Counting overlap-free binary words

Julien Cassaigne

École Normale Supérieure, Paris

AMS Mathematics Subject Classification: 05A15

Abstract

A word on a finite alphabet A is said to be overlap-free if it contains no factor of the form $xuxux$, where x is a letter and u a (possibly empty) word. In this paper we study the number u_n of overlap-free binary words of length n, which is known to be bounded by a polynomial in n. First, we describe a bijection between the set of overlap-free words and a rational language. This yields recurrence relations for u_n, which allow to compute u_n in logarithmic time. Then, we prove that the numbers $\alpha = \sup \{ r \mid n^r = \mathcal{O}(u_n) \}$ and $\beta = \inf \{ r \mid u_n = \mathcal{O}(n^r) \}$ are distinct, and we give an upper bound for α and a lower bound for β. Finally, we compute an asymptotically tight bound to the number of overlap-free words of length less than n.

1 Introduction

In general, the problem of evaluating the number u_n of words of length n in the language U consisting of words on some finite alphabet A with no factors in a certain set F is not easy. If F is finite, it amounts to counting words in the rational language U and there are well-known techniques for this, e.g. computing the generating function. But if for instance F is the set of images of a pattern p by non-erasing morphisms, it is not even known how to decide whether U is finite or infinite (i.e. whether p is unavoidable or avoidable on A). However, Brandenburg [2] proved that u_n grows exponentially for the pattern $p = \alpha^2$ on a ternary alphabet A (then U is the set of square-free ternary words), and Goralcik and Vanicek [6] proved the same for any 2-avoidable binary pattern p on a binary alphabet A (also proved by Brandenburg for $p = \alpha^3$).

Here we shall study the growth of u_n in the case where F is the set

$$F = \{ \, xuxux \mid x \in A \text{ and } u \in A^* \, \}$$

with $A = \{a, b\}$; the elements of U are then called *overlap-free binary words*. After recalling what is known about u_n (end of this section), we describe (Section 2) a bijection between U and a rational language. From the automaton recognizing this language we construct (Section 3) explicit recurrence relations verified by u_n (showing that u_n can be computed in logarithmic time). In Section 4, we study the consequences of these relations on the asymptotic behaviour of u_n. In particular we prove that, although u_n is bounded from below and from above by polynomial quantities, it is not itself equivalent to a polynomial.

1.1 Overlap-Free Binary Words

We consider words on the alphabet $A = \{a, b\}$, i.e. elements of the monoid A^*. The letter ε denotes the empty word. We shall use the notation $\overline{a} = b$ and $\overline{b} = a$.

A word w contains an *overlap* if some factor v appears at two overlapping positions. It is equivalent to say that w contains a factor of the form $xuxux$ with $x \in A$ and $u \in A^*$.

A word w is called *overlap-free* if it contains no overlap.

In the language of *patterns* an overlap-free word is a word which simultaneously *avoids* the patterns $\alpha\beta\alpha\beta\alpha$ and α^3.

1.2 Previous Results

In 1906, Axel Thue [10, 11] proved that there are infinitely many overlap-free binary words, (or, equivalently: there are arbitrarily long overlap-free binary words). To do this, he constructed an infinite overlap-free word, $\theta^\omega(a)$, where θ is the morphism:

$$\begin{array}{rccc} \theta\colon & A^* & \longrightarrow & A^* \\ & a & \longmapsto & ab \\ & b & \longmapsto & ba \end{array}$$

and $\theta^\omega(a)$ means the limit of the sequence $\left(\theta^k(a)\right)$:

$$\theta^\omega(a) = abbabaabbaababbabaababbaabbabaabba\ \ldots\ .$$

This infinite word $\theta^\omega(a)$ is overlap-free because the morphism θ is itself overlap-free, which means that if w is an overlap-free word, then $\theta(w)$ is overlap-free too.

The problem arising naturally is the following: let u_n be the number of overlap-free binary words of length n. Is it possible to find a bound to u_n? An equivalent? Recurrence relations? Here are some results about this problem.

A linear lower bound to u_n is given by the number of factors of length n in $\theta^\omega(a)$ (see [3, 5]):

$$\begin{cases} \text{If } 2.2^k \le n \le 3.2^k,\ \theta^\omega(a) \text{ has } 4n - 2.2^k \text{ factors of length } n+1 \\ \text{If } 3.2^k \le n \le 4.2^k,\ \theta^\omega(a) \text{ has } 2n + 4.2^k \text{ factors of length } n+1 \end{cases}$$

(this holds for all $k \ge 0$).

Restivo and Salemi [9] gave a polynomial upper bound to u_n: $u_n \le Cn^r$ with $C > 0$ and $r = \log_2 15 \simeq 3.906$.

Kobayashi [8] improved these bounds:

$$C_1 n^{1.155} \le u_n \le C_2 n^{1.587}\ .$$

The lower bound is obtained by counting the overlap-free words that are infinitely extensible on the right.

Carpi [4] proved that a finite automaton can be used to compute u_n, and gave a way to find upper bounds Cn^r with r arbitrarily near of the optimal value. However he didn't publish any numerical result.

2 A Bijection Between the Language of Overlap-Free Binary Words and a Rational Language

2.1 A Decomposition of Overlap-Free Binary Words

As we shall see later, we need to consider not only overlap-free words, but also words that contain only one overlap, which covers them entirely. We call them *simple overlaps.*

Definition: A word $w \in A^*$ is called a *simple overlap* if it can be written $w = xuxux$ with $x \in A$ and $u \in A^+$ such that $xuxu$ and $uxux$ are overlap-free.

Moreover, to avoid special cases as much as possible, we shall restrict ourselves to words of length at least 4. We define three sets U, V and S as follows:

U is the set of overlap-free words of length greater than 3 (13 words are excluded);

V is the set of simple overlaps of length greater than 3 (only aaa and bbb are excluded);

S is the finite set containing the elements of $U \cup V$ of length less than 8. This set has 76 elements.

Kfoury [7], Kobayashi [8] and Carpi [4] use results similar to the following lemma:

Lemma 2.1. *A word $w \in (U \cup V) \setminus S$ can be written $w = r_1\theta(u)r_2$ with u overlap-free and $r_i \in \{\varepsilon, a, b, aa, bb\}$. This decomposition is unique.*

Using this lemma, they construct overlap-free words by starting with a small word and alternatively applying θ and adding letters to the left or to the right of the current word. We shall use a similar approach, but instead of adding new letters we shall remove or modify existing letters.

Definition: Let $E = \{\delta, \iota, \kappa\}$, and G be the monoid $(E \times E)^*$. The unit of G is denoted by 1. We define an action of G on A^* in the following way:

For all $g \in G$, let $\varepsilon.g = \varepsilon$.

Given $\gamma = (\gamma_1, \gamma_2) \in E \times E$ and $w \in A^+$, we write the image of w by θ as $\theta(w) = x_1 u x_2$ with $u \in A^*$, $x_1 \in A$ and $x_2 \in A$, and $w.\gamma$ is given by the table:

	$\gamma_2 = \delta$	$\gamma_2 = \iota$	$\gamma_2 = \kappa$
$\gamma_1 = \delta$	u	$u\overline{x_2}$	ux_2
$\gamma_1 = \iota$	$\overline{x_1}u$	$\overline{x_1}u\overline{x_2}$	$\overline{x_1}ux_2$
$\gamma_1 = \kappa$	x_1u	$x_1u\overline{x_2}$	x_1ux_2

(recall that $\overline{a} = b$ and $\overline{b} = a$). For instance, if $w = aabb$ and $\gamma = (\delta, \iota)$, to compute $w.\gamma$ we take $\theta(w) = ababbaba$, we delete the first letter and invert the last one: $w.\gamma = babbabb$.

For all $w \in A^*$, let $w.1 = w$.

Finally, for $g \in G$, $\gamma \in E \times E$ and $w \in A^*$, let $w.(\gamma g) = (w.\gamma).g$.

Lemma 2.2. *A word $w \in (U \cup V) \setminus S$ can be uniquely written $w = v.\gamma$ with $\gamma \in E \times E$ and $v \in (U \cup V)$. Moreover, $|v| < |w|$.*

Proof: Lemma 2.1 gives a unique decomposition $w = r_1\theta(u)r_2$. We define $v = x_1ux_2$ and $\gamma = (\gamma_1, \gamma_2)$ as follows:

	$r_1 = \varepsilon$	$r_1 = a$	$r_1 = b$	$r_1 = aa$	$r_1 = bb$
x_1	ε	b	a	b	a
γ_1	κ	δ	δ	ι	ι

	$r_2 = \varepsilon$	$r_2 = a$	$r_2 = b$	$r_2 = aa$	$r_2 = bb$
x_2	ε	a	b	a	b
γ_2	κ	δ	δ	ι	ι

and it is easy to check that $w = v.\gamma$ and $3 < |v| < |w|$ (in fact $\frac{|w|}{2} \leq |v| \leq \frac{|w|}{2} + 1$).

The word u is overlap-free. If x_1ux_2 were not in $U \cup V$, x_1u would begin with an overlap, or ux_2 would end with one. Suppose that x_1u begins with an overlap: then $\overline{x_1}\theta(u)$ begins with an overlap too; this is only possible if $w = \overline{x_1}\theta(u) \in V$, and then $x_2 = \varepsilon$, hence $x_1ux_2 = x_1u \in V$. The other case is symmetric. Note that v can be in V even if w is in U, for instance $aabbaabb = babab.(\delta, \delta)$. That's why we have to consider $U \cup V$ instead of simply U.

The unicity of (v, γ) is a consequence of the unicity of (u, r_1, r_2). ∎

Let Z be the subset of $S \times G$ containing all elements except those of the form $(s, \gamma g)$ with $|s.\gamma| < 8$. To any $z = (s, g) \in Z$, we associate the word $[z] = s.g \in A^*$. Conversely, we can represent certain words by elements of Z:

Theorem 2.1. *Every word $w \in U \cup V$ has a unique decomposition $w = s.g$ with $(s, g) \in Z$.*

Proof: Let's prove this theorem by induction on the length of w. If $|w| < 8$, $w \in S$ and $(s, g) = (w, 1)$ is the only possibility according to the definition of Z. Now suppose that the theorem has been proved for words of length less than n; then if w is a word of length n, by Lemma 2.2 this word can be uniquely written $w = v.\gamma$ with $\gamma \in E \times E$, and $|v| < n$. By the inductive hypothesis, v has a unique decomposition $v = s.g$ hence w also has a unique decomposition $w = s.(g\gamma)$. ∎

Elements of Z should be viewed as words on the alphabet $S \cup (E \times E)$, with only their first letter in S. The set Z is clearly a rational language on this alphabet: the restriction $(s, \gamma g) \notin Z$ if $|s.\gamma| < 8$ (needed for the unicity in Theorem 2.1) amounts to defining Z as

$$Z = (S \times G) \setminus (S_4 \times \{(\delta, \delta), (\delta, \iota), (\delta, \kappa), (\iota, \delta), (\kappa, \delta)\}G)$$

where $S_4 = \{aaba, aabb, abaa, abab, abba, baab, baba, babb, bbaa, bbab\}$ is the set of the elements of length 4 in S.

2.2 The Rational Languages Representing U and V

Let $L = \{ z \in Z \mid [z] \in U \}$ and $M = \{ z \in Z \mid [z] \in V \}$. We shall prove that L and M are rational languages. To do this we have to answer the question: under which conditions (on z) is the word $[z]$ an element of U or of V?

First, let's see what can be said about $[z\gamma] = [z].\gamma$ when the status of $[z]$ is known, and $\gamma = (\gamma_1, \gamma_2) \in E \times E$.

- If $[z]$ is not in $U \cup V$, then $[z\gamma]$ cannot be in $U \cup V$ either.
- If $[z] \in U$, $\gamma_1 \in \{\delta, \kappa\}$ and $\gamma_2 \in \{\delta, \kappa\}$, then $[z\gamma] \in U$.
- If $[z] \in V$, then $[z(\iota, \kappa)]$, $[z(\kappa, \iota)]$ and $[z(\kappa, \kappa)]$ are not in $U \cup V$, $[z(\delta, \kappa)]$ and $[z(\kappa, \delta)]$ are in V, and $[z(\delta, \delta)]$ is in U.
- If $[z]$ does not begin with *aaba* or *bbab*, then $[z(\iota, \gamma_2)] \notin U \cup V$.
- If $[z]$ does not end with *abaa* or *babb*, then $[z(\gamma_1, \iota)] \notin U \cup V$.
- If $[z]$ begins with *aaba* and either $\gamma_2 \in \{\delta, \kappa\}$ and $[z] \in U$, or $\gamma_2 = \delta$ and $[z] \in V$, then $[z(\iota, \gamma_2)] \in U$: it begins with *bbabbaa*, and if there were an overlap, *bbabba* would appear somewhere else in this word, which is impossible. This result holds also if $[z]$ begins with *bbab*.
- If $[z]$ ends with *abaa* or *babb* and either $\gamma_1 \in \{\delta, \kappa\}$ and $[z] \in U$, or $\gamma_1 = \delta$ and $[z] \in V$, then by a similar argument $[z(\gamma_1, \iota)] \in U$.
- If $[z]$ begins with *aaba* or *bbab*, ends with *abaa* or *babb*, and is in $U \cup V$, then $[z(\iota, \iota)] \in U$.

We can see that we need to know a small prefix (at most four letters) and a small suffix of $[z]$ to conclude.

Definition: We define the *type* of a word $w \in U \cup V$ as the couple (i, j) where

$$\begin{cases} i = 1 & \text{if } w \text{ begins with } aba \text{ or } bab \\ i = 2 & \text{if } w \text{ begins with } abb \text{ or } baa \\ i = 3 & \text{if } w \text{ begins with } aaba \text{ or } bbab \\ i = 4 & \text{if } w \text{ begins with } aabb \text{ or } bbaa \end{cases} \quad \text{and} \quad \begin{cases} j = 1 & \text{if } w \text{ ends with } aba \text{ or } bab \\ j = 2 & \text{if } w \text{ ends with } bba \text{ or } aab \\ j = 3 & \text{if } w \text{ ends with } abaa \text{ or } babb \\ j = 4 & \text{if } w \text{ ends with } bbaa \text{ or } aabb \end{cases}$$

(as w contains neither *aaa* nor *bbb*, this is always defined). We call i the *prefix type* and j the *suffix type* of w.

Knowing the type (i, j) of $[z]$, one can compute the type (i', j') of $[z(\gamma_1, \gamma_2)]$ (when this word is in $U \cup V$) using the formula

$$(i', j') = (\varphi(i, \gamma_1), \varphi(j, \gamma_2))$$

where φ is the function defined by the following table:

	1	2	3	4
δ	4	3	1	1
ι			3	
κ	2	2	1	1

where an empty box means that $[z(\gamma_1, \gamma_2)]$ cannot be in $U \cup V$.

Note that φ is nothing else than the transition function of a 4-state automaton that computes the prefix type or the suffix type of $[z]$. Taking the direct product of this automaton by itself, we get a 16-state automaton computing the type of $[z]$: as far as type is concerned, there is no interaction between the beginning and the end of $[z]$.

If now we take two copies of this automaton (one for L and one for M), and we place arrows between these two copies according to the rules given at the beginning of the present section, we get an automaton that recognizes L or M depending on what final states we choose. This automaton as 32 states, plus a few states to initialize the process, minus a few states that can be suppressed because they are not reached.

3 Using Matrices to Compute u_n

We consider two sequences of 4 by 4 matrices, U_n and V_n:

$U_n(i,j)$ is the number of words in U of length n and of type (i,j),
$V_n(i,j)$ is the number of words in V of length n and of type (i,j).

Let δ, ι, κ be the matrices associated with the transformations δ, ι, κ, so that for example $\left(\kappa U_n{}^t\iota\right)(i,j)$ is the number of words $u.(\kappa,\iota)$ of type (i,j) for any $u \in U$ of length n:

$$\delta = \begin{pmatrix} 0&0&1&1\\ 0&0&0&0\\ 0&1&0&0\\ 1&0&0&0 \end{pmatrix}, \quad \iota = \begin{pmatrix} 0&0&0&0\\ 0&0&0&0\\ 0&0&1&0\\ 0&0&0&0 \end{pmatrix}, \quad \kappa = \begin{pmatrix} 0&0&1&1\\ 1&1&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{pmatrix}.$$

Taking into account the conditions given in Section 2, we can establish the

Theorem 3.1. *The matrices U_n and V_n defined above verify the following recurrence formulas:*

$$\left\{\begin{array}{lcccccccc} V_{2n} &=& 0 &&&&&& \\ V_{2n+1} &=& \kappa V_{n+1}{}^t\delta &+& \delta V_{n+1}{}^t\kappa &&&& \\ U_{2n} &=& \iota V_n{}^t\iota &+& \delta V_{n+1}{}^t\delta &+& (\kappa+\iota)U_n{}^t(\kappa+\iota) &+& \delta U_{n+1}{}^t\delta \\ U_{2n+1} &=& \iota V_{n+1}{}^t\delta &+& \delta V_{n+1}{}^t\iota &+& (\kappa+\iota)U_{n+1}{}^t\delta &+& \delta U_{n+1}{}^t(\kappa+\iota) \end{array}\right.$$

for any integer n greater than 3. ■

Given the first few values:

$$V_4 = 0, \quad V_5 = \begin{pmatrix} 2&0&0&0\\ 0&0&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{pmatrix}, \quad V_6 = 0, \quad V_7 = \begin{pmatrix} 0&0&2&0\\ 0&2&0&0\\ 2&0&0&0\\ 0&0&0&0 \end{pmatrix},$$

$$U_4 = \begin{pmatrix} 2&0&2&0\\ 0&2&0&0\\ 2&0&0&0\\ 0&0&0&2 \end{pmatrix}, \quad U_5 = \begin{pmatrix} 0&2&2&0\\ 2&0&0&2\\ 2&0&2&0\\ 0&2&0&0 \end{pmatrix}, \quad U_6 = \begin{pmatrix} 2&2&0&2\\ 2&2&2&0\\ 0&2&2&0\\ 2&0&0&2 \end{pmatrix}, \quad U_7 = \begin{pmatrix} 4&2&0&2\\ 2&0&2&2\\ 0&2&0&2\\ 2&2&2&0 \end{pmatrix}$$

it is therefore possible to compute U_n and V_n in logarithmic time, and hence u_n:

Theorem 3.2. *The number u_n of overlap-free binary words can be computed with the recurrence formulas of Theorem 3.1 and the relation*

$$u_n = \sum_{i,j} U_n(i,j)$$

which holds for $n > 3$ ($u_0 = 1$, $u_1 = 2$, $u_2 = 4$, $u_3 = 6$).

The sequence u_n is therefore a 2-regular sequence *in the sense of [1].* ■

Remark: The recurrence for V_n can be solved:

$$V_5,\ V_7 \text{ as above, } V_9 = \begin{pmatrix} 0&0&0&0\\ 0&0&0&2\\ 0&0&0&0\\ 0&2&0&0 \end{pmatrix}, \quad V_{13} = \begin{pmatrix} 0&2&0&2\\ 2&0&2&0\\ 0&2&0&0\\ 2&0&0&0 \end{pmatrix},$$

$$V_{2^k+1} = \tfrac{2^{k-3}}{3}\begin{pmatrix} 0&2&1&1\\ 2&0&1&1\\ 1&1&0&0\\ 1&1&0&0 \end{pmatrix} + \tfrac{2(-1)^k}{3}\begin{pmatrix} 0&1&2&-1\\ 1&0&-1&-1\\ 2&-1&0&0\\ -1&-1&0&0 \end{pmatrix},$$

$$V_{3.2^{k-1}+1} = 2^{k-3}\begin{pmatrix} 0&2&1&1\\ 2&0&1&1\\ 1&1&0&0\\ 1&1&0&0 \end{pmatrix} \text{ for } k > 3;\ V_n = 0 \text{ otherwise.}$$

Summing the coefficients we get the value of the number of simple overlaps, $v_n = \sum_{i,j} V_n(i,j)$ for $n > 3$: the only nonzero values are $v_{2^k+1} = 2^{k-1}$ and $v_{3.2^k+1} = 3.2^k$ for $k \geq 1$. Hence $v_n < n$.

4 Asymptotic Study of u_n

4.1 Simpler Recurrence Relations

It is easier to deal with vector sequences than matrix sequences: taking into account the fact that U_n and V_n are symmetric, and eliminating the components of V_n which are always zero, we can code both matrices by a 15-dimensional vector:

$$\begin{aligned} X_n \quad = \quad & \big(U_n(1,1), U_n(1,2), U_n(1,3), U_n(1,4), U_n(2,2), U_n(2,3), U_n(2,4), \\ & U_n(3,3), U_n(3,4), U_n(4,4), V_n(1,2), V_n(1,3), V_n(1,4), V_n(2,3), V_n(2,4)\big) \ . \end{aligned}$$

The recurrence relations become, for $n > 6$:

$$\begin{cases} X_{2n} = AX_n + BX_{n+1} \\ X_{2n+1} = CX_{n+1} \end{cases}$$

where A, B and C are 15 by 15 matrices:

$$A = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} ,$$

$$B = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} ,$$

$$C = \begin{pmatrix} 0&0&0&0&0&0&0&2&4&2&0&0&0&0&0 \\ 0&0&1&1&0&1&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&1&1&1&0&0&0&0&0&0 \\ 0&0&1&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&1&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&2&0&0&0&0&0&0&0&2&0 \\ 0&0&1&0&0&0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&1&1&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&1 \\ 0&0&0&0&0&0&0&0&0&0&0&1&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&0 \end{pmatrix} .$$

We can also write $Y_n = \begin{pmatrix} X_{n+1} \\ X_{n+2} \end{pmatrix}$, $F = \begin{pmatrix} C & 0 \\ A & B \end{pmatrix}$ and $G = \begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$. Then $Y_{2n} = FY_n$ and $Y_{2n+1} = GY_n$ for $n > 6$.

Remark: These recurrence relations may seem much simpler than those given in Theorem 3.1; however the complexity of the problem is hidden in the huge matrices F and G.

4.2 The Exponents α and β

As the growth of u_n is polynomial, it is natural to study the numbers

$$\alpha = \sup \{ \ r \ \mid \exists C > 0, \forall n, u_n \geq Cn^r \ \}$$

and

$$\beta = \inf \{ \ r \ \mid \exists C > 0, \forall n, u_n \leq Cn^r \ \} \quad .$$

Theorem 4.1. *$\alpha < \beta$.*

Note: We say that a sequence f_n *grows as* another sequence g_n if there are two positive constants C_1 and C_2 such that $C_1 g_n \leq f_n \leq C_2 g_n$ for n large enough (i.e. $f_n = \Theta(g_n)$).

Proof: First consider the subsequence $Y_{7.2^k} = F^k Y_7$. The matrix F has one simple eigenvalue of maximal modulus $\rho(F)$, (the spectral radius of F, which is approximately 2.421), and it is easy to check that Y_7 is not in the sum of the characteristic spaces corresponding to the other eigenvalues. Hence at least one element of $Y_{7.2^k}$ grows as $\rho(F)^k$. As $u_{7.2^k+1} + v_{7.2^k+1} + u_{7.2^k+2} + v_{7.2^k+2}$ is a combination with positive coefficients of these elements, it also grows as $\rho(F)^k$, therefore either $u_{7.2^k+1}$ or $u_{7.2^k+2}$ grows as $\rho(F)^k$ (we have seen that v_n has a linear upper bound). We have found a subsequence of u_n growing as n^r with $r = \log_2 \rho(F) \simeq 1.275$, therefore $\alpha < 1.276$.

Then consider the subsequence $Y_{\frac{22.4^k-1}{3}} = (GF)^k Y_7$. It grows, by similar arguments, as $\rho(GF)^k$ with $\rho(GF) \simeq 6.340$. Hence there is a subsequence of u_n growing as n^r with $r = \log_4 \rho(GF) \simeq 1.3322$, therefore $\beta > 1.332$. ■

Remark: These results, together with Kobayashi's, yield the following bounds:

$$1.155 < \alpha < 1.276 < 1.332 < \beta < 1.587 \ .$$

4.3 Partial Sums

Let $s_n = \sum_{k=0}^{n-1} u_k$be the number of overlap-free words of length less than n. We can precisely describe the asymptotic behaviour of this sequence:

Theorem 4.2. *The sequence s_n grows as n^r, with $r = \log_2 \zeta \simeq 2.310$, where*

$$\zeta = \rho(F+G) = \frac{3}{2} + \sqrt{3} + \sqrt{\frac{5}{4} + \sqrt{3}}$$

is the largest root of the polynomial $X^4 - 6X^3 + 5X^2 + 4$.

Proof: Let $S_n = \sum_{k=7}^{n-1} Y_k$, for $n > 6$. Then

$$S_{2n} = \sum_{k=7}^{13} Y_k + \sum_{k=7}^{n-1} Y_{2k} + \sum_{k=7}^{n-1} Y_{2k+1} = S_{14} + (F+G)S_n \ .$$

One can check that $F + G - 1$ is a regular matrix, so let $T = (F + G - 1)^{-1} S_{14}$: then $S_{2n} + T = (F+G)(S_n + T)$, and $S_{7.2^k} = (F+G)^k T - T$, therefore at least one element of $S_{7.2^k}$ grows as ζ^k. As before, we deduce that $s_{7.2^k}$ also grows as ζ^k.

If $7.2^k \leq n \leq 7.2^{k+1}$, then $s_{7.2^k} \leq s_n \leq s_{7.2^{k+1}}$, so s_n grows as $\zeta^{\log_2 n}$, i.e. $n^{\log_2 \zeta}$. ∎

References

[1] J.-P. ALLOUCHE AND J. SHALLIT, The ring of k-regular sequences, *Theoret. Comput. Sci.* **98** (1992), 163–197.

[2] F.-J. BRANDENBURG, Uniformely growing k-th power-free homomorphisms, *Theoret. Comput. Sci.* **23** (1983), 69–82.

[3] S. BRLEK, Enumeration of factors in the Thue-Morse word, *Discr. Appl. Math.* **24** (1989), 83–96.

[4] A. CARPI, Overlap-free words and finite automata, *Theoret. Comput. Sci.* **115** (1993), 243–260.

[5] A. DE LUCA AND S. VARRICCHIO, Some combinatorial properties of the Thue-Morse sequence and a problem in semigroups, *Theoret. Comput. Sci.* **63** (1989), 333–348.

[6] P. GORALČIK AND T. VANIČEK, Binary patterns in binary words, *Inter. J. Algebra Comput.* **1**,3 (1991), 387–391.

[7] A. J. KFOURY, A linear time algorithm to decide whether a binary word contains an overlap, *Theoret. Inform. Appl.* **22** (1988), 135–145.

[8] Y. KOBAYASHI, Enumeration of irreducible binary words, *Discr. Appl. Math.* **20** (1988), 221–232.

[9] A. RESTIVO AND S. SALEMI, Overlap-free words on two symbols, in *Automata on Infinite Words*, M. Nivat and D. Perrin, eds, pp. 198–206, no. 192 in *Lect. Notes Comp. Sci.*, Springer-Verlag, 1985.

[10] A. THUE, Über unendliche Zeichenreihen, *Norske Vid. Selsk. Skr., I. Mat. Nat. Kl., Christiana* **7** (1906), 1–22.

[11] A. THUE, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, *Norske Vid. Selsk. Skr., I. Mat. Nat. Kl., Christiana* **1** (1912), 1–67.