International Journal of Foundations of Computer Science
Vol. 23, No. 2 (2012) 303–321

DOI: 10.1142/S012905411240014X

ASYMPTOTIC BEHAVIOUR OF THE MAXIMAL NUMBER OF SQUARES IN STANDARD STURMIAN WORDS

MARCIN PIATKOWSKI*

Faculty of Mathematics and Computer Science, Nicolaus Copernicus University
Torun, Poland
martinp@mat.umk.pl

WOJCIECH RYTTER

Institute of Informatics, Warsaw University
Warsaw, Poland
rytter@mimuw.edu.pl

Received 15 February 2010
Accepted 6 August 2010
Communicated by J. Holub

Denote by $sq(w)$ the number of distinct squares in a string w and let $\mathcal{S}$ be the class of standard Sturmian words. They are generalizations of Fibonacci words and are important in combinatorics on words. For Fibonacci words the asymptotic behaviour of the number of runs and the number of squares is the same. We show that for Sturmian words the situation is quite different. The tight bound $\frac{8}{10}|w|$ for the number of runs was given in [3]. In this paper we show that the tight bound for the maximal number of squares is $\frac{9}{10}|w|$. We use the results of [11], where exact (but not closed) complicated formulas were given for $sq(w)$ for $w \in \mathcal{S}$. We show that for all $w \in \mathcal{S}$ we have $sq(w) \leq \frac{9}{10}|w|$ and there is an infinite sequence of words $w_k \in \mathcal{S}$ such that $\lim_{k\to\infty} |w_k| = \infty$ and $\lim_{k\to\infty} \frac{sq(w_k)}{|w_k|} = \frac{9}{10}$.

Surprisingly the maximal number of squares is reached by the words with recurrences of length only 5. This contrasts with the situation of Fibonacci words, though standard Sturmian words are natural extension of Fibonacci words. If this length drops to 4, the asymptotic behaviour of the maximal number of squares falls down significantly below $\frac{9}{10}|w|$. The structure of Sturmian words rich in squares has been discovered by us experimentally and verified theoretically. The upper bound is much harder, its proof is not a matter of simple calculations. The summation formulas for the number of squares are complicated, no closed formula is known. Some nontrivial reductions were necessary.

Keywords: Sturmian words; repetitions; data compression.

*The research supported by Ministry of Science and Higher Education of Poland, grant N N206 58035.

1. Introduction

A square in a string is a subword of the form ww, where w (called a period) is nonempty. The squares are the simplest forms of repetitions, despite the simple formulation many combinatorial problems related to squares are not well understood. The subject of computing the maximal number of squares and repetitions in words is one of the fundamental topics in combinatorics on words [18, 21] initiated by A. Thue [27], as well as it is important in other areas: lossless compression, word representation, computational biology, etc.

Let $sq(w)$ be the number of distinct squares in the word w, $sq(n)$ be the maximal number of distinct squares in words of length n, and $|w|$ denotes the length of the word w. The behaviour of the function $sq(n)$ is not well understood, though the subject of squares was studied by many authors, see [9, 10, 17]. The best known results related to the value of $sq(n)$ are, see [13, 15, 16]:

$$n - o(n) \ \leq \ sq(n) \ \leq \ 2n - O(\log n).$$

In this paper we concentrate on the asymptotic behaviour of the maximal number of distinct squares in class of standard Sturmian words $\mathcal{S}$. We show that for all $w \in \mathcal{S}$ we have

$$\frac{sq(w)}{|w|} \leq \frac{9}{10}$$

and there is an infinite sequence of strictly growing words $\{w_k\} \in \mathcal{S}$ such that

$$\lim_{k \to \infty} \ \frac{sq(w_k)}{|w_k|} \ = \ \frac{9}{10}.$$

There are known efficient algorithms for the computation of integer powers in words, see [2, 6, 11, 22, 23]. The powers in words are related to maximal repetitions, also called *runs*. It is surprising that the known bounds for the number of runs are much tighter than for squares, this is due to the work of many people [3, 7, 8, 14, 19, 20, 24, 25, 26].

One of interesting questions related to squares is the relation of their number to the number of runs. In case of Fibonacci words the number of squares and runs differ only by 1 and have the same asymptotic behaviour.

The results of this paper show that the maximal number of distinct squares and the maximal number of runs are possibly not closely related, since in case of well structured words (Sturmian words) the *density ratio* of distinct squares (the asymptotic quotient of the maximal number of squares by the length of the string) is $\frac{9}{10}$ and for runs the *density ratio* is $\frac{8}{10}$. Moreover both limits could be reached for different types of words (see section 6 for details).

2. Standard Sturmian Words

The *standard Sturmian words* (*standard words*, in short) are aperiodic words of minimal combinatorial complexity. They are generalization of Fibonacci words and have a very simple *grammar-based* representation which has some algorithmic consequences.

Let $\mathcal{S}$ denote the set of all standard Sturmian words. These words are defined over a binary alphabet $\Sigma = \{a, b\}$ and are described by recurrences (or grammar-based representation) corresponding to so called *directive sequences*: integer sequences

$$\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n),$$

where $\gamma_0 \geq 0$, $\gamma_i > 0$ for $0 < i \leq n$.

The word x_{n+1} corresponding to γ, denoted by $\mathrm{Sw}(\gamma)$, is defined by recurrences:

$$\begin{array}{ll} x_{-1} = b, & x_0 = a, \\ x_1 = x_0^{\gamma_0} x_{-1}, & x_2 = x_1^{\gamma_1} x_0, \\ \vdots & \vdots \\ x_n = x_{n-1}^{\gamma_{n-1}} x_{n-2}, & x_{n+1} = x_n^{\gamma_n} x_{n-1}. \end{array} \tag{1}$$

Fibonacci words are standard Sturmian words given by the directive sequences of the form $\gamma = (1, 1, \ldots, 1)$ (n-th Fibonacci word F_n corresponds to a sequence of n ones). We consider here standard words starting with the letter a, hence assume $\gamma_0 > 0$. The case $\gamma_0 = 0$ can be considered similarly.

For even $n > 0$ a standard word x_n has the suffix ba, and for odd $n > 0$ it has the suffix ab. The number $N = |x_{n+1}|$ is the (real) size, while $n + 1$ can be thought as the compressed size.

Example 1.
Consider the directive sequence $\gamma = (1, 2, 1, 3, 1)$. We have:

$$\begin{array}{lll} x_{-1} & = b & \\ x_0 & = a & \\ x_1 & = (x_0)^1 \cdot x_{-1} & = a \cdot b \\ x_2 & = (x_1)^2 \cdot x_0 & = ab \cdot ab \cdot a \\ x_3 & = (x_2)^1 \cdot x_1 & = ababa \cdot ab \\ x_4 & = (x_3)^3 \cdot x_2 & = ababaab \cdot ababaab \cdot ababaab \cdot ababa \\ x_5 & = (x_4)^1 \cdot x_3 & = ababaababaababaababaababa \cdot ababaab \end{array}$$

and finally

$$\mathrm{Sw}(1, 2, 1, 3, 1) \;=\; ababaababaababaababaababaababaab.$$

The *grammar-based compression* consists in describing a given word by a context-free grammar G generating this (single) word. The size of the grammar G is the total length of all productions of G. In particular each directive sequence of a standard Sturmian word corresponds to such a compression – the sequence of recurrences corresponding to the directive sequence. In this case the size of the grammar is proportional to the length of the directive sequence.

For some lexicographic properties and structure of repetitions of standard Sturmian words see [5, 3, 1, 4].

3. Summation Formulas for the Number of Squares

The exact formulas for the number of squares in standard Sturmian words were given by Damanik and Lenz in [11]. In this section we reformulate their equations to have compact version more suitable for the asymptotic analysis. The formulas are rather complicated and such an analysis is nontrivial. It will be done in the section 5.

Denote $q_i = |x_i|$, where x_i are as in equation (1). We have then

$$q_{-1} \;=\; q_0 \;=\; 1 \qquad \text{and} \qquad q_{i+1} \;=\; \gamma_i q_i + q_{i-1}. \tag{2}$$

The following lemma characterize the possible lengths of periods of squares in Sturmian words.

Lemma 1. *(*[11]*)*
Let $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word. Each primitive period of a square in w has the length kq_i for $1 \le k \le \gamma_i$ or $kq_i + q_{i-1}$ for $1 \le k < \gamma_i$.

The squares in the standard Sturmian word w with period of the length kq_i for $1 \le k \le \gamma_i$ or $kq_i + q_{i-1}$ for $1 \le k < \gamma_i$ are said to be of the *type* i. The squares with the period of the form a^+ are said to be of the *type* 0.

Example 2.
Consider the word from Example 1:

$$\mathrm{Sw}(1, 2, 1, 3, 1) \;=\; abababaababababaababababaabababaab.$$

We have:

one square of type 0: $a{\cdot}a$,

three squares of type 1 (period 2, 3): $ab{\cdot}ab$, $ba{\cdot}ba$, $aba{\cdot}aba$,

three squares of type 2 (period 5): $ababa{\cdot}ababa$, $babaab{\cdot}babaab$, $abaab{\cdot}abaab$,

and eleven squares of type 3 (with periods 7, 14):

$ababaab \cdot ababaab$, $babaaba \cdot babaaba$, $abaabab \cdot abaabab$, $baababa \cdot baababa$,

$aababab \cdot aababab$, $ababaab \cdot ababaab$, $bababaa \cdot bababaa$,

$ababaababab aab \cdot ababaababab aab$, $babaababab aaba \cdot babaababab aaba$,

$abaababab aabab \cdot abaababab aabab$, $baababab aababa \cdot baababab aababa$,

see Fig. 1.

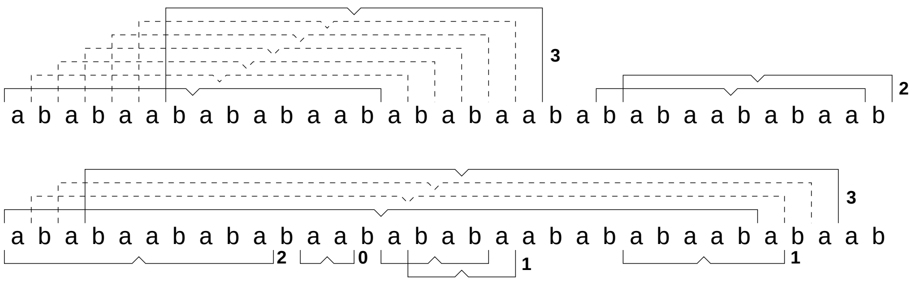

Fig. 1. The squares in word $\mathrm{Sw}(1, 2, 1, 3, 1)$ with their types.

We slightly abuse the notation and denote

$$sq(\gamma_0, \gamma_1, ..., \gamma_n) \;=\; sq\big(\mathrm{Sw}(\gamma_0, \gamma_1, ..., \gamma_n)\big).$$

Let $sq_i(\gamma_0, \gamma_1, \ldots, \gamma_n)$, for $0 \le i \le n$, be the number of distinct squares of the type i in the word $w = \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$. We count all squares in w by counting separately all squares of each type:

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \;=\; \sum_{i=0}^{n} \; sq_i(\gamma_0, \gamma_1, \ldots, \gamma_n).$$

For the directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ and $1 \le i \le n$ denote:

$$d(0) \;=\; \left\lfloor \tfrac{\gamma_0+1}{2} \right\rfloor,$$

$$d_1(i) \;=\; \begin{cases} \frac{\gamma_i}{2} q_i + q_{i-1} - 1 & \text{for even } \gamma_i \\ \frac{\gamma_i}{2} q_i + \frac{1}{2}\, q_i & \text{for odd } \gamma_i \end{cases}, \tag{3}$$

$$d(i) \;=\; d_1(i) + \gamma_i\, q_i - q_i - \gamma_i + 1,$$

where q_i are as in equation 2.

Let $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word. Then $sq(\gamma_0, \gamma_1, \ldots, \gamma_n)$ is determined as follows, see [11]:

Summation formulas:

$$(\mathbf{1}) \qquad sq(\gamma_0, \gamma_1, \ldots, \gamma_n) = \sum_{i=0}^{n} sq_i(\gamma_0, \gamma_1, \ldots, \gamma_n).$$

$$(\mathbf{2}) \quad (0 \le i \le n-3) \text{ or } (i = n-2 \;\&\; \gamma_n \ge 2) \;\Rightarrow\; sq_i(\gamma) = d(i).$$

$$(\mathbf{3}) \quad \gamma_n = 1 \;\Rightarrow\; sq_{n-2}(\gamma) = \begin{cases} d(n-2) - q_{n-3} + 1 & \text{for even } \gamma_{n-2} \\ d(n-2) - q_{n-2} + q_{n-3} + 1 & \text{otherwise} \end{cases}$$

$$(\mathbf{4}) \quad \gamma_n = 1 \;\Rightarrow\; sq_{n-1}(\gamma) = \begin{cases} d_1(n-1) - q_{n-2} + 1 & \text{for even } \gamma_{n-1} \\ d_1(n-1) - q_{n-1} + q_{n-2} - 1 & \text{otherwise} \end{cases} \tag{4}$$

$$(\mathbf{5}) \quad \gamma_n > 1 \;\Rightarrow\; sq_{n-1}(\gamma) = \begin{cases} d(n-1) - q_{n-2} + 1 & \text{for even } \gamma_{n-1} \\ d(n-1) - q_{n-1} + q_{n-2} - 1 & \text{otherwise} \end{cases}$$

$$(\mathbf{6}) \qquad sq_n(\gamma) = \begin{cases} d_1(n) - q_n + 2 & \text{for even } \gamma_n \\ d_1(n) - q_n & \text{otherwise} \end{cases}$$

4. Sturmian Words with Many Squares

In this section we present and analyze the sequence $\{w_k\}$ of strictly growing Sturmian words achieving asymptotically maximal ratio of the number of distinct squares to the length of the word:

$$\lim_{k\to\infty} |w_k| = \infty \qquad \text{and} \qquad \lim_{k\to\infty} \frac{sq(w_k)}{|w_k|} = \frac{9}{10}.$$

Recall that the squares with periods of the length kq_i for $1 \le k \le \gamma_i$ or $kq_i + q_{i-1}$ for $1 \le k < \gamma_i$, where q_i are as in equation 2, are said to be of the type i and the squares with periods of the form a^+ are said to be of the type 0.

Consider a directive sequence $\gamma_k = (k, k, 2, 1, 1)$ and a word $w_k = \mathrm{Sw}(k, k, 2, 1, 1)$, where $k > 0$.

Example 3.

$$
\begin{aligned}
w_1 &= \mathrm{Sw}(1,1,2,1,1) = (aba)^2ab(aba)^3ab,\\
w_2 &= \mathrm{Sw}(2,2,2,1,1) = \Big((aab)^2a\Big)^2 aab\Big((aab)^2a\Big)^3 aab,\\
w_3 &= \mathrm{Sw}(3,3,2,1,1) = \Big((aaab)^3a\Big)^2 aaab\Big((aaab)^3a\Big)^3 aaab,\\
&\vdots\\
w_k &= \mathrm{Sw}(k,k,2,1,1) = \Big(\big(a^kb\big)^ka\Big)^2 a^kb\Big(\big(a^kb\big)^ka\Big)^3 a^kb.
\end{aligned}
$$

Sw(3, 3, 2, 1, 1) is illustrated in Fig. 2.

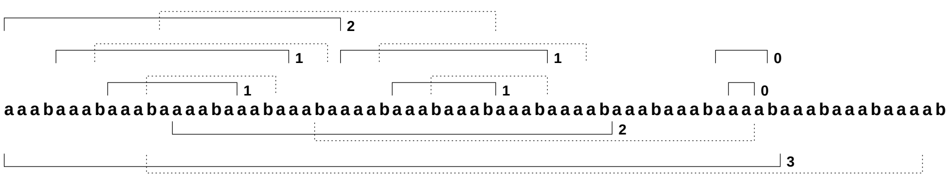

Fig. 2. The squares in word Sw(3, 3, 2, 1, 1) with their *shifts* and types.

Theorem 2.
We have $sq(k,k,2,1,1) \longrightarrow \frac{9}{10} \cdot \Big|\mathrm{Sw}(k,k,2,1,1)\Big|$ *for* $k \longrightarrow \infty$.

Proof.
Let $\gamma_k = (k,k,2,1,1)$. We have:

$$\mathrm{Sw}(\gamma_k) = \Big((a^kb)^ka\Big)^2 a^kb\Big((a^kb)^ka\Big)^3 a^kb$$

and

$$|\mathrm{Sw}(\gamma_k)| = 5k^2 + 7k + 7.$$

We compute separately the number of distinct squares of each type $sq_i(\gamma_k)$ for $0 \le i \le 4$ in the word w_k.

There are two cases depending on the parity of the parameter k. We are interested in asymptotic behaviour of the number of distinct squares, hence we can assume that $k > 1$.

Case 1: k is odd.

We have (according to formulas (1-6) from the equation 4):

$$\begin{aligned} sq_0(\gamma_k) &= \frac{1}{2}\Big(k+1\Big), \\ sq_1(\gamma_k) &= \frac{1}{2}\Big(3k^2+1\Big), \\ sq_2(\gamma_k) &= 2k^2+2k+1, \\ sq_3(\gamma_k) &= k^2+k, \\ sq_4(\gamma_k) &= 0. \end{aligned}$$

Summing altogether we have:

$$sq(\gamma_k) = \frac{1}{2}\Big(9k^2+7k+4\Big),$$

and finally

$$\lim_{k\to\infty} \frac{sq(\gamma_k)}{|\mathrm{Sw}(\gamma_k)|} = \lim_{k\to\infty} \frac{9k^2+7k+4}{10k^2+14k+14} = \frac{9}{10}.$$

Case 2: k is even.

We have (according to formulas (1-6) from the equation 4):

$$\begin{aligned} sq_0(\gamma_k) &= \frac{1}{2}k, \\ sq_1(\gamma_k) &= \frac{1}{2}\Big(3k^2-k\Big), \\ sq_2(\gamma_k) &= 2k^2+2k+1, \\ sq_3(\gamma_k) &= k^2+k, \\ sq_4(\gamma_k) &= 0. \end{aligned}$$

Summing altogether we have:

$$sq(\gamma_k) = \frac{1}{2}\Big(9k^2+6k+2\Big),$$

and finally

$$\lim_{k\to\infty} \frac{sq(\gamma_k)}{|\mathrm{Sw}(\gamma_k)|} = \lim_{k\to\infty} \frac{9k^2+6k+2}{10k^2+14k+14} = \frac{9}{10}. \qquad \square$$

5. Asymptotic Behaviour of the Maximal Number of Squares

The formulas (1-6) from the equation 4 give together the value of $sq(\gamma)$, however there is no close simple formula. Therefore tight asymptotic estimations are nontrivial.

We start with two lemmas, that allow us to restrict the value of the last two elements of the directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ in the asymptotic estimation of the maximal number of distinct squares in $\mathrm{Sw}(\gamma)$.

Lemma 3. ***[Reduction of** γ_n**]***
Let $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *be a standard Sturmian word. If* $\gamma_n > 1$ *then*

$$sq(\gamma_0, \ldots, \gamma_{n-1}, \gamma_n) \ \leq \ sq(\gamma_0, \ldots, \gamma_{n-1}, \gamma_n - 1, 1) + 2.$$

Proof.
The words $w_1 = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n)$ and $w_2 = \mathrm{Sw}(\gamma_0, \ldots, \gamma_n - 1, 1)$ differ only on the last two letters – ab or ba, see [4, 18] for more details.

The squares of the types 0, 1, ..., $n-1$ and short squares of the type n are the same for w_1 and w_2 (see Fig. 3). The difference is possible only for the longest squares of the type n. Exchange of the last two letters enables (or disables respectively) the shift of the longest square of the type n by one and two positions (see the squares marked in bold on Fig. 3). In w_2 we have $\gamma_{n+1} = 1$ and, due to formulas (1-6) from the equation 4, there is no squares of the type $n+1$ in w_2. Therefore the difference between the numbers of squares in w_1 and w_2 is not greater than 2, what follows the thesis. □

Lemma 4. ***[Reduction of** γ_{n-1}**]***
Let $w = \mathrm{Sw}(\gamma_0, \ldots, \gamma_{n-2}, \gamma_{n-1}, 1)$, $w_1 = \mathrm{Sw}(\gamma_0, \ldots, \gamma_{n-2}, 1, 1)$,
$w_2 = \mathrm{Sw}(\gamma_0, \ldots, \gamma_{n-2}, 2, 1)$ *be standard Sturmian words.*
Then we have:

$$\left(sq(w_1) \ \leq \ \frac{9}{10}\,|w_1| - 2 \quad \textit{and} \quad sq(w_2) < \frac{9}{10}\,|w_2| - 2 \right) \quad \Longrightarrow \quad sq(w) \ \leq \ \frac{9}{10}\,|w| - 2.$$

Proof.
If γ_{n-1} is odd then let $\Delta = \gamma_{n-1} - 1$ otherwise let $\Delta = \gamma_{n-1} - 2$.

Consider what happens when we change γ_{n-1} by the quantity Δ (see formula (4) from the equation 4). The increase of the number of distinct squares is $\frac{\Delta}{2}\, q_{n-1}$,

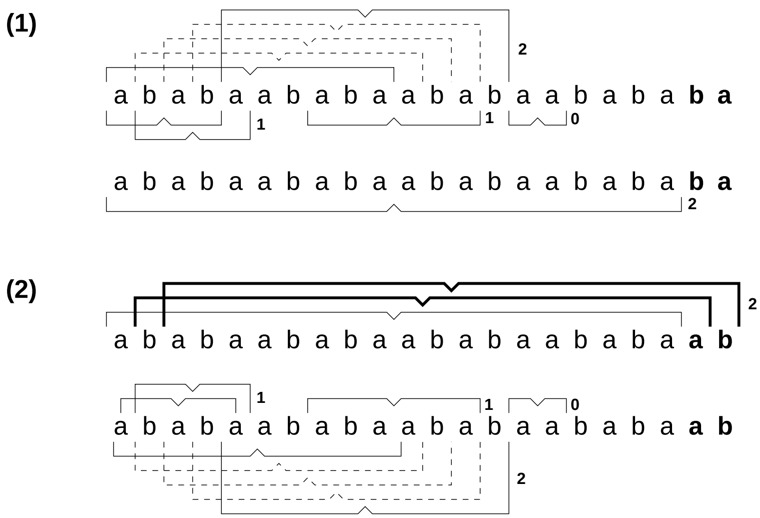

Fig. 3. The squares in the words Sw(1, 2, 3, 1) – (1) and Sw(1, 2, 4) – (2). Two additional squares of the type 2 in Sw(1, 2, 4) are marked in bold.

while the increase of the length of the word is $\Delta\ q_{n-1}$. The increase of the number squares is amortized by half of the increase of the length of the word. Therefore we can subtract Δ from γ_{n-1} and the thesis follows. □

Now we are ready to estimate the number of distinct squares in standard Sturmian words defined by short directive sequences.

Lemma 5. *[Short γ]*
If $n < 3$ then

$$sq(\gamma_0, \ldots, \gamma_n) \;\leq\; \frac{9}{10}\big|\mathrm{Sw}(\gamma_0, \ldots, \gamma_n)\big|.$$

Proof.
There are three types of short directive sequences: $\gamma^I = (\gamma_0)$, $\gamma^{II} = (\gamma_0, \gamma_1)$ and $\gamma^{III} = (\gamma_0, \gamma_1, \gamma_2)$. We consider them separately.

Case 1: $\gamma^I = (\gamma_0)$.
We have (due to formulas (1-6) from the equation 4):

$$sq\big(\gamma^I\big) \leq \frac{1}{2}\big(\gamma_0 + 1\big) \qquad \text{and} \qquad \big|\mathrm{Sw}\big(\gamma^I\big)\big| = \gamma_0 + 1.$$

Therefore

$$sq\big(\gamma^I\big) \;\;\leq\;\; \frac{1}{2}\Big|\mathrm{Sw}\big(\gamma^I\big)\Big| \;\;<\;\; \frac{9}{10}\Big|\mathrm{Sw}\big(\gamma^I\big)\Big|.$$

Case 2: $\gamma^{II} = (\gamma_0, \gamma_1)$.
We have (due to formulas (1-6) from the equation 4):

$$\begin{aligned} \left|\mathrm{Sw}\left(\gamma^{II}\right)\right| &= \left(\gamma_0 + 1\right)\gamma_1 + 1, \\ sq\left(\gamma^{II}\right) &= sq_0\left(\gamma^{II}\right) + sq_1\left(\gamma^{II}\right), \\ sq_0\left(\gamma^{II}\right) &\leq \tfrac{1}{2}\left(\gamma_0 + 1\right). \end{aligned}$$

There are two cases depending on the value of γ_1.

I: If $\gamma_1 = 1$ then $sq_1\left(\gamma^{II}\right) = 0$ and we have:

$$sq\left(\gamma^{I}\right) \leq \frac{1}{2}\left(\gamma_0 + 1\right),$$

and consequently

$$\frac{sq\left(\gamma^{II}\right)}{\left|\mathrm{Sw}\left(\gamma^{II}\right)\right|} \leq \frac{1}{2} - \frac{3}{2\gamma_0 + 4} \leq \frac{9}{10}.$$

II: If $\gamma_1 > 1$ then

$$sq_1\left(\gamma^{II}\right) \leq \frac{1}{2}\left(\gamma_0 + 1\right)\left(\gamma_1 - 1\right) + 2,$$

and we have:

$$sq\left(\gamma^{II}\right) \leq \frac{1}{2}\left(\gamma_0 + 1\right)\gamma_1 + 2,$$

and finally

$$\frac{sq\left(\gamma^{II}\right)}{\left|\mathrm{Sw}\left(\gamma^{II}\right)\right|} \leq \frac{1}{2} + \frac{3}{2\left(\gamma_0 + 1\right)\gamma_1 + 2} \leq \frac{9}{10}.$$

Case 3: $\gamma^{III} = (\gamma_0, \gamma_1, \gamma_2)$.
There are two cases depending on the value of γ_2.

I: If $\gamma_2 = 1$ then we have (due to formulas (1-6) from the equation 4):

$$\begin{aligned} \left|\mathrm{Sw}\left(\gamma^{III}\right)\right| &= \left(\gamma_0 + 1\right)\left(\gamma_1 + 1\right) + 1, \\ sq\left(\gamma^{III}\right) &= sq_0\left(\gamma^{III}\right) + sq_1\left(\gamma^{III}\right) + sq_2\left(\gamma^{III}\right), \\ sq_0\left(\gamma^{III}\right) &\leq \tfrac{1}{2}\left(\gamma_0 + 1\right), \\ sq_1\left(\gamma^{III}\right) &\leq \tfrac{1}{2}\left(\gamma_0 + 1\right)\left(\gamma_1 + 1\right), \\ sq_2\left(\gamma^{III}\right) &= 0. \end{aligned}$$

Therefore

$$sq(\gamma_0, \gamma_1, 1) \leq \frac{1}{2}\left(\gamma_0 + 1\right)\left(\gamma_1 + 2\right),$$

and finally

$$\frac{sq\big(\gamma^{III}\big)}{\big|\mathrm{Sw}\big(\gamma^{III}\big)\big|} \;\le\; \frac{1}{2} \;+\; \frac{\gamma_0}{2\big(\gamma_0+1\big)\big(\gamma_1+1\big)+2} \;\le\; \frac{9}{10}.$$

II: If $\gamma_2 > 1$ then we have (due to theorem 3)

$$sq(\gamma_0, \gamma_1, \gamma_2) \;\le\; sq(\gamma_0, \gamma_1, \gamma_2 - 1, 1) + 2,$$

and the proof is similar to the proof of the theorem 7. □

The next two facts will be useful in estimation of the number of distinct squares for longer directive sequences.

Observation

$$d(i) \;\le\; \begin{cases} \left(\frac{3}{2}\,\gamma_i - 1\right)\, q_i + q_{i-1} - 1 & \text{for even } \gamma_i \\ \left(\frac{3}{2}\,\gamma_i - \frac{1}{2}\right)\, q_i & \text{for odd } \gamma_i \end{cases}.$$

Lemma 6.
For $0 \le r \le n-3$ *we have*

$$\sum_{i=0}^{r} d(i) \;<\; \frac{3}{2}\, q_{r+1} + q_r - 2.$$

Proof.
Recall that $q_i \;=\; |x_i|$ (see equations 1 and 2). According to the observation above and implication

$$\gamma_i \ge 2 \quad\Rightarrow\quad q_i - q_{i+1} < -\,\frac{1}{2}\, q_{i+1},$$

we have:

$$d(i) \le \frac{3}{2}\,\gamma_i\; q_i - \frac{1}{2}\, q_i.$$

Observe now that $\gamma_i\; q_i = q_{i+1} - q_{i-1}$. Therefore

$$\begin{aligned}\sum_{i=0}^{r} \gamma_i\; q_i \;&=\; q_{r+1} + q_r - q_0 - q_{-1} \\ &=\; q_{r+1} + q_r - 2.\end{aligned}$$

Consequently

$$\begin{aligned}\sum_{i=0}^{r} d(i) \; &< \; \frac{3}{2} \sum_{i=0}^{r} \gamma_i \, q_i \; - \frac{1}{2} \, q_r \\ &\leq \; \frac{3}{2} \Big(q_{r+1} + q_r - 2 \Big) - \frac{1}{2} \, q_r \qquad (5) \\ &\leq \; \frac{3}{2} \, q_{r+1} + q_r - 2.\end{aligned}$$

This completes the proof. □

Now we are ready to prove the tight bound for the number of squares in standard Sturmian words.

Theorem 7.
Let $\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ *be a standard Sturmian word. Then*

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \; \leq \; \frac{9}{10} \cdot \Big| \mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n) \Big| .$$

Proof.
If $n < 3$ then the thesis follows from the lemma 5, hence we can assume that $n \geq 3$.

We start with assumption that:

$$\boxed{\gamma_n = 1 \quad \text{and} \quad \gamma_{n-1} \in \{1, 2\} .}$$

Let us shorten the notation and denote:

$$A = q_{n-2}, \quad B = q_{n-3}, \quad \alpha = \gamma_{n-2}.$$

Claim 1.
Recall that $q_i = |x_i|$ and due the equation (1) we have $A > B$ and the values of A and B increase exponentially. The smallest growth of A and B we have for Fibonacci words, where A and B are consecutive Fibonacci numbers. For other standard Sturmian words the growth of A and B is significantly larger and the difference between A and B is bigger.

We have, due to Lemma 6, the following fact (in terms of A and B):

Claim 2.

$$\sum_{i=0}^{n-3} sq_i(\gamma) \; = \; \sum_{i=0}^{n-3} d(i) \; \leq \; \frac{3}{2} \, A + B - 2.$$

This, together with the fact that $sq_n(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}, 1) = 0$, implies:

Claim 3.

$$sq(\gamma) \;\le\; \Phi(\gamma) \;\stackrel{def}{=}\; \frac{3}{2}\,A + B - 2 + sq_{n-1}(\gamma) + sq_{n-2}(\gamma).$$

Our goal is to prove the inequality

$$\Phi(\gamma) \le \frac{9}{10}\,|w| - 2.$$

Using our terminology we can write:

$$\textbf{(a)}\ \big|\mathrm{Sw}(\gamma)\big| \;=\; \begin{cases} 2\,\alpha\,A + A + 2B & \text{for }\ \gamma_{n-1} = 1 \\ 3\,\alpha\,A + A + 3B & \text{for }\ \gamma_{n-1} = 2 \end{cases}.$$

$$\textbf{(b)}\ sq_{n-2}(\gamma) \;\le\; \begin{cases} \frac{3}{2}\,\alpha\,A - A & \text{for even } \gamma_{n-2} \\ \frac{3}{2}\,\alpha\,A - \frac{3}{2}A + B + 1 & \text{for odd } \gamma_{n-2} \end{cases}.$$

$$\textbf{(c)}\ sq_{n-1}(\gamma) \;\le\; \begin{cases} \alpha\,A + B & \text{for }\ \gamma_{n-1} = 2 \\ A - 1 & \text{for }\ \gamma_{n-1} = 1 \end{cases}.$$

There are 4 cases depending on $\gamma_{n-1} \in \{1, 2\}$ and the parity of α.

Case 1: ($\gamma_{n-1} = 1$, α is even)

In this case inequality $\Phi(\gamma) \le \frac{9}{10}\,|w|$ reduces to:

$$\frac{3}{2}\Big(\alpha + 1\Big)\ A + B \;\le\; \frac{9}{10}\ \Big((2\,\alpha + 1)\ A + 2\ B\Big).$$

This reduces to:

$$(3\alpha - 6)A + 8B \ge 0,$$

which obviously holds for $\alpha \ge 2$.

This completes the proof of this case.

Case 2: ($\gamma_{n-1} = 1$, α is odd)

In this case the inequality $\Phi(\gamma) \le \frac{9}{10}\,|w| - 2$ reduces to:

$$\left(\frac{3}{2}\,\alpha+1\right)\,A+2\,B\leq\;\frac{9}{10}\,\Big((2\,\alpha+1)\,A+2\,B\Big).$$

This reduces to

$$(3\alpha-1)A\geq 2B,$$

which holds since $\alpha\geq 1$ and due the Claim 1.

This completes the proof of this case.

Case 3: ($\gamma_{n-1}=2$, α is even)

In this case

$$\Phi(\gamma)\leq\left(\frac{5}{2}\,\alpha+\frac{1}{2}\right)\,A+2\,B-2.$$

Consequently the inequality $\Phi(\gamma)\leq\frac{9}{10}\,|w|-2$ reduces to:

$$\left(\frac{5}{2}\,\alpha+\frac{1}{2}\right)\,A+2\,B\;\leq\;\frac{9}{10}\,\Big(3\,\alpha\,A+A+3B\Big).$$

This reduces to

$$(2\alpha+4)A+7B\geq 0$$

and holds since $\alpha\geq 2,\;A>B>0$.

Case 4: ($\gamma_{n-1}=2$, α is odd)

In this case

$$\Phi(\gamma)\;\leq\;\frac{5}{2}\,\alpha\,A+3\,B-1.$$

Now the inequality $\Phi(\gamma)\leq\frac{9}{10}\,|w|-2$ reduces to:

$$\frac{5}{2}\,\alpha\,A+3\,B+1\;\leq\;\frac{9}{10}\,\Big(3\,\alpha\,A+A+3B\Big).$$

This reduces to

$$3B+10\leq(2\alpha+9)A,$$

which holds since $\alpha\geq 1$ and due the Claim 1.

We proved that

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 1, 1) \;\le\; \frac{9}{10}\,|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 1, 1)| - 2$$

and

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 2, 1) \;\le\; \frac{9}{10}\,|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_{n-2}, 2, 1)| - 2.$$

This implies, that in general case, due to Lemma 3 and Lemma 4, we have

$$sq(\gamma_0, \gamma_1, \ldots, \gamma_n) \;\le\; \frac{9}{10}\,|\mathrm{Sw}(\gamma_0, \gamma_1, \ldots, \gamma_n)|,$$

which completes the proof of the theorem. □

6. Squares versus Maximal Repetitions

The *maximal repetition* (the *run*, in short) in a word w is a nonempty subword $w[i..j] = u^k v$ ($k \ge 2$), where u is of the minimal length and v is proper prefix (possibly empty) of u, that can not be extended (neither $w[i-1..j]$ nor $w[i..j+1]$ is a run with period $|u|$).

Let $\rho(w)$ be the number of runs in the word w. For n-th Fibonacci word F_n we have:

$$sq(F_n) \;=\; 2|F_{n-2}| - 2,$$
$$\rho(F_n) \;=\; 2|F_{n-2}| - 3,$$

hence $sq(F_n) = \rho(f_n)+1$, consequently $\frac{sq(F_n)}{|F_n|}$ and $\frac{\rho(F_n)}{|F_n|}$ have the same asymptotic behaviour, see [12, 20].

For standard Sturmian words the situation is different. We have:

$$\frac{\rho(w)}{|w|} \longrightarrow 0.8 \qquad \text{and} \qquad \frac{sq(w)}{|w|} \longrightarrow 0.9,$$

see [3] for more details. Below we will investigate three different sequences of standard Sturmian words to see that the number of squares and the number of runs are not so closely related as in Fibonacci words.

Case 1: $w_k = \mathrm{Sw}(k, k, 2, 1, 1)$.
The word w_k has the form:

$$w_k \;=\; \Big((a^k b)^k a\Big)^2 a^k b \Big((a^k b)^k a\Big)^3 a^k b$$

and length

$$|w_k| = 5k^2 + 7k + 7.$$

In the section 4 we have computed the number of squares for the word w_k, and we have seen that

$$\frac{sq(w_k)}{|w_k|} \;\longrightarrow\; \frac{9}{10}.$$

Now we compute the number of runs for w_k using formulas from [3]. We have:

$$\rho(w_k) = 9k + 7,$$

hence

$$\frac{\rho(w_k)}{|w_k|} \longrightarrow 0.$$

We can see that w_k is an example of a word that is rich in squares and at the same time has very small number of runs.

Case 2: $v_k = \mathrm{Sw}(1, 2, k, k)$.
The word v_k has the form

$$v_k \;=\; \Big((ababa)^k ab\Big)^k ababa,$$

and length

$$|v_k| = 5k^2 + 2k + 5.$$

Using formulas from [3] we we have

$$\rho(v_k) = 4k^2 - k + 3,$$

hence

$$\frac{\rho(v_k)}{|v_k|} \longrightarrow \frac{8}{10}.$$

Using the formulas (1-6) from the section 3 we have:

$$sq(v_k) = \begin{cases} \frac{5}{2}k^2 + \frac{5}{2}k + 4 & \text{for even } \; k \\ \frac{5}{2}k^2 + 5k - \frac{5}{2} & \text{for odd } \; k \end{cases}$$

consequently

$$\frac{sq(v_k)}{|v_k|} \longrightarrow \frac{1}{2}.$$

We can see that v_k is an example of a word for which the number of squares is significantly smaller than the number of runs.

Case 3: $z_k = \mathrm{Sw}(1, 2, k, k, 2, 1, 1)$.
The word z_k has the form

$$z_k \;=\; \Big[\Big((ababa)^k ab\Big)^k ababa\Big]^2 (ababa)^k ab \Big[\Big((ababa)^k ab\Big)^k ababa\Big]^3 (ababa)^k ab$$

and length

$$|z_k| \;=\; 25k^2 + 20k + 29.$$

Using the formulas (1-6) from the section 3 we compute the number of squares:

$$sq(z_k) = \begin{cases} \frac{45}{2}k^2 + \frac{19}{2}k + 17 & \text{for even } \; k \\ \frac{45}{2}k^2 + 12k + \frac{31}{2} & \text{for odd } \; k \end{cases}$$

and consequently:

$$\frac{sq(z_k)}{|z_k|} \longrightarrow \frac{9}{10}.$$

Using formulas from [3] we compute the number of runs:

$$\rho(z_k) = 20k^2 + 11k + 20,$$

hence

$$\frac{\rho(z_k)}{|z_k|} \longrightarrow \frac{8}{10}.$$

We can see that z_k is an example of a word for which both the number of squares and the number of runs are high.

The results above show that the maximal number of squares and the maximal number of runs for standard Sturmian words are not closely related. The asymptotic limits are close, but for different types of words the number of squares and the number of runs could have different asymptotic behaviour.

References

[1] J. Allouche, J. Shallit, Automatic Sequences: Theory, Applications, Generalizations, Cambridge University Press, 2003.

[2] A. Apostolico, F. P. Preparata, Optimal off-line detection of repetitions in a string, *Theoretical Computer Science*, 22 1983, pp.297-315.

[3] P. Baturo, M. Piatkowski, W. Rytter, The number of runs in Sturmian words, *Proceedings of International Conference on Implementation and Application of Automata* 2008, pp.252-261.

[4] P. Baturo, M. Piatkowski, W. Rytter, Usefulness of directed acyclic subword graphs in problems related to standard Sturmian words, *International Journal of Foundations of Computer Science*, 20(6) 2009, pp. 1005-1023.

[5] P. Baturo, W. Rytter, Occurrence and lexicographic properties of standard Sturmian words, *Proceedings of International Conference on Language and Automata Theory and Applications*, 2007, pp. ???

[6] M. Crochemore, An optimal algorithm for computing the repetitions in a word, *Information Processing Letters* 12(5) 1981, pp. 244-250.

[7] M. Crochemore, L. Ilie, Maximal repetitions in strings, *Journal of Computer and System Sciences* 74(5) 2008, pp. 796-807.

[8] M. Crochemore, L. Ilie, L. Tinta, Towards a solution to the "runs" conjecture, *Lecture Notes in Computer Science* 5029 2008, pp.290-302.

[9] M. Crochemore, W. Rytter, Squares, cubes, and time-space efficient string searching, *Algorithmica* 13(5) 1995, pp. 405-425.

[10] M. Crochemore, W. Rytter, Jewels of stringology, World Scientific, 2003.

[11] D. Damanik, D. Lenz, Powers in Sturmian sequences, *European Journal of Combinatorics* 24(4) 2003, pp. 377-390.

[12] A. S. Fraenkel, J. Simpson, The exact number of squares in Fibonacci words, *Theoretical Computer Science* 218(1) 1999, pp. 95-106.

[13] A. S. Fraenkel, J. Simpson, How many squares can a string contain?, *Journal of Combinatorial Theory Series A* 82 1998, pp. 112-120.

[14] M. Giraud, Not so many runs in a string, *Proceedings of International Conference on Language and Automata Theory and Applications*, 2008, pp.232-239.

[15] L. Ilie, A simple proof that a word of length n has at most $2n$ distinct squares, *Journal of Combinatorial Theory Series A* 112 2005, pp.163-164.

[16] L. Ilie, A note on the number of squares in a word, *Theoretical Computer Science* 380 2007, pp. 373-376.

[17] C. S. Iliopoulos, D. Moore, W. F. Smyth, A characterization of the squares in a Fibonacci string, *Theoretical Computer Science* 172(1-2) 1997, pp. 281-291.

[18] J. Karhumaki, Combinatorics on words, notes in pdf.

[19] R. M. Kolpakov, G. Kucherov, Finding maximal repetitions in a word in linear time, *Proceedings of Symposium on Foundations of Computer Science* 1999, pp. 596-604.

[20] R. M. Kolpakov, G. Kucherov, On maximal repetitions in words, *Lecture Notes in Computer Science* 1999, pp. 374-385.

[21] M. Lothaire, Applied Combinatorics on Words, Cambridge University Press, 2005.

[22] M. G. Main, Detecting leftmost maximal periodicities, *Discrete Applied Mathematics* 25(1-2) 1989, pp. 145-153.

[23] M. G. Main, J. Lorentz, An $o(n \log n)$ algorithm for finding all repetitions in a string, *Journal of Algorithms* 5(3) 1984, pp. 422-432.

[24] S. J. Puglisi, J. Simpson, W. F. Smyth, How many runs can a string contain?, *Theoretical Computer Science* 4001(1-3) 2008, pp.165-171.

[25] W. Rytter, The number of runs in a string: Improved analysis of the linear upper bound, *Lecture Notes in Computer Science* 3884 2006, pp. 184-195.

[26] W. Rytter, The number of runs in a string, *Information and Computation* 205(9) 2007, pp. 1459-1469.

[27] A. Thue, Uber unendliche zeichenreihen, *Norske Vid. Selsk. Skr. I Math-Nat.* 7 1906, pp. 1-22.