p(r_1)$, $p(r_2)$ respectively. By the remark above, $p(r_1) \neq p(r_2)$, and assume that $p(r_1) > p(r_2)$, and $\Delta = p(r_1) - p(r_2)$. Consider the (non-empty) word $v = w[\pi_b + 1..\pi_e]$, where $\pi_b = \max\{initpos(r_1) + p(r_1), initpos(r_2) + p(r_2)\}$ and $\pi_e = \min\{endpos(r_1), endpos(r_2)\}$. Observe that v is a subword of both r_1 and r_2 which occurs, in each of them, at least one period away from the beginning. Then v has two other occurrences at positions $\pi_b - p(r_1)$ and $\pi_b - p(r_2)$. Consider word $v' = w[\pi_b - p(r_1) + 1..\pi_b - p(r_2) + |v|]$. Observe that $|v'| = |v| + \Delta$, and v' has a period Δ, as v occurs both as a prefix and a suffix of v'. The situation is illustrated in Figure 1.[2] Since $w \notin \mathcal{BR}$, we can bound $\frac{|v'|}{\Delta} = \frac{|v|}{\Delta} + 1 \leq \lceil \sqrt{n} \rceil$. Since v contains the subword $w[\lceil \frac{n}{2} \rceil + 1..\pi_e]$ of length at least $p(r_2)/2$, we have $|v| \geq p(r_2)/2$. We then have $\frac{p(r_2)}{2\Delta} \leq \sqrt{n}$ which implies $\frac{p(r_2)}{p(r_1)} \leq 1 - \frac{1}{2\sqrt{n}+1}$. Turning to logarithms, $\log p(r_2) - \log p(r_1) \leq \log(1 - \frac{1}{2\sqrt{n}+1}) \leq -\frac{1}{2\sqrt{n}+1}$, as $\log(1-x) \leq -x$ for $0 \leq x < 1$. Therefore, $\log p(r_1) - \log p(r_2) \geq \frac{1}{2\sqrt{n}+1}$. Recall that each repetition r from $CRL1(w)$ has a distinct period $p(r)$ and hence a distinct value $\log p(r)$. On the other hand, $\log p(r)$ can vary from 0 to $(\log n - 1)$. By Lemma 2, there are at most $(\log n - 1)(2\sqrt{n} + 1) + 1 = O(\sqrt{n} \log n)$ distinct values $\log p(r)$, and therefore that many repetitions in $CRL1(w)$.

Case 1.1.2 ($CRL2(w)$): For $CRL2(w)$, the proof is exactly the same as in Case 1.1.1 except that here v contains the subword $w[\max\{endpos(r_1) - p(r_1), endpos(r_2) - p(r_2)\} + 1..\lceil \frac{n}{2} \rceil]$ of length at least $p(r_2)/2$ which implies that $|v| \geq p(r_2)/2$. Thus, $CRL2(w)$ contains at most $O(\sqrt{n} \log n)$ repetitions too.

[2] Note that Figure 1 depicts only one of possible situations. E.g., the end position of r_1 may be smaller than that of r_2, left occurrences of v may not intersect, etc.

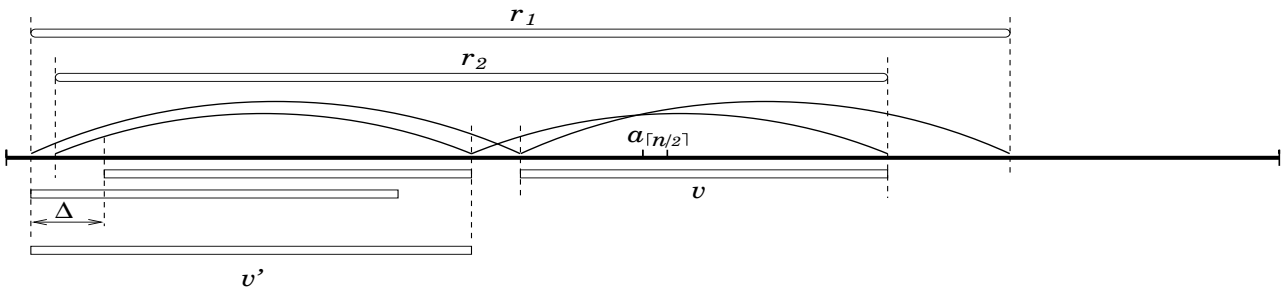

Figure 1: Illustration to Case 1.1.1

We obtain that $CRL(w)$ contains at most $O(\sqrt{n}\log n)$ repetitions.

Case 1.2 ($CRR(w)$): By symmetry, there are at most $O(\sqrt{n}\log n)$ repetitions in $CRR(w)$ too.

We conclude that $\#CR(w) \le \#CRL(w) + \#CRR(w) = O(\sqrt{n}\log n)$. Therefore,

$$Sexp(w) \le C_1 n - \left(\sqrt{2} - \frac{1}{2n^2}\right) C_2\sqrt{n}\log\frac{n}{2} + O(\sqrt{n}\log n).$$

To complete the induction step and verify that $Sexp(w)$ satisfies inequation (4), we have to make the expression $\left(\sqrt{2} - \frac{1}{2n^2}\right) C_2\sqrt{n}\log\frac{n}{2} - O(\sqrt{n}\log n)$ bigger than $C_2\sqrt{n}\log n$. This can be done by picking a sufficiently large constant C_2. The proof of Case 1 is completed.

Case 2 ($w \in \mathcal{BR}$): Let us now turn to the case $w \in \mathcal{BR}$. Let $w = w_1 r w_2$, where r is a maximal repetition in w with period $p_r = p(r)$ and exponent $e(r) = \frac{|r|}{p_r} \ge \lceil\sqrt{n}\rceil$. Denote $e_r = e(r)$. Note that $p(r) \le \sqrt{n}$, so $p(r) \le e(r)$. Since $n \ge 81$, then $e_r \ge 9$. Denote $\pi_{init} = initpos(r)$, $\pi_{end} = endpos(r)$. We now split r into three approximately equal parts, each having at least 3 periods. Formally, we find positions $\pi_{left} = \pi_{init} + \lfloor\frac{|r|}{3}\rfloor$, $\pi_{right} = \pi_{end} - \lfloor\frac{|r|}{3}\rfloor$. Denote by $w' = w[1..\pi_{left}]$, $w'' = w[\pi_{right}+1..n]$, and $r_0 = w[\pi_{left}+1..\pi_{right}]$. By Corollary 2,

$$Sexp(w) < Sexp(w') + Sexp(w'') + Sexp(r_0) + 4\cdot\#LR(w) + 4\cdot\#RR(w),$$

where

$LR(w)$: maximal repetitions, crossing position π_{left},

$RR(w)$: maximal repetitions, crossing position π_{right}.

The goal is now to estimate the cardinality of each of these classes.

Case 2.1 ($LR(w)$): Our goal is to prove that $\#LR(w) = O(e_r)$. Split the set $LR(w)$ into subset $SLR(w)$ of repetitions with a period smaller or equal to p_r, and subset $BLR(w)$ of repetitions with a period larger than p_r.

Case 2.1.1 ($SLR(w)$): If two repetitions from $SLR(w)$ have a period on the right (on the left) of π_{left}, then by Corollary 1, they cannot have the same period length. Therefore, each of these two subsets cannot have more than p_r distinct elements and there are no more than $2p_r$ overall maximal repetitions crossing π_{left}. So $\#SLR(w) \le 2p_r \le 2e(r) = O(e_r)$.

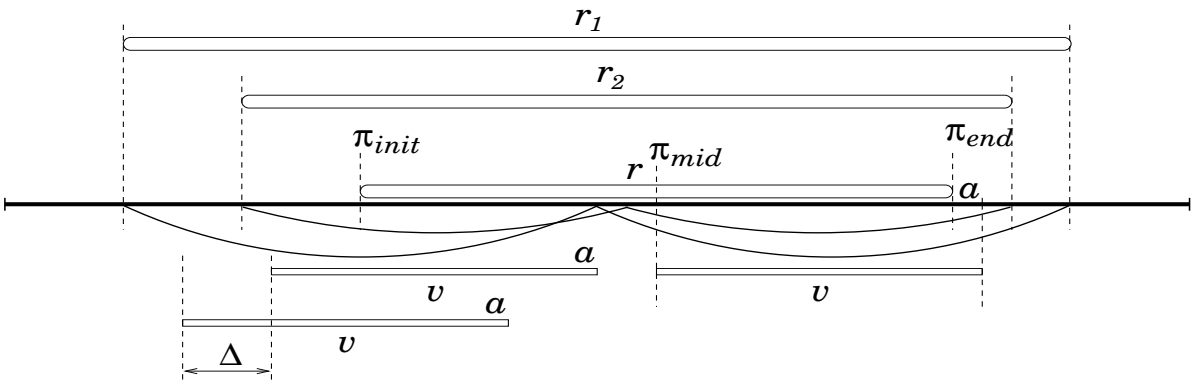

Figure 2: Illustration to Case 2.1.2.1.1

Case 2.1.2 ($BLR(w)$): The first observation is that repetitions of $BLR(w)$ cannot lie entirely inside r as this would contradict Lemma 1(ii). Thus, any repetition of $BLR(w)$ contains at least one of the letters $a_{\pi_{init}}, a_{\pi_{end}+1}$. We further split $BLR(w)$ according to different possibilities:

$BLR0(w)= \{u \in BLR(w) | initpos(u) < \pi_{init}$ and $endpos(w) > \pi_{end}\}$,

$BLR1(w)= \{u \in BCR(w) | initpos(u) \geq \pi_{init}$ and $endpos(w) > \pi_{end}\}$,

$BLR2(w)= \{u \in BLR(w) | initpos(u) < \pi_{init}$ and $endpos(w) \leq \pi_{end}\}$.

Then $\#BLR(w) = \#BLR0(w) + \#BLR1(w) + \#BLR2(w)$. We proceed by analyzing each of this classes separately.

Case 2.1.2.1 ($BLR0(w)$): Let us pick the position $\pi_{mid} = \pi_{init} + \lfloor |r|/2 \rfloor$ in the middle of r. We further divide $BLR0(w)$ into two (possibly intersecting) subsets. Let $BLR0'(w)$ consist of those repetitions of $BLR0(w)$ that have a period on the left of π_{mid}, and $BLR0''(w)$ of those that have a period on the right of π_{mid}.

Case 2.1.2.1.1 ($BLR0'(w)$): Consider two repetitions $r_1, r_2 \in BLR0'(w)$. By Corollary 1, $p(r_1) \neq p(r_2)$. Assume $p(r_1) > p(r_2)$. Consider the word $v = w[\pi_{mid} + 1..\pi_{end} + 1]$. Note that $a_{\pi_{end}+1}$ is the letter right after the end of repetition r, which implies that $a_{\pi_{end}+1} \neq a_{\pi_{end}+1-p_r}$. Note also that any proper prefix of v is a part of r and then has a period p_r. Word v belongs to both r_1 and r_2 and starts, in each of them, at least one period away from the beginning. Then v has two other occurrences starting at positions $\pi_{mid} - p(r_1)$ and $\pi_{mid} - p(r_2)$ (see Figure 2). The shift between these occurrences is $\Delta = p(r_1) - p(r_2)$ and we claim that $\Delta \geq |v| - p_r$. Otherwise, if $\Delta < |v| - p_r$, then the two occurrences of v have an overlap of length at least $p_r + 1$. Since this overlap is a prefix of the occurrence of v starting at $\pi_{mid} - p(r_2)$, it has a period p_r. Since the overlap is also a suffix of the occurrence of v starting at $\pi_{mid} - p(r_1)$ (see Figure 2), we have that $a_{\pi_{end}+1} = a_{\pi_{end}+1-p_r}$ which is a contradiction.

Thus, $\Delta = p(r_1) - p(r_2) \geq |v| - p_r \geq \frac{|r|}{2} - p_r$. As for any $u \in BLR0'(w)$, $p_r < p(u) \leq \frac{n}{2}$, by Lemma 2 we have that $BLR0'(w)$ contains at most $\frac{n/2-p_r}{|r|/2-p_r} + 1$ repetitions.

Case 2.1.2.1.2 ($BLR0''(w)$): This case is symmetric to Case 2.1.2.1.1. $BLR0''(w)$ contains at most $\frac{n/2-p_r}{|r|/2-p_r} + 1$ repetitions too.

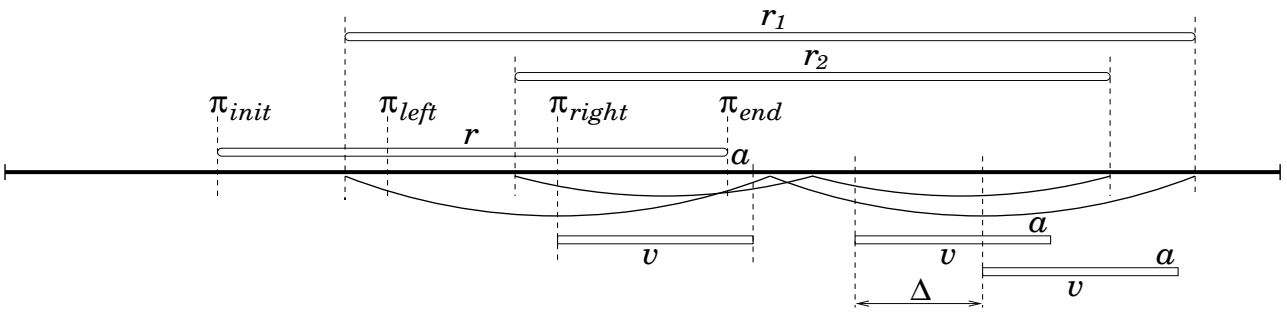

Figure 3: Illustration to Case 2.1.2.2.1

Summing up $\#BLR0'(w)$ and $\#BLR0''(w)$, we get $\#BLR0(w) \leq \frac{n-2p_r}{|r|/2-p_r} + 2 \leq \frac{n}{p_r(e_r/2-1)} + 2 \leq \frac{{e_r}^2}{e_r/2-1} + 2 = O(e_r)$.

Case 2.1.2.2 ($BLR1(w)$): Recall that $BLR1(w)$ consists of the maximal repetitions of period more than p_r, starting at a position between π_{init} and π_{left}, and ending at a position greater than π_{end}. We divide the repetitions of $BLR1(w)$ into two subsets according to their overlap with r. Let $BLR1'(w)$ be the set of those repetitions r' of $BLR1(w)$ which overlap with r on less than $p(r')$ letters, and $BLR1''(w)$ the set of those repetitions r' of $BCR1(w)$ which overlap with r on at least $p(r')$ letters.

Case 2.1.2.2.1 ($BLR1'(w)$): Clearly, any maximal repetition $u \in BLR1'(w)$ has at least $p(u)+1$ letters on the right of π_{end}. This implies, by Corollary 1, that no two repetitions of $BLR1'(w)$ have the same period. Consider two repetitions $r_1, r_2 \in BLR1'(w)$, and assume $p(r_1) > p(r_2)$. Consider the word $v = w[\pi_{right}+1..\pi_{end}+1]$ that is contained in both r_1 and r_2. Similar to Case 2.1.2.1.1, observe that any proper prefix of v has period p_r, and $a_{\pi_{end}+1} \neq a_{\pi_{end}+1-p_r}$. Since v is located in the prefix cyclic root of both r_1 and r_2, it has two copies in w at positions $\pi_{right}+p(r_2)$ and $\pi_{right}+p(r_1)$ (see Figure 3). The shift between these occurrences is $\Delta = p(r_1) - p(r_2)$. Similar to Case 2.1.2.1.1, if $\Delta < |v| - p_r$, we obtain the contradiction with $a_{\pi_{end}+1} \neq a_{\pi_{end}+1-p_r}$, and therefore we conclude that $\Delta \geq |v| - p_r$.

Since for $u \in BLR1'(w)$, $p_r < p(u) \leq (|r|+|w_2|)/2$, we conclude, by Lemma 2, that $BLR1'(w)$ contains at most $\frac{(|r|+|w_2|)/2-p_r}{|v|-p_r} + 1 \leq \frac{(|r|+|w_2|)/2}{|r|/3-p_r} + 1 \leq \frac{n}{2p_r(e_r/3-1)} + 1 \leq \frac{{e_r}^2}{2e_r/3-2} + 1 = O(e_r)$ repetitions.

Case 2.1.2.2.2 ($BLR1''(w)$): $BLR1''(w)$ consists of those repetitions of $BLR1(w)$ which have their prefix cyclic root completely inside r. First, show that for any repetition $r' \in BLR1''(w)$, its overlap with r is smaller than $p(r')+p_r$. Indeed, let $\pi_b = initpos(r')$, and consider $u = w[\pi_b+1..\pi_{end}]$ – the overlap of r' and r. Assume $|u| \geq p(r')+p_r$. Observe that $p(r') \neq p(r)$ as this would contradict Lemma 1(ii). Thus, $p(r') > p(r)$. By Fine and Wilf's theorem, u has a period $gcd(p(r'), p_r)$, and this implies that a root of r' of length $p(r')$ is not primitive. This contradicts the fact that $p(r')$ is the minimal period of r'. Thus, the end position π_{end} of r is less than p_r letters away from the end position of the prefix cyclic root of r'.

Consider the word $v = w[\pi_{right}+1..\pi_{right}+2p_r]$. By construction, $\pi_{end} - \pi_{right} \geq 3p_r$, and therefore by the above paragraph, v occurs inside the prefix cyclic root of both r_1 and r_2. Consider

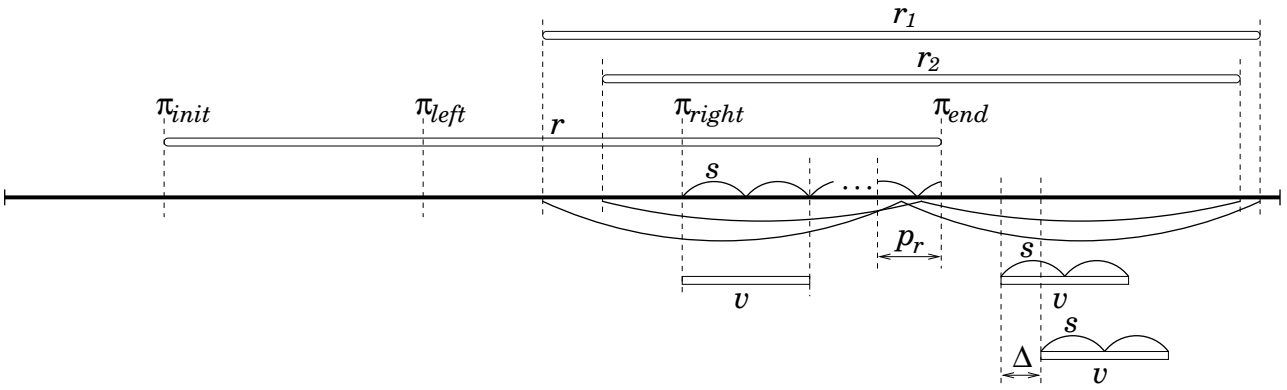

Figure 4: Illustration to Case 2.1.2.2.2

now $r_1, r_2 \in BLR1''(w)$. By Corollary 1, $p(r_1) \neq p(r_2)$, and assume that $p(r_1) > p(r_2)$. Again, v has two copies at positions $\pi_{right} + p(r_2)$ and $\pi_{right} + p(r_1)$ (see Figure 4). Note that $v = s^2$, where s is a cyclic root of r. Now if the two copies of v have an offset smaller than p_r (see Figure 4), then s occurs as a proper subword of $v = s^2$ which is impossible by Proposition 1 as s primitive. Therefore, $\Delta = p(r_1) - p(r_2) \geq p_r$.

As for any $u \in BLR1''(w)$, $2|r|/3 - p_r \leq \pi_{end} - \pi_{left} - p_r + 1 \leq p(u) \leq |r|$, by Lemma 2, $BLR1''(w)$ contains at most $\frac{|r|/3+p_r}{p_r} + 1 = \frac{|r|}{3p_r} + 2 = O(e_r)$ repetitions.

In total, $BLR1(w)$ has $O(e_r)$ maximal repetitions.

Case 2.1.2.3 ($BLR2(w)$): Similarly to Case 2.1.2.2, we split $BLR2(w)$ into set $BLR2'$ of maximal repetitions which don't have a period in r, and set $BLR2''$ of maximal repetitions which have a period in r.

Case 2.1.2.3.1 ($BLR2'(w)$): This case is symmetric to Case 2.1.2.2.1. Thus, $BLR2'(w)$ contains $O(e_r)$ repetitions too.

Case 2.1.2.3.2 ($BLR2''(w)$): Similarly to Case 2.1.2.2.2, we obtain that $BLR2''(w)$ contains at most $\frac{2|r|/3+p_r}{p_r} + 1 = \frac{2|r|}{3p_r} + 2 = O(e_r)$ repetitions.

Summing up $\#BLR2'(w)$ and $\#BLR2''(w)$, we get $\#BLR2 = O(e_r)$.

Putting together cases 2.1.2.1, 2.1.2.2, 2.1.2.3 and 2.1.1, we conclude that $LR(w)$ contains $O(e_r)$ maximal repetitions overall.

Case 2.2 ($RR(w)$): By symmetry, $RR(w)$ contains also at most $O(e_r)$ repetitions.

Thus,

$$Sexp(w) < Sexp(w') + Sexp(w'') + Sexp(r_0) + O(e_r), \tag{6}$$

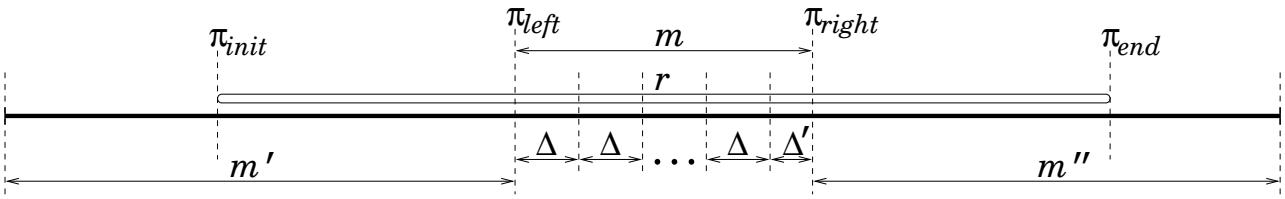

Figure 5: Estimation of $Sexp(r_0)$

Case 2.3 ($Sexp(r_0)$) Let us now turn to estimating $Sexp(r_0)$. Note that since r_0 is a repetition with period p_r, by Lemma 1(ii), r_0 cannot contain repetitions with a period larger than p_r. We introduce $m = \pi_{right} - \pi_{left}$ (the length of r_0), and $e_{mid} = m/p_r$ (the exponent of r_0).

We now split r_0 into consecutive blocks of length Δ (that will be defined later) with a possible remainder block of length $\Delta' < \Delta$ (see Figure 5). There are then $k = (m - \Delta')/\Delta$ blocks $u_1, \ldots, u_k$ of length Δ and one block u_{k+1} of length Δ'. Then by Corollary 2,

$$\begin{aligned} Sexp(r_0) &< \sum_{i=1,\ldots k+1} Sexp(u_i) + 4 \cdot \sum_{i=1,\ldots k} \#CR_i \\ &\leq kSexp(\Delta) + Sexp(\Delta') + 4 \cdot \sum_{i=1,\ldots k} \#CR_i, \end{aligned}$$

where CR_i is the set of all repetitions in r_0 which cross the boundary between blocks u_i and u_{i+1}, $i = 1, \ldots k$. Since r_0 has no repetitions of period greater than p_r, for any fixed position π in r_0, there are at most $2p_r$ repetitions in r_0, crossing this position. This can be seen using our usual technique: If two repetitions in r_0 have one period on the right (on the left) of π, they cannot have the same period length. Therefore, each of these two subsets cannot have more than p_r distinct repetitions and there are no more than $2p_r$ overall repetitions crossing π. Thus, $\#CR_i \leq 2p_r$ for any $i = 1, \ldots k$, and

$$Sexp(r_0) < kSexp(\Delta) + Sexp(\Delta') + 8kp_r. \tag{7}$$

By induction hypothesis, $Sexp(\Delta) \leq C_1\Delta - C_2\sqrt{\Delta}\log\Delta$, and $Sexp(\Delta') \leq C_1\Delta' - C_2\sqrt{\Delta'}\log\Delta'$. We then have the estimate

$$\begin{aligned} Sexp(r_0) &\leq k(C_1\Delta - C_2\sqrt{\Delta}\log\Delta) + C_1\Delta' - C_2\sqrt{\Delta'}\log\Delta' + 8kp_r \\ &= C_1 m - C_2(k\sqrt{\Delta}\log\Delta + \sqrt{\Delta'}\log\Delta') + 8kp_r \\ &= C_1 m - C_2\frac{m}{\Delta}\sqrt{\Delta}\log\Delta + C_2\left(\frac{\Delta'\log\Delta}{\sqrt{\Delta}} - \sqrt{\Delta'}\log\Delta'\right) + 8kp_r \end{aligned}$$

Considering the expression in parentheses as a real function of Δ' defined on the interval $[1..\Delta]$, we observe that this function is convex, and then reaches its maximum on one of the ends of the interval. We then conclude that $\frac{\Delta'\log\Delta}{\sqrt{\Delta}} - \sqrt{\Delta'}\log\Delta'$ reaches the maximum for $\Delta' = 1$ and is then bounded from above by $\frac{\log\Delta}{\sqrt{\Delta}} \leq \frac{2\log e}{e} \leq \frac{3}{2}$. We then get

$$Sexp(r_0) \leq C_1 m - C_2 m\frac{\log\Delta}{\sqrt{\Delta}} + \frac{3}{2}C_2 + 8kp_r. \tag{8}$$

Putting everything together. We now count together all the values $Sexp(w')$,$Sexp(w'')$,$Sexp(r_0)$. Recall that according to (6), our goal is to prove

$$Sexp(w') + Sexp(w'') + Sexp(r_0) + O(e_r) \leq C_1 n - C_2\sqrt{n}\log n. \tag{9}$$

which would conclude the induction argument.

The numbers $Sexp(w')$, $Sexp(w'')$ will be estimated by induction. Denote $m' = \pi_{left}$ (the length of w'), $m'' = n - \pi_{right}$ (the length of w''). Thus, $m' + m + m'' = n$. By induction, we have

$$\begin{aligned} Sexp(w') &\leq C_1 m' - C_2\sqrt{m'}\log m' & (10)\\ Sexp(w'') &\leq C_1 m'' - C_2\sqrt{m''}\log m'' & (11) \end{aligned}$$

Substituting (10), (11) into (9), we have

$$C_1 m' - C_2\sqrt{m'}\log m' + C_1 m'' - C_2\sqrt{m''}\log m'' + Sexp(r_0) + O(e_r) \leq C_1 n - C_2\sqrt{n}\log n,$$

or, since $n - m' - m'' = m$,

$$C_2(\sqrt{n}\log n - \sqrt{m'}\log m' - \sqrt{m''}\log m'') + Sexp(r_0) + O(e_r) \leq C_1 m \tag{12}$$

Let us now estimate the expression in parentheses in the left-hand side. The following lemma follows from routine calculus considerations.

Lemma 4

$$\sqrt{n}\log n - \sqrt{m'}\log m' - \sqrt{m''}\log m'' \leq \frac{m(\log n + 1)}{\sqrt{n}}$$

Proof: The proof is given in Appendix A. □

From Lemma 4 it follows that to prove (12) it is sufficient to prove the following inequality.

$$C_2\frac{m(\log n + 1)}{\sqrt{n}} + Sexp(r_0) + O(e_r) \leq C_1 m. \tag{13}$$

Note that by construction, $e_r \leq 3e_{mid}$, and we can then rewrite (13) into

$$C_2\frac{m(\log n + 1)}{\sqrt{n}} + O(e_{mid}) + Sexp(r_0) \leq C_1 m. \tag{14}$$

Completing the proof for $p_r \geq 8$. To finish the proof, we have to prove inequality (14) using estimate (8) for $Sexp(r_0)$. For technical reasons we now assume that $p_r \geq 8$ (the case $p_r < 8$ will be considered separately), and we choose $\Delta = \lfloor \frac{{p_r}^2}{4} \rfloor$. With this choice of Δ, inequation (8) can be transformed as follows. First note that $\frac{\log x}{\sqrt{x}}$ is decreasing for $x \geq e^2$, and $\Delta = \lfloor \frac{{p_r}^2}{4} \rfloor \geq 16$ as $p_r \geq 8$. Therefore, $\frac{\log \Delta}{\sqrt{\Delta}} \geq \frac{\log \frac{p_r^2}{4}}{\sqrt{\frac{p_r^2}{4}}} = \frac{4\log p_r - 4}{p_r}$. Turn now to the term $8kp_r$ in (8). Since $k = O(\frac{m}{p_r^2}) = O(\frac{e_{mid}}{p_r})$, then $p_r k = O(e_{mid})$, and we estimate $8kp_r$ as $O(e_{mid})$. We then rewrite (8) as

$$Sexp(r_0) \leq C_1 m - 4C_2 e_{mid}(\log p_r - 1) + \frac{3}{2}C_2 + O(e_{mid}) \tag{15}$$

We further estimate $\frac{m(\log n+1)}{\sqrt{n}}$ in (14). The term $\frac{m\log n}{\sqrt{n}}$ is estimated as follows using the fact that $\frac{\log x}{\sqrt{x}}$ is decreasing for $x \geq e^2$. Since $p_r^2 \leq n$, we then have $\frac{m\log n}{\sqrt{n}} \leq m\frac{2\log p_r}{p_r} = 2e_{mid}\log p_r$. The term $\frac{m}{\sqrt{n}}$ can be simply estimated by $\frac{m}{\sqrt{n}} = e_{mid}\frac{p_r}{\sqrt{n}} \leq e_{mid}$. Therefore,

$$\frac{m(\log n+1)}{\sqrt{n}} \leq e_{mid}(2\log p_r + 1),$$

and to prove (14), it then suffices to prove

$$C_2 e_{mid}(2\log p_r + 1) + O(e_{mid}) + Sexp(r_0) \leq C_1 m. \tag{16}$$

Substituting (15) into (16), we have the inequality

$$C_2 e_{mid}(2\log p_r + 1) + C_1 m - 4C_2 e_{mid}(\log p_r - 1) + \frac{3}{2}C_2 + O(e_{mid}) \leq C_1 m,$$

or

$$O(e_{mid}) \leq C_2 e_{mid}(2\log p_r - 5) - \frac{3}{2}C_2. \tag{17}$$

Recalling that $\log p_r \geq 3$ and $e_{mid} \geq 3$, inequation (17) can be satisfied by choosing a sufficiently large constant C_2.

Completing the proof for $p_r < 8$. The above analysis used the assumption $p_r \geq 8$, and to complete the proof, we are left with analyzing $Sexp(r_0)$ for $p_r < 8$.

If $p_r < 8$, we split r_0 into blocks of length $\Delta = p_r$ (instead of $\Delta = \lfloor \frac{{p_r}^2}{4} \rfloor$ as in the case $p_r \geq 8$). Then from (7), we get

$$\begin{aligned} Sexp(r_0) &< (k+1)Sexp(\Delta) + 8kp_r = (k+1)Sexp(p_r) + 8kp_r \\ &\leq Sexp(7)(k+1) + 56k = O(k). \end{aligned}$$

Since now $k = \lfloor \frac{m}{\Delta} \rfloor \leq \frac{m}{p_r} = e_{mid}$, then we have $Sexp(r_0) = O(e_{mid})$. We have then to prove

$$C_2\frac{m(\log n+1)}{\sqrt{n}} + O(e_{mid}) \leq C_1 m.$$

We also need to estimate $\frac{m(\log n+1)}{\sqrt{n}}$ differently. We simply note that $\frac{\log n+1}{\sqrt{n}} \leq \frac{5}{6}$ for $n \geq 81$, and then $\frac{m(\log n+1)}{\sqrt{n}} \leq \frac{5}{6}m$. So we have

$$\frac{5}{6}C_2 m + O(e_{mid}) \leq C_1 m.$$

Since we assumed $C_1 \geq 2C_2$ (see the beginning of the proof) and by definition $m \geq e_{mid}$, an appropriate choice of C_2 makes hold the above inequality.

The case analysis is exhausted. Choosing a sufficiently large constant C_2 which satisfies all the cases above allows to satisfy inequality (9). The proof of Theorem 4 is completed. Theorem 3 follows.

As a corollary of Theorem 3, we obtain the main result of [KK98] asserting that the number of maximal repetitions in a word over arbitrary alphabet is linearly-bounded in the length of the word.

Theorem 5

$$\max_{|w|=n} \#R(w) = O(n)$$

5 Some consequences and algorithmic applications

From the combinatorial perspective, Theorem 3 provides a better understanding of properties of repetitions in the word. For example, it explains a trade-off, observed in [KK98], between the number of repetitions and their exponents : informally, if a word contains many maximal repetitions, they are typically of short exponent. Fibonacci words provide a typical example of this situation.

A major algorithmic application of Theorem 5 proved in [KK98], has been a proof that all the maximal repetitions in a word can be found in linear time. This was achieved by an algorithm which is a modification of Main's linear-time algorithm [Mai89] for finding all leftmost occurrences of *distinct* maximal repetitions. Once the set of all maximal repetitions is found, it provides exhaustive information about the repetitive structure of the word. It allows easily to extract all repetitions of other types, such as (primitively- or non-primitively-rooted) squares, cubes, or integer powers. Thus, all these tasks can be done in time $O(n+T)$ where T is the output size (cf also [Kos94, SG98b]).

Theorem 3 provides new applications to the above method. One of them concerns the notion of *branching tandem repeats* studied in [SG98a]. In our terminology, branching tandem repeats are (non necessarily primitively-rooted) square suffixes of maximal repetitions. In [SG98a], the authors conjecture that the maximal number of branching tandem repeats in a word is linearly-bounded in the length. Our Theorem 3 confirms that conjecture, since each maximal repetition r contains $\lfloor e(r)/2 \rfloor$ branching tandem repeats, and therefore their total number, in any word w, is at most half of $Sexp(w)$. Clearly, the set of maximal repetitions allows to extract all branching tandem repeats. Since their number is linear, finding all branching tandem repeats takes linear time.

As another application, the set of maximal repetitions allows to determine, in linear time, the number of (primitively-rooted) integer powers of a given exponent k, starting at each position of the word. Here is how this can be done. For each position $i \in 1..|w|$, we create two counters $c^b(i)$ and $c^e(i)$, initially set to 0. For each repetition $r = w[m..l]$, we increment $c^b(m)$ and $c^e(l-kp(r)+1)$ by 1 ($[m..l-kp(r)+1]$ is the interval, where primitively-rooted k-powers induced by repetition r start). By Theorem 5, the number of updates is linear. To compute the numbers $d^k(i)$ of k-powers starting at each character i, we scan all characters from left to right applying the following iterative procedure: $d^k(1) = c^b(1)$, $d^k(i+1) = d^k(i) + c^b(i) - c^e(i-1)$, $i = 2..|w|$. Note that the algorithm can be extended to all (not necessarily primitively-rooted) k-powers. In this case, we increment $c^b(m)$ by $\lfloor e(r)/k \rfloor$, and we increment by 1 $c^e(j)$, for each $j = l-kp(r)+1, l-2kp(r)+1, \ldots, l-\lfloor e(r)/k \rfloor kp(r)+1$. Here, Theorem 3 guarantees that the number of updates is linear. Finally, note that the procedure can be easily modified in order to count arbitrary (non-integer) repetitions of given exponent, as well as repetitions ending (or centered) at each position.

6 Conclusions

The main drawback of our proof of Theorem 3 is that it does not allow to extract a "reasonable" constant factor in the linear bound. It seems however that this constant factor is quite small. Computer experiments suggest that the number of maximal repetitions is actually smaller than n and the sum of their exponents smaller than $2n$, at least for the binary alphabet. It would be interesting to find a simpler proof of Theorems 3,5 implying a small multiplicative constant in the linear bound.

References

[Cro81] M. Crochemore. An optimal algorithm for computing the repetitions in a word. *Information Processing Letters*, 12:244–250, 1981.

[FS99] A.S. Fraenkel and J. Simpson. The exact number of squares in Fibonacci words. *Theoretical Computer Science*, 218(1):83–94, 1999.

[IMS97] C.S. Iliopoulos, D. Moore, and W.F. Smyth. A characterization of the squares in a Fibonacci string. *Theoretical Computer Scince*, 172:281–291, 1997.

[KK98] R. Kolpakov and G. Kucherov. Maximal repetitions in words or how to find all squares in linear time. Rapport Interne 98-R-227, Laboratoire Lorrain de Recherche en Informatique et ses Applications, 1998. available from `http://www.loria.fr/~kucherov/res_activ.html`.

[Kos94] S. R. Kosaraju. Computation of squares in string. In M. Crochemore and D. Gusfield, editors, *Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, number 807 in Lecture Notes in Computer Science, pages 146–150. Springer Verlag, 1994.

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

[Mai89] M. G. Main. Detecting leftmost maximal periodicities. *Discrete Applied Mathematics*, 25:145–153, 1989.

[MP92] F. Mignosi and G. Pirillo. Repetitions in the Fibonacci infinite word. *RAIRO Theoretical Informatics and Applications*, 26(3):199–204, 1992.

[Séé85] P. Séébold. Propriétés combinatoires des mots infinis engendrés par certains morphismes. Rapport 85-16, LITP, Paris, 1985.

[SG98a] J. Stoye and D. Gusfield. Simple and flexible detection of contiguous repates using a suffix tree. In M. Farach-Colton, editor, *Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching*, number 1448 in Lecture Notes in Computer Science, pages 140–152. Springer Verlag, 1998.

[SG98b] J. Stoye and D. Gusfield. Linear time algorithms for finding and representing all the tandem repeats in a string. Technical Report CSE-98-4, Computer Science Department, University of California, Davis, 1998.

Appendix A

Here we prove Lemma 4 asserting that for $n \geq 81$ and $m, m', m'' \geq 1$ such that $m + m' + m'' = n$, we have

$$\sqrt{n}\log n - \sqrt{m'}\log m' - \sqrt{m''}\log m'' \leq \frac{m(\log n + 1)}{\sqrt{n}}. \tag{18}$$

Without loss of generality assume that $m' \geq m''$. Consider the function $f(x) = \sqrt{x}\log x$. $f(x)$ is positive ($f(x) \geq 0$) increasing ($f'(x) > 0$) and concave ($f''(x) < 0$) for $x \geq 1$. Then $f(m'') \geq$

$f(n) - f(n - m'')$ and therefore $f(n) - f(m') - f(m'') \leq f(n - m'') - f(m')$. This difference is maximal when m' is minimal, that is $m' = m'' = \frac{n-m}{2}$. We than have

$$f(n) - f(m') - f(m'') \leq f(\frac{n}{2} + \frac{m}{2}) - f(\frac{n}{2} - \frac{m}{2}) \tag{19}$$

The right-hand side of (19) is equal to

$$\frac{f^2(\frac{n}{2} + \frac{m}{2}) - f^2(\frac{n}{2} - \frac{m}{2})}{f(\frac{n}{2} + \frac{m}{2}) + f(\frac{n}{2} - \frac{m}{2})}. \tag{20}$$

Since $f(x)$ is concave for $x \geq 1$, this implies that for sufficiently large x, y (actually, for $x, y \geq e^2$) we have $f(x) + f(y) \geq f(x + y)$. Then, the denominator of (20) can be estimated from below as $f(\frac{n}{2} + \frac{m}{2}) + f(\frac{n}{2} - \frac{m}{2}) \geq f(n) = \sqrt{n} \log n$. The numerator of (20) is equal to

$$\int_{\frac{n}{2} - \frac{m}{2}}^{\frac{n}{2} + \frac{m}{2}} g(x)\, dx, \tag{21}$$

where $g(x) = (f^2(x))' = \log^2(x) + 2 \log e \log x$. It is easy to check that $g(x)$ is also positive, increasing and concave for $x \geq 1$. Then the estimation $\int_a^b g(x)\, dx \leq (b-a) g(\frac{b+a}{2})$ holds, and above integral can be estimated by $m g(\frac{n}{2}) = m(\log^2(\frac{n}{2}) + 2 \log e \log \frac{n}{2}) \leq m(\log^2 n + \log n)$. Substituting into (20) the estimates for the numerator and denominator, inequation (18) follows.