On closed-rich words*

Olga Parshina[1] and Svetlana Puzynina[1,2]

[1] Saint Petersburg State University, Russia
[2] Sobolev Institute of Mathematics, Russia
{parolja,s.puzynina}@gmail.com

Abstract. A word is called closed if it has a prefix which is also its suffix and there is no internal occurrences of this prefix in the word. In this paper we study the maximal number of closed factors in a word of length n. We show that it is quadratic and give lower and upper bounds for a constant.

Keywords: Closed word · return word · rich word.

1 Introduction

Various questions that concern counting factors of a specific form in a word of length n have been studied in combinatorics on words. Several studies have been devoted to the words that are extremal with respect to the proportion of factors with a given property. For example, an extensive study has been performed on the problem of counting the maximal repetitions (runs) in a word of length n. It has been shown in [17] that the maximal number of runs in a word is linear, and it was conjectured to be n. Subsequently, there was much research performed to find the bound [8]. Recently, the conjecture has been proved with a remarkably simple argument, considering numerous attempts to solve it [4]. We remark that questions about counting regular factors in a word are often non-trivial. For example, the problem of bounding the number of distinct squares in a string: A.S. Fraenkel and J. Simpson showed in 1998 [15] that a string of length n contains at most $2n$ distinct squares, and conjectured that the bound is actually n. After several improvements, the bound of $\frac{11}{6}n$ has been proved in [9], but the conjecture remains unsolved.

A related problem concerns counting palindromic factors. It is easy to see that a word of length n can contain at most $n + 1$ distinct palindromes (see e.g. [10]). Such words are called *rich in palindromes*, and there also exist infinite words such that all their factors are rich. Words rich in palindromes have been characterized in [16]. Words containing few palindromes were studied in [6, 14]. Recently some related questions about counting generalizations of palindromes have been studied, e.g. privileged factors [22] and k-abelian palindromes [7].

* The first author is supported by Ministry of Science and Higher Education of the Russian Federation, agreement 075–15–2019–1619. The second author is supported by Foundation for the Advancement of Theoretical Physics and Mathematics "BASIS".

We are interested in counting the factors that are called *closed*. A finite word is called *closed* if it has length ≤ 1 or it is a complete first return to some proper factor, i.e. it starts and ends with the same word that has no other occurrences but these two. Otherwise the word is called *open*. The terminology closed and open was introduced by G. Fici in [12]; for more information on closed words see [13]. The notion of closed word is actually the same as the notion of complete return word. The name return word is usually referred to factors of an infinite word and is used to study its properties. It can be regarded as a discrete analogue of the first return map in dynamical systems. For example, F. Durand characterized primitive substitutive words using the notion of a return word [11]. Return words also provide a nice characterization of the family of Sturmian words [24]. The explicit formulae for the functions of closed and open complexities for the family of Arnoux–Rauzy words, which encompass Sturmian words, were obtained in [21]. In [20], the authors prove a refinement of the Morse–Hedlund theorem (see [19]) providing a criterion of periodicity of an infinite word in terms of closed and open complexities.

The concept of closed factor has recently found applications in string algorithms. The *longest closed factor array* (LCF array) of a string x stores for every suffix of x the length of its longest closed prefix. It was introduced in [3] in connection with closed factorizations of a string. Among other things, the authors presented algorithms for the factorization of a given string into a sequence of longest closed factors and for computing the longest closed factor starting at every position in the string. In [5], the authors present the algorithm of reconstructing a string from its LCF array. See also [1] for some generalizations.

It is easy to show that each word of length n contains at least $n+1$ distinct closed factors [2]. In this paper, we study closed-rich words, i.e., words containing the maximal number of distinct closed factors among words of the same length. We prove an upper bound of $\sim \frac{3-\sqrt{5}}{4} n^2$ on this number (see Theorem $1'$), and we show that a word can contain $\sim \frac{n^2}{6}$ distinct closed factors (see Proposition 3). We also extend the notion of closed-rich words to infinite words, requiring that each factor contains a quadratic number of distinct closed factors. We find a sufficient condition on an infinite word to be closed-rich (see Proposition 5), and provide some families of infinite closed-rich words.

2 Preliminaries

Let $\mathbb{A}$ be a finite set called an alphabet. A finite or an infinite word $w = w_0 w_1 \cdots$ on $\mathbb{A}$ is a finite or infinite sequence of symbols from $\mathbb{A}$. For a finite word $w = w_0 \cdots w_{n-1}$, its *length* is $|w| = n$. We let ε denote the empty word, and we set $|\varepsilon| = 0$. A word v is a *factor* of a finite or an infinite word w if there exist words u and y such that w can be represented as their concatenation $w = uvy$. If $u = \varepsilon$, then v is a prefix, and if $y = \varepsilon$, then v is a suffix of w. If a finite word w has a proper prefix v which is also its suffix, then v is called a *border* of w. If the longest border of a word w occurs in w only twice (as a prefix and as a suffix),

then w is *closed*. By convention, if w is the empty word or a letter, then it is closed.

It is not hard to see that a word of length n contains at least $n+1$ distinct closed factors; G. Fici and Z. Lipták characterized words having exactly $n+1$ closed factors [2] . In the same paper they showed that there are words containing $\Theta(n^2)$ many distinct closed factors. The example they provided is a binary word with $\sim \frac{n^2}{32}$ closed factors. We say that a finite word w is *closed-rich* if it contains at least as many distinct closed factors as any other word of the same length and on the alphabet of the same cardinality.

If there exists an integer t such that for each i ($i < |w| - t$ in the case w is finite) we have $w_{i+t} = w_i$, then t is called a *period* of w. Let $s = \frac{|w|}{t}$ and let u be the prefix of w of length t. We say that w has *exponent* s and write $w = u^s$. The notation $w = u^{k+}$ means that w has exponent $s > k$ for an integer k. The word u is called the *fractional root* of w. The word w is *primitive* if its only integer exponent is 1. Hereinafter we always assume t to be the shortest period of w, and thus, s to be the largest exponent of w.

The following properties follow directly from the definitions.

Proposition 1. *Any word with exponent at least two is closed.*

Proposition 2. *Let w be a word of exponent 3 and of length n. Then all its factors of length at least $\frac{2n}{3}$ are closed, and moreover, all of them except for one of length $\frac{2n}{3}$ are unioccurrent.*

De Bruijn graph of order n on an alphabet $\mathbb{A}$ is the directed graph whose set of vertices (resp. edges) consists of all words over $\mathbb{A}$ of length n (resp. $n+1$). There is a directed edge from u to v labeled w if u is a prefix of w and v a suffix of w. We call a Hamiltonian path in this graph a *de Bruijn word*.

Proposition 3. *Let $n = 3 \cdot |\mathbb{A}|^k$ for an integer k, and v be a de Bruijn word of length $\frac{n}{3}$. Then $w = v^3$ has $\sim \frac{n^2}{6}$ distinct closed factors.*

Proof. Due to Proposition 2, all factors that are longer than $\frac{2n}{3}$ are closed and distinct (there are $\frac{n^2}{18}$ of those). All words of length $\frac{n}{3} + \log(\frac{n}{3}) \leq l \leq \frac{2n}{3}$ are also closed with corresponding border of length $l - n/3$ (there are $\sim (\frac{n}{3})^2$ of distinct factors of these lengths).

If a factor of length less than $\frac{n}{3} + \log(\frac{n}{3})$ is closed, then its border is shorter than $\log(\frac{n}{3})$, because all factors of de Bruijn word of length at least $\log(\frac{n}{3})$ are unioccurrent. Thus, there are not more than $\frac{n}{3} \cdot \log(\frac{n}{3})$ closed factors that are shorter than $\frac{n}{3} + \log(\frac{n}{3})$.

The construction from the previous proposition gives only words of length $n = 3 \cdot |\mathbb{A}|^k, k \geq 0$, but it could be easily modified to other lengths. For lengths n divisible by 3 we can e.g. take cubes of prefixes of de Bruijn words (we omit technical details here). Words of lengths not divisible by 3 can be obtained by shortening a word of next length divisible by 3 — clearly, a prefix of length 1 or 2 can add at most linear number of closed factors. However, if we change one

letter in the middle of a word, the total number of closed factors can change dramatically:

Example 1. Let us show that the number of closed factors can change from linear to quadratic when changing only one letter in a word. It is easy to see that the word $a^nba^nba^n$ has quadratic number of closed factors. After replacing the leftmost occurrence of b with a, we obtain the word $a^naa^nba^n$ with linear number of closed factors.

For a finite word w, we let $\mathrm{Cl}(w)$ denote the number of distinct closed factors of w. We now provide a trivial upper bound on $\mathrm{Cl}(w)$.

Proposition 4. *For each word w of length $n \geq 7$, one has $\mathrm{Cl}(w) \leq \frac{n^2}{4}$.*

Proof. For a word u and a letter a, let us denote by t the longest repeated suffix of ua, and z denote the longest repeated suffix of t. Clearly, the number of new closed factors ending in the last letter of ua is at most $|t|-|z| \leq \left\lfloor \frac{|ua|}{2} \right\rfloor$ when u is non-empty. For $n = 0$ we have one closed factor (the empty word), for $n = 1$ we add one closed factor (a letter). So, building w letter by letter, we get at most $2 + \sum_{i=2}^{n} \lfloor \frac{i}{2} \rfloor = 2 + \frac{n(n+1)}{4} - \lceil \frac{n}{2} \rceil$ closed factors in w. The claim follows.

In the next section we prove a tighter upper bound with the leading coefficient $\frac{3-\sqrt{5}}{4} \approx 0.19$. We believe it can be improved to $\frac{1}{6} = 0.1\overline{6}$.

3 Finite closed-rich words

The main goal of this section is to prove Theorem 1 providing an upper bound on the number of closed factors that a finite word can contain. We start with some auxiliary lemmas. Sometimes it will be convenient for us to consider cyclic words. For a normal word w, we can consider a corresponding cyclic word as the class of all its cyclic shifts. Then by a closed factor of a cyclic word we mean a closed factor of some its shift.

Lemma 1. *Let u be a primitive finite word of length k, then its cyclic square has at most k^2 distinct closed factors.*

Proof. Let $\hat{u}$ be the cyclic square of u with the first letter $\hat{u}_0 = u_0$. The following observations constitute the proof of the lemma. Basically, we count (left) borders giving rise to distinct closed words.

Each occurrence of a factor of $\hat{u}$ is a (left) border of at most one closed factor. Since $\hat{u}$ is a cyclic square, in order to count borders giving rise to distinct closed words, we can only consider the factors of $\hat{u}$ starting in its first half $\hat{u}_0 \cdots \hat{u}_{k-1} = u_0 \cdots u_{k-1}$ (borders starting in the second half of $\hat{u}$ give rise to the same closed words). The border of a closed factor of $\hat{u}$ cannot be longer than k, otherwise u is not a primitive word. The number of factors of $\hat{u}$ that start in its first half and are not longer than k is k^2. The statement follows.

Lemma 2. *Let w be a word with exponent at least 3, i.e., if we denote its period by k and its length by n, then $n \geq 3k$. Then the number of closed factors in the word is at most $k^2 + (n-3k)k + \frac{1}{2}(k+1)k$.*

Proof. We count closed factors by their lengths: k^2 is the upper bound for the number of closed factors of length at most $2k$ (given by the bound for a cyclic word of length $2k$ from Lemma 1); $(n-3k)k$ counts k closed factors of each length $2k+1, \ldots, n-k$; and $\frac{1}{2}(k+1)k$ counts long ones as the sum of an arithmetic progression (k factors for length $n-k+1$, $k-1$ for $n-k+1$, ..., and 1 for length n).

Corollary 1. *For words of length n and of exponent greater than 3, we have less than $\frac{n^2}{6}$ factors asymptotically.*

Proof. We estimate the bound from Lemma 2: for words with exponent $t \geq 3$ we count closed factors and get the function $c(t) = n^2(\frac{1}{t} - \frac{3}{2t^2})$. The maximum value of $c(t)$ is $\frac{n^2}{6}$ and it is achieved for $t = 3$.

We say that a finite word w' is a *cyclic shift* (or a conjugate) of a finite word w if there exist words u and v such that $w = uv$ and $w' = vu$.

Lemma 3. *Let w be a word of exponent $\alpha \geq 3$ and v its primitive root, so that $w = v^\alpha$. Then for any cyclic shift v' of v, the word v'^α contains the same number of closed factors as w.*

Proof. The proof follows from the counting in the proof of Lemma 2: the set of closed factors consists of all long factors (their numbers are clearly the same since we only take lengths into account) and short closed factors which are also closed factors of the cyclic square, so their sets are the same.

The following example shows that the statement of Lemma 3 does not hold for squares. Moreover, it is possible that taking a cyclic shift of the root changes the number of closed factors in a word from linear to quadratic.

Example 2. The number of closed factors in $a^{n/2}ba^nba^{n/2}$ is quadratic, while in its cyclic shift a^nba^nb it is linear.

The set of distinct closed factors of a word w can naturally be split into two sets, the set of closed words of length at least 2, which have border, and the set of closed words of length at most 1, i.e., letters and the empty word. We let $\mathrm{Cl}'(w)$ denote the number of closed words of length at least 2 ("long" closed factors), and $\mathrm{Cl}^0(w)$ denote the number of "short" closed factors, so that $\mathrm{Cl}(w) = \mathrm{Cl}'(w) + \mathrm{Cl}^0(w)$.

Now we can state the main result of this section.

Theorem 1. *For a finite word w of length n, the following holds:*

$$Cl'(w) < \frac{3-\sqrt{5}}{4}n^2 + \frac{\sqrt{5}-1}{4}n.$$

Clearly, since $\mathrm{Cl}(w) = \mathrm{Cl}'(w) + |\mathbb{A}| + 1 \leq \mathrm{Cl}'(w) + n + 1$, we can rewrite the statement of Theorem 1 as follows.

Theorem 1′ *For a finite word w of length n, the following holds:*

$$Cl(w) < \frac{3-\sqrt{5}}{4}n^2 + \frac{3+\sqrt{5}}{4}n + 1.$$

For a finite word w of length n, we denote its prefix of length $n-1$ by w^-. The following lemma constitutes the key part of the proof of Theorem 1.

Lemma 4. *Let w be a word of length n. If $\mathrm{Cl}'(w) - \mathrm{Cl}'(w^-) \geq Cn$ for some $\frac{1}{3} < C \leq \frac{1}{2}$, then $\mathrm{Cl}'(w) \leq \frac{(1-C)^2}{2}n^2 + \frac{1-C}{2}n$.*

Proof. In the proofs of the lemma and of Theorem 1, we only talk about long closed factors (of length at least 2). Let t denote the longest repeated suffix of w, z denote the longest repeated suffix of t, and c be the letter preceding the last occurrence of z in w, so that cz is the shortest unrepeated suffix of t. Clearly,

$$\mathrm{Cl}'(w) - \mathrm{Cl}'(w^-) \leq |t| - |z|. \tag{1}$$

Two cases are possible: the last and the penultimate occurrences of t in w might intersect, or not. If they intersect, we have a power as a suffix of w; let l be the period of this power. In this case we denote the suffix of t of length l by x. If they do not intersect, we set $x = t$.

There are several possibilities of how the word w can look like depending on whether the occurrences of t intersect or not and whether the penultimate occurrence of t starts from the beginning of the word or not. The case $w = x^{3+}$ is treated in Lemma 2 and is not considered here. Other possibilities follow.

1. $w = x^s$, $2 \leq s < 3$ where x is the shortest period of w;
2. $w = xvx$ for a non-empty word v;
3. $w = uxvx$ for non-empty words u, v;
4. $w = ux^s$ for $2 \leq s < 3$ and a non-empty word u.

Let us treat each case separately.

1. Let $w = x^s$, $2 \leq s < 3$. We use the following notation: $|x| = l$, $r = n - 2l$. In this case the longest border t ending in the last position of w is of length $r + l$. We should consider two subcases depending on the length of cz.

(a) The shortest unrepeated suffix cz of t occurs in w three times (see Fig. 1). Due to inequality (1), in this case $l \geq Cn$.

Let us count closed factors of w; it is easier to do by counting their borders. Let us count borders that are suffixes of the rightmost occurrences of closed factors of w. It is easy to see that every factor $w_i \cdots w_j$ for $i \geq l$ and $j \geq l + r$ is a border of the factor $w_{i-l} \dots w_j$. So, such border (suffix of a rightmost occurrence of a closed factor) can start at the earliest at position l. There are

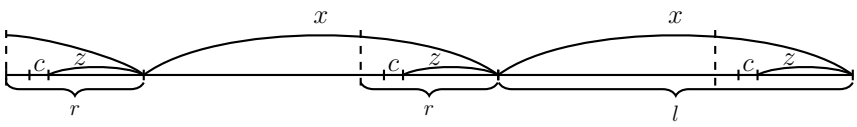

Fig. 1. The case $w = x^s, 2 \le s < 3, |w| = n = 2l + r, |cz| \le r$.

$\sum_{i=l}^{n-1}(n-i) - \sum_{i=l}^{l+r-1}(l+r-i)$ such borders in w. The second sum stands for borders occurring inside $w_l \dots w_{l+r-1}$; each one of them defines the same closed factor of w as the borders occurring in $w_{n-r} \dots w_{n-1}$. Since the word w is a power, borders appearing in $w_0 \dots w_{l-1}$ do not define any closed factors different from the ones already counted in the sum above. Hence,

$$\mathrm{Cl}'(w) \le \sum_{i=l}^{n-1}(n-i) \; - \sum_{i=l}^{l+r-1}(l+r-i) = \frac{1}{2}((n-l)(n-l+1) - (r+1)r)$$
$$= \frac{1}{2}((n-l)^2 + (n-l) - r^2 - r) = -\frac{3}{2}l^2 + \left(n + \frac{1}{2}\right)l.$$

The last equality is due to the equality $n - l = l + r$. This expression reaches its maximum when $l = \frac{1}{3}n + \frac{1}{6}$ and is equal to $\frac{1}{6}n^2 + \frac{1}{6}n + \frac{1}{24}$. Thus, in this case $\mathrm{Cl}'(w) \le \frac{1}{6}n^2 + \frac{1}{6}n + \frac{1}{24}$.

(b) The shortest unrepeated suffix cz of t occurs in w twice (Fig. 2).

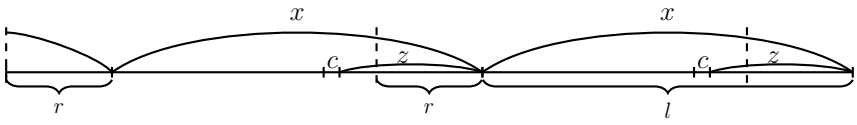

Fig. 2. The case $w = x^s, 2 \le s < 3, |w| = n = 2l + r, |cz| > r$.

Due to inequality (1), in this case $l + r - |z| \ge Cn$.

Let us count the borders that are suffixes of the rightmost occurrences of closed factors of w. All of them are located in the suffix of length $n - l - |z|$ of w. In a similar to 1.(a) way, using the relations $r = n - 2l$ and $|z| \le l + r - Cn$ we obtain the following.

$$\mathrm{Cl}'(w) \le \sum_{i=l+r-|z|}^{n-1}(n-i) - \sum_{i=l+r-|z|}^{n-l-1}(n-l-i) = \frac{l^2}{2} + l|z| + \frac{l}{2} \le -\frac{l^2}{2} + (1-C)nl + \frac{l}{2}.$$

This function reaches its maximum when $l = (1 - C)n + \frac{1}{2}$. Since we deal with integers in all inequalities, we obtain $\mathrm{Cl}'(w) \le \frac{(1-C)^2}{2}n^2 + \frac{(1-C)}{2}n$.

2. If $w = xvx$ for some word v (Fig. 3), inequality (1) gives $l - |z| \geq Cn$.

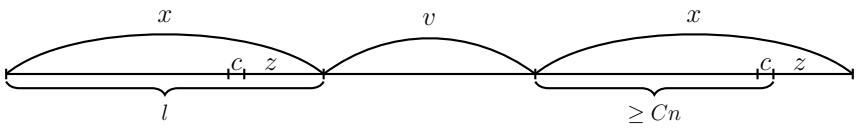

Fig. 3. The case $w = xvx, |w| = n$.

Let us count borders that are suffixes of the rightmost occurrences of closed factors of w. In fact, any such border cannot contain the first occurrence of cz in w (the one starting with $w_{l-|z|}$). Thus, we have the following inequality.

$$\mathrm{Cl}'(w) \leq \sum_{j=l-|z|}^{n-1} (n-j) = \frac{1}{2}(n-l+|z|)(n-l+|z|+1)$$
$$\leq \frac{1}{2}(n-Cn)(n-Cn+1) \leq \frac{(1-C)^2}{2}n^2 + \frac{(1-C)}{2}n.$$

3. Let $w = uxvx$ for non-empty words u, v (Fig. 4).

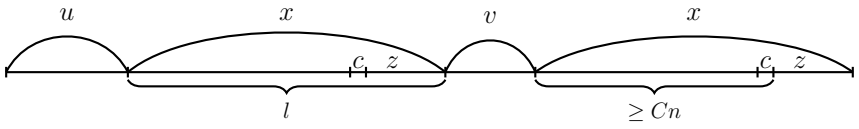

Fig. 4. The case $w = uxvx, |w| = n$.

Due to inequality (1), in this case $l \geq Cn + |z|$. We will count borders of closed factors that start in u, and the borders that are factors of xvx separately.

Let us show that the border of a closed factor starting in u cannot contain x as a factor. Suppose it is the case and consider the next occurrence of x in this closed word. It must end before the index $n - Cn$, otherwise its suffix cz is not unioccurrent in x. By the same reasoning it cannot start before $|u| + Cn$. The distance between these two points is $n - Cn - |u| - Cn = (1 - 2C)n - |u|$. Provided with $1/3 < C \leq 1/2$ we have $1 - 2C < C$. Thus, $(1 - 2C)n - |u| < Cn - |u| < Cn < l$, and x cannot be placed in the indicated gap.

We will make a more generous rounding up saying that the number of borders beginning in u is not greater than the number of factors in the prefix of w of length $n - Cn$. This number is $\sum_{j=0}^{|u|-1} (n - Cn - j) = n|u| - Cn|u| - \frac{|u|^2}{2} + \frac{|u|}{2}$.

The number of closed factors in xvx is not greater than $\sum_{i=|u|+l-|z|}^{n-1}(n-i) = \frac{1}{2}(n-|u|-l+|z|)(n-|u|-l+|z|+1) \le \frac{1}{2}(n-|u|-Cn)(n-|u|-Cn+1)$.

Summing the two expressions, we obtain $\mathrm{Cl}'(w) \le \frac{(1-C)^2}{2}n^2 + \frac{(1-C)}{2}n$.

4. Let $w = ux^s$, $2 < s < 3$. Analogously to the first case we should consider two situations, when the the shortest unrepeated suffix cz is shorter than r, and when it is longer than r.

(a) The shortest unrepeated suffix cz of t occurs in x^s three times.

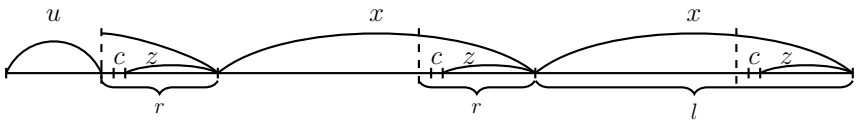

Fig. 5. The case $w = ux^s, 2 \le s < 3, |cz| \le r, |w| = n$.

Here $n = |u| + 2l + r$, and $l \ge Cn$.

The border of every closed factor that starts in u must end before the index $|u| + l - 1$, otherwise x would occur in x^2 three times, what contradicts the assumption on x to be the smallest period of x^s (see e.g. Problem 8.1.6. in [18]). Thus, the number of closed factors starting in u is not greater than the sum $\sum_{j=0}^{|u|-1}(|u| + l - 1 - j) = \frac{|u|^2}{2} - \frac{|u|}{2} + l|u|$.

We will count the rightmost occurrences of borders of closed words that are factors of x^s. It is enough to count the borders that end in the last occurrence of x in w. Again, the borders cannot start before the index $n - l - r$, otherwise x would occur in x^2 three times. Thus, the number of closed factors of x^s is not greater than $\sum_{j=n-l-r}^{n-1}(n-j) - \frac{r(r+1)}{2} = \frac{l^2}{2} + \frac{l}{2} + lr$.

Summing the two expressions, we get $\mathrm{Cl}'(w) \le \frac{|u|^2}{2} - \frac{|u|}{2} + l|u| + \frac{l^2}{2} + \frac{l}{2} + lr = l(|u| + 2l + r) - \frac{3}{2}l^2 + \frac{l}{2} - \frac{|u|}{2} + \frac{|u|^2}{2}$. Using the inequality $|u| \le n - 2l$, we obtain

$$\mathrm{Cl}'(w) < ln - \frac{3}{2}l^2 + \frac{l}{2} - \frac{n}{2} + l + \frac{n^2}{2} + 2l^2 - 2ln = \frac{(n-l)^2}{2} - \frac{n}{2} + \frac{3l}{2} < \frac{(1-C)^2}{2}n^2 + \frac{n}{4}.$$

(b) The shortest unrepeated suffix cz of t occurs in x^s twice (Fig. 6).

In this case $l + r - |z| \ge Cn$. Since $n = |u| + 2l + r$, we have $n - l - |u| - |z| \ge Cn$, and thus,

$$l + |u| + |z| \le (1 - C)n. \qquad (*)$$

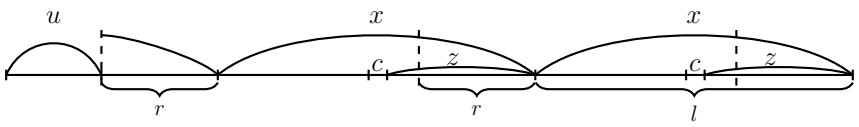

Fig. 6. The case $w = ux^s, 2 \le s < 3, |w| = n$.

As in the case 4.(a), the number of closed factors starting in u is

$$\sum_{j=0}^{|u|-1} (|u| + l - j - 1) = \frac{|u|^2}{2} + l|u| - \frac{|u|}{2}.$$

The number of closed factors in the suffix x^s is less than

$$\sum_{j=n-l-|z|}^{n-1} (n-j) - \frac{|z|(|z|+1)}{2} = \frac{l^2}{2} + l|z| + \frac{l}{2}.$$

Thus, the number of closed factors in this case is

$$\mathrm{Cl}'(w) \le \frac{|u|^2}{2} + l|u| - \frac{|u|}{2} + \frac{l^2}{2} + l|z| + \frac{l}{2} = l(l + |u| + |z|) - \frac{l^2}{2} + \frac{|u|^2}{2} + \frac{l}{2} - \frac{|u|}{2}.$$

Using $(*)$ and $|u| \le n - 2l$ we obtain

$$\begin{aligned}\mathrm{Cl}'(w) &\le (1-C)nl + \frac{3l^2}{2} + \frac{n^2}{2} - 2ln + \frac{l}{2} - \frac{n}{2} + l \\ &= \frac{3l^2}{2} + \left(-Cn - n + \frac{3}{2}\right) l + \frac{n^2}{2} - \frac{n}{2} < \frac{3l^2}{2} - (C+1)nl + \frac{n^2}{2} + \frac{n}{4}.\end{aligned}$$

This expression reaches its minimum when $l = \frac{1+C}{3}n$.

Let us compare the values at the endpoints of its domain $\left(Cn, \frac{n}{2}\right)$.

When $l = Cn$, the expression is $\frac{(1-C)^2}{2}n^2 + \frac{n}{4}$. When $l = \frac{n}{2}$, the expression is $\frac{3-4C}{8}n^2 + \frac{n}{4}$. Let us note that the latter value is smaller than the former for all possible values of C. Thus, in this case $\mathrm{Cl}'(w) \le \frac{(1-C)^2}{2}n^2 + \frac{n}{4}$.

The maximal bound among the obtained ones is $\frac{(1-C)^2}{2}n^2 + \frac{(1-C)}{2}n$.

We let $\mathrm{pref}_i(w)$ and $\mathrm{suff}_i(w)$ denote the prefix and the suffix of w of length i, respectively.

Proof (of Theorem 1). Let us suppose that w is a word of length n with more than $\frac{Cn(n+1)}{2} + \left(\frac{\sqrt{5}}{2} - 1\right) n$ long closed factors, for some $C \in (\frac{1}{3}, \frac{1}{2})$. Clearly, for the proof we only need to consider C in these bounds due to Propositions 3 and 4. In other words, $\sum_{j=1}^{n} (\mathrm{Cl}'(\mathrm{pref}_j(w)) - \mathrm{Cl}'(\mathrm{pref}_{j-1}(w)) \ge \frac{Cn(n+1)}{2} + \left(\frac{\sqrt{5}}{2} - 1\right) n$. It

would mean that one of the terms in the sum, let us say the i-th one, is at least Ci.

Let us consider the largest index i satisfying $\mathrm{Cl}'(\mathrm{pref}_i(w)) - \mathrm{Cl}'(\mathrm{pref}_{i-1}(w)) \geq Ci$, i.e., for all $j > i$, we have $\mathrm{Cl}'(\mathrm{pref}_j(w)) - \mathrm{Cl}'(\mathrm{pref}_{j-1}(w)) < Cj$. Using Lemma 4, we can bound the number of distinct long closed factors of w the following way.

$$\mathrm{Cl}'(w) < \frac{(1-C)^2}{2}i^2 + \frac{(1-C)}{2}i + \sum_{j=i+1}^{n} Cj = \frac{(1-C)^2}{2}i^2 + \frac{(1-C)}{2}i + C\frac{(n+i+1)(n-i)}{2} = \left(\frac{C^2}{2} - \frac{3C}{2} + \frac{1}{2}\right)i^2 + \left(\frac{1}{2} - C\right)i + \frac{C}{2}n^2 + \frac{C}{2}n.$$

The last expression in the formula is smaller than $\frac{Cn(n+1)}{2} + \left(\frac{\sqrt{5}}{2} - 1\right)n$ when $C \geq \frac{3-\sqrt{5}}{2}$. Thus, the word w has less than $\frac{Cn(n+1)}{2} + \left(\frac{\sqrt{5}}{2} - 1\right)n$ long closed factors. Moreover, this expression reaches its maximum when $C = \frac{3-\sqrt{5}}{2}$.

Thus, the number of long closed factors in a word of length n is bounded by $\frac{3-\sqrt{5}}{4}n^2 + \frac{\sqrt{5}-1}{4}n$.

4 Infinite rich words

We say that an infinite word w is *closed-rich* if there is a constant C such that for each $n \in \mathbb{N}$ each factor of w of length n contains at least Cn^2 distinct closed factors. We remark that in this definition we do not require the constant to be optimal; however, a natural question is optimizing this constant (see Question 2). In this section, we show that infinite closed-rich words exist, and provide some families of examples.

Proposition 5. *Let w be an infinite word, and let $C > 2$, $\alpha < 1$ be two constants. If for each n each factor of w of length n contains a factor of exponent at least C and of length of period at least αn, then w is infinite closed-rich.*

Proof. Let v be a factor of w of length n. By the condition of the lemma, it contains a factor u of period $k \geq \alpha n$ and of exponent $C' \geq C > 2$, hence its length $l = C'k \geq C\alpha n$.

To count closed factors of u we use the following two observations. All factors of u of length greater than $l - k$ are distinct. Each factor of u of length at least $2k$ has exponent at least 2 and hence is closed by Proposition 1.

If $C' \leq 3$, there are at least $\sum_{j=2k}^{l}(l-j) \geq \frac{(C'-2)^2}{2}k^2 \geq \frac{(C-2)^2(\alpha n)^2}{2}$.

If $C' > 3$, then, in addition to the closed factors longer than $l-k$, the word u has k distinct closed factors of each length between $2k$ and $l-k$. Thus, there are at least $(l-3k)k + \sum_{j=l-k}^{l}(l-j) \geq (C'-3)k^2 + \frac{k^2}{2} = \frac{C'-5}{2}k^2 > \frac{(\alpha n)^2}{2} = \frac{\alpha^2 n^2}{2}$.

Therefore, the constant in the definition of infinite rich words is given by $\min(\frac{\alpha^2}{2}, \frac{(C-2)^2\alpha^2}{2})$.

A *morphism* φ is a map on the set of all finite words on the alphabet $\mathbb{A}$ such that $\varphi(uv) = \varphi(u)\varphi(v)$ for all finite words u, v on $\mathbb{A}$. The domain of the morphism φ can be naturally extended to infinite words by $\varphi(w_0w_1w_2\cdots) = \varphi(w_0)\varphi(w_1)\varphi(w_2)\cdots$. A morphism φ is primitive if there exists a positive integer l such that the letter a occurs in the word $\varphi^l(b)$ for each pair of letters $a, b \in \mathbb{A}$. A fixed point of a morphism φ is an infinite word w such that $\varphi(w) = w$.

Example 3. Let w be a fixed point of the morphism $\varphi : a \to abbba, b \to abbbb$. We show that it is infinite closed-rich. Indeed, each block $\varphi^k(c)$ for $c \in \{a, b\}$ has length 5^k and contains a cube with the period 5^{k-1}. Clearly, each factor of length at least $2 \cdot 5^k - 1$ contains a block $\varphi^k(c)$. The maximal length, where we cannot guarantee the next block $\varphi^{k+1}(c)$, is $2{\cdot}5^{k+1}-2$. Thus, we can apply Proposition 5 with $C = 3$ and $\alpha = \frac{1}{(2\cdot 5^2)} = 0.02$. We remark that this word can also be seen as a Toeplitz word [23] with pattern *baaa*?, and that this construction can be easily generalized to other morphic and Toeplitz words.

A large subclass of the family of Sturmian words also turns out to be infinite closed-rich. *Sturmian words* are usually defined as infinite words with the smallest possible number of distinct factors of each length among aperiodic words ($n + 1$ factor of each length $n \geq 1$). Sturmian words are known to be rich in palindromes [16]. Sturmian words admit various characterizations. The one we use here is via standard words. Let $(d_1, d_2, \ldots, d_n, \ldots)$ be a sequence of integers, with $d_1 \geq 0$ and $d_n > 0$ for $n > 1$. To such a sequence, we associate a sequence $(s_n)_{n\geq 1}$ of words by

$$s_{-1} = 1, \qquad s_0 = 0, \qquad s_n = s_{n-1}^{d_n} s_{n-2} \quad (n \geq 1).$$

The sequence $(s_n)_{n\geq -1}$ is a *standard sequence*, and the sequence $(d_1, d_2, \ldots)$ is its *directive sequence*. This sequence defines a limit: $s = \lim_{n\to\infty} s_n$.

It is well known that a word is Sturmian if and only if it has the same set of factors as the limit of some standard sequence. For more information on Sturmian words we refer to Chapter 2 of [18].

Proposition 6. *Let D be an integer and s be a Sturmian word with a directive sequence $(d_i)_{i\geq 1}$ such that $d_i \leq D$ for every $i \geq 1$. Then s is infinite closed-rich.*

Proof. For a sketch of proof, consider a factorization of s to standard words s_k and s_{k-1}. If $d_{k+1} \geq 2$, then between each two consecutive s_{k-1} in the factorization there are at least two s_k (in fact, d_k or d_{k+1}); that gives a power s_k^α with $2 < \alpha \leq D + 1$. In the case when $d_{k+1} = 1$, we can factorize s_k to $s_{k-1}^{d_k} s_{k-2}$ and get a power s_{k-1}^γ with $2 < \gamma \leq D + 1$. By Proposition 5, one can see that the quadratic number of closed factors is achieved inside these powers. Here the constant C from the definition of infinite closed-rich words depends on D. Note that if the sequence of d_i's is unbounded, then the Sturmian word can be made not rich due to presence of powers with big exponents and relatively short periods.

5 Concluding remarks

In this paper, we showed an upper bound $\sim \frac{3-\sqrt{5}}{4}n^2$ for the number of closed factors in a finite word of length n, and we constructed examples with $\sim \frac{n^2}{6}$. We conjecture that this gives an asymptotic bound:

Conjecture 1. Finite closed-rich words contain $\sim \frac{n^2}{6}$ closed factors.

Based on numerical experiments, we also conjecture that they are cubes or words of exponent close to 3. Table 1 shows the maximal number of closed factors that a binary word of given length can contain.

Table 1. The maximal number of closed factors for binary words of length n.

n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$\max_{\|w\|=n} \mathrm{Cl}(w)$	2	3	4	6	8	10	12	15	18	21	25	29
n	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
$\max_{\|w\|=n} \mathrm{Cl}(w)$	33	37	42	48	54	60	66	72	79	86	93	101

Similar calculations have been made in [2], but there were some errors. We made corrections for values $n = 16, 17, \ldots, 20$. For the lengths we computed, closed-rich words are cubes or close to cubes by their structure. For example, the word $u = (100101)^3$ of length 18 has 60 closed factors (one can easily verify it). The word u^- has 54 closed factors, and the word $u1$ has 66 closed factors. The following question remains open even for binary alphabet:

Question 1. What is the exact formula for the maximal number of distinct closed factors in a finite rich word?

In the last section we defined infinite closed-rich words as words for which there exists a constant C such that each factor of length n contains Cn^2 distinct closed factors. A question that naturally arises is that of optimizing the constant:

Question 2. What is the supremum of the constant for infinite rich words?

References

1. Alamro, H., Alzamel, M., Iliopoulos, C.S., Pissis, S.P., Sung, W.K., Watts, S.: Efficient identification of k-closed strings. Int. J. Found. Comput. Sci. **31**(05), 595–610 (2020)
2. Badkobeh, G., Fici, G., Lipták, Z.: On the number of closed factors in a word. In: LATA 2015. LNCS. vol. 8977 (2015)
3. Badkobeh, G., Bannai, H., Goto, K., I, T., Iliopoulos, C.S., Inenaga, S., Puglisi, S.J., Sugimoto, S.: Closed factorization. Discrete Appl. Math. **212**, 23–29 (2016)
4. Bannai, H., I, T., Inenaga, S., Nakashima, Y., Takeda, M., Tsuruta, K.: The "runs" theorem. SIAM J. Comput. **46**(5), 1501–1514 (2017)

5. Bannai, H., Inenaga, S., Kociumaka, T., Lefebvre, A., Radoszewski, J., Rytter, W., Sugimoto, S., Waleń, T.: Efficient algorithms for longest closed factor array. In: SPIRE 2015. LNCS, vol. 9309, pp. 95–102 (2015)
6. Brlek, S., Hamel, S., Nivat, M., Reutenauer, C.: On the palindromic complexity of infinite words. Int. J. Found. Comput. Sci. **15**, 293–306 (2004)
7. Cassaigne, J., Karhumäki, J., Puzynina, S.: On k-abelian palindromes. Inf. Comput. **260**, 89–98 (2018)
8. Crochemore, M., Ilie, L., Tinta, L.: The "runs" conjecture. Theor. Comput. Sci. **412**(27), 2931–2941 (2011)
9. Deza, A., Franek, F., Thierry, A.: How many double squares can a string contain? Discret. Appl. Math. **180**, 52–69 (2015)
10. Droubay, X., Justin, J., Pirillo, G.: Episturmian words and some constructions of de Luca and Rauzy. Theor. Comput. Sci. **255**(1), 539–553 (2001)
11. Durand, F.: A characterization of substitutive sequences using return words. Discret. Math. **179**(1–3), 89–101 (1998)
12. Fici, G.: A classification of trapezoidal words. In: Ambroz, P., Holub, S., Masáková, Z. (eds.) Words 2011. EPTCS, vol. 63, pp. 129–137 (2011)
13. Fici, G.: Open and closed words. Bulletin of the European Association for Theoretical Computer Science **123**, 140–149 (2017)
14. Fici, G., Zamboni, L.Q.: On the least number of palindromes contained in an infinite word. Theor. Comput. Sci. **481**, 1–8 (2013)
15. Fraenkel, A.S., Simpson, J.: How many squares can a string contain? J. Comb. Theory, Ser. A **82**(1), 112–120 (1998)
16. Glen, A., Justin, J., Widmer, S., Zamboni, L.Q.: Palindromic richness. Eur. J. Comb. **30**(2), 510–531 (2009)
17. Kolpakov, R.M., Kucherov, G.: Finding maximal repetitions in a word in linear time. In: FOCS '99. pp. 596–604. IEEE Computer Society (1999)
18. Lothaire, M.: Algebraic combinatorics on words, Encycl. Math. Appl., vol. 90. Cambridge University Press (2002)
19. Morse, M., Hedlund, G.A.: Symbolic dynamics. Am. J. Math. **60**(4), 815–866 (1938)
20. Parshina, O., Postic, M.: Open and closed complexity of infinite words (2020), arXiv:2005.06254
21. Parshina, O., Zamboni, L.Q.: Open and closed factors in Arnoux–Rauzy words. Adv. Appl. Math. **107**, 22–31 (2019)
22. Peltomäki, J.: Introducing privileged words: Privileged complexity of Sturmian words. Theor. Comput. Sci. **500**, 57–67 (2013)
23. Toeplitz, O.: Ein beispiel zur theorie der fastperiodischen funktionen. Math. Ann. **98**, 281–295 (1928)
24. Vuillon, L.: A characterization of Sturmian words by return words. Eur. J. Comb. **22**(2), 263–275 (2001)