Thesis for the Degree of Licentiate of Philosophy

Symbolic Sequences, Crucial Words and Iterations of a Morphism

Sergey Kitaev

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden

Göteborg, October 2000

Symbolic Sequences, Crucial Words and
Iterations of a Morphism
Sergey Kitaev

ISSN 0347-2809/NO 2000:62

Department of Mathematics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden

Göteborg, Sweden, October 2000

Abstract

In the present thesis there are included three papers. In the first paper we introduce the concepts of sets of prohibitions, complete sets and crucial words. The basic purpose of the paper is to investigate the extremal lengths of crucial words for four chosen sets of prohibitions. Two of these sets have been considered in different branches of mathematics such as number theory, algebra and dynamical systems. The other two papers are concerned with the concept of morphism. Since there are many techniques to study sequences generated by morphisms, it is reasonable to ask if a given sequence can be obtained by iteration of a morphism. We consider two sequences that have been used to solve different problems. These sequences are the Arshon sequence and the sequence of σ. We prove that these sequences cannot be defined by iteration of a morphism.

Keywords: Symbolic sequence, combinatorics on words, morphism, iteration of a morphism, set of prohibitions, complete set of prohibitions, crucial word, Arshon sequence, sequence of σ

AMS 2000 subject classification: 68R15, 11B85

Acknowledgements

First of all I would like to thank Alexander Evdokimov for suggesting interesting unsolved problems to me, for his permanent interest in my work and his encouragement and support during my student life. I would also like to thank Sergei Avgustinovich for useful discussions as well as being the first one to show me the beauty of discrete mathematics when I was a pupil of the Physics and Mathematics School at Novosibirsk State University. I would like to thank Einar Steingrímsson for helpful remarks on my papers, for good advice and for making me feel very welcome during my stay in Sweden.

I thank all the people at the Laboratory of Discrete Analysis of Sobolev Institute of Mathematics and all the people at the Department of Mathematics of Chalmers University of Technology and Göteborg University for creating such a friendly atmosphere.

Special thanks to all my friends for making my life full of fun.

Finally, I would like to thank my parents for supporting me during all my life, the Shuiskii family, my aunt Valentina Semenova for pointing out the right way of living and a special thanks to my wife, Dariya, for continuous support and patience.

Sergey Kitaev
Göteborg, October 2000

Contents

Preface

Combinatorics on words is a relatively new subject. The reasons to study it come from different branches of mathematics (algebra, discrete mathematics, number theory, dynamical systems and probability), computer science (coding theory and complexity of formal languages), physics (structure of quasicrystals) and biology (investigating DNA).

The goal of the subject is to study various combinatorial properties of finite and infinite words on a finite set of symbols which usually is called an *alphabet*.

The history of combinatorics on words begins with work of Axel Thue on nonrepetitive sequences of symbols at the beginning of this century. However, only in 1983, after the publication of Lothaire's book (*Combinatorics on Words*) has the field developed systematically and rapidly.

In the present thesis there are included three papers.

In the first paper we introduce the concepts of *sets of prohibitions*, *complete sets* and *crucial words*.

A *set of prohibitions* $\mathbf{S}$ is a set of words in some alphabet $\mathbf{A}$.

If there exists an infinite sequence that has no words from $\mathbf{S}$ as subword, then the set of prohibitions $\mathbf{S}$ is called an *incomplete set*. Otherwise it is called a *complete set*.

A *crucial word* X is a word in the alphabet $\mathbf{A}$ that has two properties:
1) X has no words from $\mathbf{S}$ as subwords
2) For any letter $a \in \mathbf{A}$ the word Xa has a word from $\mathbf{S}$ as its subword (in other words Xa is not free from $\mathbf{S}$).

The notion of a complete set is related to the notion of *unavoidable set* which appears in the literature.

The basic purpose of the paper is to investigate the extremal *lengths* of crucial words for four chosen sets of prohibitions. These sets are $\mathbf{S}_1^n$, $\mathbf{S}_2^n$, $\mathbf{S}_3^{n,k_1,\ldots,k_n}$, $\mathbf{S}_4^{n,k}$, where the parameter n is the number of letters in an alphabet.

To be more precise, by $\mathbf{S}_1^n$ we prohibit the repetition of two equal consecutive subwords; by $\mathbf{S}_2^n$ we prohibit the repetition of two equal consecutive subwords that have the same number of occurrences of $a_i \in \mathbf{A}$ for $i = 1, \ldots, n$; by $\mathbf{S}_3^{n,k_1,\ldots,k_n}$ we prohibit words, in which the number of letters a_i is congruent with zero modulo k_i for $i = 1, \ldots, n$; by $\mathbf{S}_4^{n,k}$ we prohibit any two consecutive subwords of the length greater then k such that the number of positions in which these words differ is less then or equal to k.

The choice of these sets is not random. The interest in $\mathbf{S}_1^n$ appeared in number theory, algebra and dynamical systems. One was considering the problems related to the question on existence or non-existence of infinite sequences without repetitions.

The set of prohibitions $\mathbf{S}_2^n$ is a generalisation of the first one. For a long time one was not able to answer to the so called "*New Problem of Four Colours*". The problem was to construct in a four letter alphabet an infinite sequence that avoids (has no subwords from) $\mathbf{S}_2^4$. A construction of such a sequence was given in 1991, with the help of computer.

The remaining two sets do not seem to have appeared in the literature, but they certainly have some combinatorial interest. For the set $\mathbf{S}_4^{n,k}$ and any $n \geq 3$ we prove the theorem of *incompleteness* of this set of prohibitions.

The other two papers are concerned with the concept of *morphism*.

For those who are not familiar with this concept we recommend the book by Salomaa *"Jewels of Formal Language Theory"*. This book is a nice introduction to the subject under consideration.

Let Σ be an alphabet. A map $\varphi : \Sigma^* \to \Sigma^*$ is called a *morphism*, if we have $\varphi(uv) = \varphi(u)\varphi(v)$ for any $u, v \in \Sigma^*$. It easy to see that a morphism φ can be defined by defining $\varphi(i)$ for each $i \in \Sigma$.

Since there are many techniques to study sequences generated by morphisms, it is reasonable to ask if a given sequence can be obtained by iteration of a morphism.

In these two papers we consider two sequences that were used to solve different problems.

The *Arshon sequence* has the property that it is free from the set of prohibitions $\mathbf{S}_1^n$. There exists a lot of sequences with this property, but the Arshon sequence was the first one, constructed in 1937. This sequence is still interesting from different points of view.

The *sequence of* σ was used by Evdokimov to build chains of maximal length in the n-dimensional unit cube in 1969. We note that this sequence has at least two different definitions.

In 1979, Berstel proved that the Arshon sequence can be generated by a *tag-system*, but cannot be obtained by iteration of a morphism. He proved the latter of these by proving a more general statement. In this thesis we give an alternative proof of this fact. An interesting aspect of this proof is that it includes a consideration of the letter frequencies in the Arshon sequence.

Finally, we prove that the sequence of σ cannot be defined by iteration of a morphism.

So far there are unfortunately no universal criteria to determine whether a given sequence can be obtained by iteration of a morphism. The purpose of the papers presented here is to demonstrate different approaches to this problem.

Chapter 1

On Crucial Words for Some Sets of Prohibitions

Abstract

Introduced the notation of a set of prohibitions; given definitions of a complete set and a crucial word with respect to a given set of prohibitions. Considered 4 particular sets which appear in different areas of mathematics and for each of them examined the length of a crucial word. One of these sets proved to be incomplete.

1.1 Introduction and Background

In defining or characterising sets of objects in discrete mathematics, "languages of prohibitions" are often used to define a class of objects by listing those prohibited subobjects that are not contained in the objects of the class. To this end the notion of a subobject is defined in different ways. The notion depends on the set under consideration. These sets are subwords for partially bounded languages, subgraphs for families of graphs and so on. One of the classes of interest that have appeared and are considered in different areas of mathematics is the class of nonrecurrent symbolic sequences defined by prohibiting strong periodicity in them, or, to be more exact, by prohibiting the repetition of subwords in these symbolic sequences, for example of type XX.

In this paper we consider 4 types of "prohibitions" connected with a generalisation of the notion of nonrecurrent symbolic sequences, and for each of these sets we consider the structure of crucial words and find their lengths.

Let $\mathbf{A} = \{a_1, \ldots, a_n\}$ be an alphabet of n letters. A *word* in the alphabet $\mathbf{A}$ is a finite sequence of letters of the alphabet. Any i consecutive letters of a word X generate a *subword* of length i. If X is a subword of a word Y, we write $X \subseteq Y$.

The set $\mathbf{A}^*$ is the set of all the words of the alphabet $\mathbf{A}$. Let $\mathbf{S} \subseteq \mathbf{A}^*$. Then $\mathbf{S}$ is called a *set of prohibited words*. A word that does not contain any words from $\mathbf{S}$ as its subwords is called *free* from $\mathbf{S}$. The set of all words that are free from $\mathbf{S}$ is denoted by $\widehat{\mathbf{S}}$.

Example 1. Let $\mathbf{A} = \{a,\ b\}$. The set of prohibitions is $\mathbf{S} = \{aa,\ ba\}$. The word $abbb$ is in $\widehat{\mathbf{S}}$.

If there exists a $k \in \mathbf{N}$ with the property that the length of any word in $\widehat{\mathbf{S}}$ is less than k, then the set of prohibitions $\mathbf{S}$ is called a *complete* set.

Example 2. $\mathbf{A} = \{1,\ 2,\ 3,\ 4\}$. The set of prohibitions is

$$\mathbf{S} = \{123,\ 13,\ 14,\ 11,\ 22,\ 33,\ 44\ \}.$$

Then $\mathbf{S}$ is incomplete, since the word $\underbrace{124124\dots124}_{3k}$ is in $\widehat{\mathbf{S}}$ for any k.

Example 3. $\mathbf{A} = \{1, 2, 3\}$. The set of prohibitions is

$$\mathbf{S} = \{12,\ 23,\ 31,\ 32,\ 11,\ 22,\ 33\}.$$

Simple sorting shows $\mathbf{S}$ to be complete.

A word $X \in \widehat{\mathbf{S}}$ is called a *crucial* word (with respect to $\mathbf{S}$), if the word Xa_i contains a prohibited subword for any letter $a_i \in \mathbf{A}$. This means that Xa_i has the structure BB_ia_i, where B is some word and $B_ia_i \in \mathbf{S}$. The subword B_i is called the *i-ending* of crucial word X. If for each letter of the alphabet we consider minimal i-ending (with respect to inclusion) we obtain a system of included i-endings, which we will use to investigate crucial words.

Example 4. $\mathbf{A} = \{a,\ b,\ c\}$. The set of prohibitions is $\mathbf{S} = \{aa,\ cab,\ acac\}$. The word $abaca$ is crucial with respect to $\mathbf{S}$.

A crucial word of minimal (maximal) length, if it exists, is called a *minimal* (*maximal*) crucial word.

Example 5. $\mathbf{A} = \{a,\ b,\ c\}$. The set of prohibitions is $\mathbf{S} = \{aa,\ cab,\ acac\}$. The word aca is a minimal crucial word with respect to $\mathbf{S}$. There do not exist any maximal crucial words, since the word $\underbrace{b\dots b}_{k}aca$ is crucial for all $k \in \mathbf{N}$.

Let $L_{min}(\mathbf{S})$ $(L_{max}(\mathbf{S}))$ denote of the length of a minimal (maximal) crucial word with respect to $\mathbf{S}$.

In this paper we consider four sets of prohibitions denoted $\mathbf{S}_1^n$, $\mathbf{S}_2^n$, $\mathbf{S}_3^{n,k_1,\ldots,k_n}$, $\mathbf{S}_4^{n,k}$. Here we use n for indicating of the number of letters of the alphabet under consideration and $k, k_1, \ldots, k_n$ are natural numbers.

We now give the definitions of these sets:

$\mathbf{S}_1^n = \{XX \mid X \in \mathbf{A}^*\}$, that is, we prohibit the repetition of two equal consecutive subwords.

$\mathbf{S}_2^n = \{XY \mid \overline{\nu}(X) = \overline{\nu}(Y)\}$, where $\overline{\nu}(\mathrm{X}) = (\nu_1(X), \ldots, \nu_n(X))$ is the *content* vector of X, in which $\nu_i(X)$ is the number of occurrences of the letter a_i in X. That is, we prohibit the repetition of two consecutive subwords of the same content.

$\mathbf{S}_3^{n,k_1,\ldots,k_n} = \{X \mid \nu_i(X) \equiv 0 \pmod{k_i},\ k_i \in \mathbf{N},\ i = 1, \ldots, n\}$, that is, we prohibit words in which the number of letters a_i is congruent to zero modulo k_i for each $i = 1, \ldots, n$.

$\mathbf{S}_4^{n,k} = \{XY \mid d(X, Y) \leq k, |X| = |Y| \geq k+1, k \in \mathbf{N}\}$, where $d(X, Y)$ is the number of letters in which the words X and Y differ (Hamming metric) and $|X|$ is the length of the word X. That is we prohibit any two consecutive subwords of the length greater then k such that the number of positions in which these words differ is less then or equal to k.

The proofs of Theorems 1–6 consist of the constructions of extremal crucial words and of the proofs of their optimality, i. e. the lower bound for $L_{min}(\mathbf{S})$ and the upper bound for $L_{max}(\mathbf{S})$.

1.2 The Set of Prohibitions $\mathbf{S}_1^n$

Theorem 1. *We have*

$$L_{min}(\mathbf{S}_1^n) = 2^n - 1.$$

Proof:

We define a crucial word X by induction:

$$X_1 = a_1,\ X_i = X_{i-1}a_iX_{i-1},\ X = X_n.$$

From this construction it follows that $|X| = 2^n - 1$. We will prove that X is a minimal crucial word with respect to $\mathbf{S}_1^n$.

Let U be an arbitrary minimal crucial word. We show that U coincides with the word X up to a permutation of letters in $\mathbf{A}$.

From the definition of a crucial word it follows that in the word Ua_i there is a prohibited word of the form $B_ia_iB_ia_i$, where B_i is a certain word and $B_ia_iB_ia_i$ is the ending of the word Ua_i (the ending may coincide with Ua_i). In this case the i-ending is the subword $B_ia_iB_i$. Let $\ell_i = B_ia_iB_i$.

We assume that $\ell_1 \subset \ell_2 \subset \ldots \subset \ell_n$, since we can make such ordering by permuting the letters of the alphabet, which obviously does not affect the cruciality and minimality of a word.

Note that the minimal crucial word U has the form

$$U = B_na_nB_n = B_na_nY_na_1,$$

where Y_n is a certain word. Actually, if on the right of $B_na_nB_n$ there is a certain word, then it contradicts the minimality of a crucial word, and if instead of a_1 there stands a_k $(k > 1)$ then it contradicts $\ell_1 \subset \ell_k$.

We show that ℓ_{n-1} coincides with B_n. We have $\ell_{n-1} = B_{n-1}a_{n-1}B_{n-1}$ and let a_nB_n be a subword of ℓ_{n-1}. Now ℓ_{n-1} has the form $Ka_nPa_{n-1}Ka_nP$, where $Ka_nP = B_{n-1}$), but then

$$\ell_n = Pa_{n-1}Ka_nPa_{n-1}Ka_nP,\ \text{where}\ Pa_{n-1}Ka_nP = B_n,$$

and the word U contains the prohibited subword a_nPa_nP. This can not be the case. It means that ℓ_{n-1} is a subword of the word B_n, and the word U has the form:

$$U = \ell_n = Z_n\ell_{n-1}a_nZ_n\ell_{n-1},$$

where Z_n is a certain word. Since we explore a minimal crucial word, we have $Z_n = \emptyset$, and then $B_n = \ell_{n-1}$. In the same way we can show that $B_i = \ell_{i-1}$ for each $i = 2, \ldots, n-1$ and $B_1 = \emptyset$.

Hence the structure of a minimal crucial word U coincides with that of the word X as required.

Remark

From the proof of Theorem 1 it follows that the word X is the unique minimal crucial word to within a transposition of the letters of the alphabet $\mathbf{A}$.

1.3 The Set of Prohibitions $\mathbf{S}_2^n$

Proposition 1.

A minimal crucial (with respect to $\mathbf{S}_2^n$) word can not have three letters, each of which appears twice in the word.

Proof:

Since the proposition is obviously true for $|\mathbf{A}| = 1, 2, 3$, we will consider the case $|\mathbf{A}| \geq 4$.

Let X be a minimal crucial word, and suppose the system of included i-endings for it is $\ell_1 \subset \ell_2 \subset \ldots \subset \ell_n = X$. Suppose the letters a_{i_1}, a_{i_2}, a_{i_3} occur twice in X and that $i_1 < i_2 < i_3 < n$ (the fact that i_1, i_2, i_3 do not equal n follows from the fact that a_n must occur an odd number of times).

When we pass from ℓ_{i_3-1} to ℓ_{i_3} (ℓ_{i_3-1} is determined, since there are i_1, $i_2 < i_3$) there must appear a letter a_{i_3}, and when we pass from ℓ_{i_3} to ℓ_{i_3+1} (ℓ_{i_3+1} is determined, since $i_3 < n$) there must appear one more letter a_{i_3}; Hence, since there are two letters a_{i_3} in X, there are no letters a_{i_3} for $2 < j < i_3$ in ℓ_j whence there are no letters a_{i_3} in the X to the left of ℓ_{i_2} (both letters a_{i_3} lie to the left respecting of ℓ_{i_2}).

Obviously, the letter a_{i_1} must be in ℓ_{i_1}. The second letter a_{i_1} appears when we pass from ℓ_{i_1} to ℓ_{i_2}. Since there are only two letters a_{i_1}, there are no letters a_{i_1} in the word X to the left of ℓ_{i_2}.

If we write the letter a_{i_3+1} to the right of the word X we obtain a prohibited word (a word from $\mathbf{S}_2^n$). Words from $\mathbf{S}_2^n$ are divided into two parts which have the same content. Obviously, the letters a_{i_3} must be in different parts of the prohibited word, and letters a_{i_1} must be in different parts of the same word which is impossible, since the letters a_{i_3} lie strictly to the left of a_{i_1}, and this contradicts the assumption.

Remark.

From the proof of proposition 1 we have that if letters a_i and a_j occur twice in a word X (in which $\ell_1 \subset \ell_2 \subset \ldots \subset \ell_n = X$), then either $i = j + 1$ or $j = i + 1$.

Theorem 2. *For any $n > 2$ we have*

$$L_{min}(\mathbf{S}_2^n) = 4n - 7.$$

Proof:

Note that a natural approach to the construction of a crucial word is possible. It consists of an algorithm of step-by-step optimisation: We ascribe to a crucial word of an n-letter alphabet a minimum number of letters to obtain a crucial word of an $(n+1)$-letter alphabet.

The algorithm can be written recursively in the following way:

$$X_n = B_{n-1} a_n B_{n-1}$$
$$B_{n-1} = B_{n-3} a_{n-1} B_{n-3}$$
$$B_1 = a_1, B_2 = a_2, B_{-1} = B_0 = X_0 = \emptyset.$$

Some initial values when implementing the algorithm are:

$$X_1 = a_1,$$
$$X_2 = a_1 a_2 a_1,$$
$$X_3 = a_2 a_3 a_1 a_2 a_1,$$
$$X_4 = a_1 a_3 a_1 a_4 a_2 a_3 a_1 a_2 a_1,$$
$$X_5 = a_2 a_4 a_2 a_5 a_1 a_3 a_1 a_4 a_2 a_3 a_1 a_2 a_1.$$

This is an algorithm by which the minimal crucial word X_n for the set of prohibitions $\mathbf{S}_1^n$ can be built. For $\mathbf{S}_2^n$ such a construction gives an upper bound of the form $exp(n/2)$, or, to be more exact,

$$(3 - (n \bmod 2))2^{\lfloor \frac{n+1}{2} \rfloor} - 3.$$

We now give an upper bound that is a linear function.

We introduce, as before, a system of included i-endings: $\ell_1 \subset \ell_2 \subset \ldots \subset \ell_n$ (we permute the letters of the alphabet if it is necessary). We show that the passage from ℓ_{i-1} to ℓ_i is possible by adding only two symbols (letters of alphabet $\mathbf{A}$).

When we passed from ℓ_{i-1} to ℓ_i let there appear symbols y and z. ℓ_{i-1} may be denoted by AB, where A is a certain word, B consists of the letters of the word A (which are somehow mixed) and B contains one letter a_{i-1} less than A does. Let x be the last letter of the word A on the right. Then ℓ_i may be denoted by $yzKxB$, where $A = Kx$. From the definition of ℓ_i we have the equation

$$y \cup z \cup K = B \cup x \cup a_i.$$

which from the definition of K and B is equivalent to

$$2x \cup a_i = y \cup z \cup a_i.$$

It follows necessarily that $x = a_{i-1}$ and either $y = a_{i-1}$, $z = a_i$ or $y = a_i, z = a_{i-1}$.

Suppose $y = a_{i-1}$, $z = a_i$.

For example, we have the following crucial word for a 6-letter alphabet:

$$a_4a_5a_3a_4a_2a_3a_1a_2|a_6a_4a_3a_2a_1a_2a_3a_4a_6,$$

(the vertical line was drawn for a more convenient visual perception of the word).

This word is crucial and its length is equal to 17.

We consider a case of an arbitrary $n \geq 3$ defining the word W as

$$W = a_{n-2}a_{n-1}a_{n-3}a_{n-2}\dots a_1a_2|a_na_{n-2}a_{n-3}\dots a_2a_1a_2\dots a_{n-3}a_{n-2}a_n.$$

Then $|W| = 2(n-2) + n - 1 + n - 2 = 4n - 7$.

Let us verify that the word W is crucial.

If we write the letters a_1, a_2, a_n to the right of the word W we will obviously have prohibited subwords. Let $2 < i < n$. Then if we write the letters a_i we will have the prohibition

$$a_{i-1}a_i\dots a_1a_2a_na_{n-2}\dots a_i|a_{i-1}\dots a_2a_1a_2\dots a_{n-2}a_na_i,$$

since the composition vectors of the left and right subwords with respect to the vertical line are equal.

Before proving that $W \in \widehat{\mathbf{S}_2^n}$ we make the following remark.

In the word W we have $\ell_n \subset \ell_1 \subset \dots \subset \ell_{n-2} \subset \ell_{n-1}$. Substituting a_1 for a_n, a_2 for a_1, ..., a_n for a_{n-1} we obtain another word

$$U = a_{n-1}a_n\dots a_2a_3|a_1a_{n-1}\dots a_3a_2a_3\dots a_{n-1}a_1,$$

for which $\ell_1 \subset \ell_2 \subset \dots \subset \ell_n$.

In both cases (before and after substitution of letters of the alphabet) we have the construction of a crucial word (which will be proved below) hence the same upper bound of the length of a minimal crucial word.

For W it is more convenient to show further that $W \in \widehat{\mathbf{S}_2^n}$.

We rewrite W making in it the marks (1),(2), ...,(2n-4), which number the gaps between letters of a word like this:

$$(2n-4)a_{n-2}(2n-5)a_{n-1}\dots(2)a_1(1)a_2|a_na_{n-2}\dots a_2a_1a_2\dots a_{n-2}a_n.$$

In a possible prohibition we mark the left and right bounds. Note that the length of a prohibition is an even number, and each letter must occur an even number of times in a prohibition. The left bound of the prohibition must lie to the right of the mark (2n-5), since the letter a_{n-1} enters W once;

It must lie to the left of the mark (1), since to the right of the mark (1) there is one letter a_1.

Note that if m is even then (m) is not the left bound of the possible prohibition. Actually in this case two variants are possible:

1) the prohibition does not cover the left letter a_n.

2) the prohibition covers the left letter a_n.

In the second case we have not a prohibition, since if the prohibition begins from the even mark, then it can not cover the second a_n.

In the first case the right bound of the prohibition lies to the left of a_n, hence the letter $a_{\frac{m}{2}+1}$ enters the prohibition only once.

Suppose the prohibition begins from the mark (m) and m is odd.

There are two possible cases.

1) The prohibition does not cover the left letter a_n (this case is impossible since the letter $a_{\lfloor\frac{m}{2}\rfloor}$ occurs the prohibition once).

2) The prohibition covers the left a_n. Then it covers the right a_n too, and the letter $a_{\lfloor\frac{m}{2}\rfloor}$ occurs an odd number of times in the prohibition. So $W \in \widehat{\mathbf{S}_2^n}$ and hence $L_{min}(\mathbf{S}_2^n) \leq 4n-7$ for $n > 2$.

We give now a lower bound.

Since the length of a minimal crucial word must be odd, and the passage from ℓ_i to ℓ_{i+1} requires at least two letters, we have that a trivial lower bound of the length of a minimal crucial word is $2n-1$.

Let us now improve the lower bound. Obviously a minimal crucial word in which $\ell_1 \subset \ell_2 \subset \ldots \subset \ell_n$ has an even number of occurrences of the letter a_i for $i = 1, \ldots, n-1$ and an odd number of occurrences of the letter a_n. The word U has two letters a_1, two letters a_2, one letter a_n and four of any other letter. From proposition 1 we know that there does not exist a crucial word that has the fewer number of letters, hence the word U gives us the lower bound of the length of a minimal crucial word.

1.4 The Sets of Prohibitions $\mathbf{S}_3^{n,k_1,\dots,k_n}$ and $\mathbf{S}_4^{n,k}$

Theorem 3. *We have*

$$L_{min}(\mathbf{S}_3^{n,k_1,\dots,k_n}) = \sum_{i=1}^{n} k_i - 1.$$

Proof:

Let us give a lower bound.

Let X be a minimal crucial word. Considering a system of included i-endings gives us that X has at least $k_i t_i$ letters a_i, $i = 1, \dots, n-1$ and $k_n t_n - 1$ letters a_n, where $t_i \in \mathbf{N}$, $i \in \{1, \dots, n\}$ and hence:

$$L_{min}(\mathbf{S}_3^{n,k_1,\dots,k_n}) \geq \sum_{i=1}^{n} k_i - 1.$$

An upper bound is given by the construction

$$a_{n-1} \underbrace{a_n \dots a_n}_{k_n - 1} a_{n-2} \underbrace{a_{n-1} \dots a_{n-1}}_{k_{n-1} - 1} \dots a_1 \underbrace{a_2 \dots a_2}_{k_2 - 1} \underbrace{a_1 \dots a_1}_{k_1 - 1}$$

Obviously this word is crucial and its length equals $\sum_{i=1}^{n} k_i - 1$.

Theorem 4. *We have*

$$L_{min}(\mathbf{S}_4^{n,k}) = 2k + 1.$$

Proof:

For the set of prohibitions $\mathbf{S}_4^{n,k}$ we must have $|A| = |B| \geq k + 1$, where AB is an arbitrary prohibition. So we have

$$L_{min}(\mathbf{S}_4^{n,k}) \geq 2k + 1.$$

An upper bound is given by the construction $p_1 p_2 \dots p_k x p_1 p_2 \dots p_k$, where $x, p_i \in \mathbf{A}$, $i = 1, \dots, k$ and $x \neq p_i$.

Theorem 5. *We have*

$$L_{max}(\mathbf{S}_3^{n,k_1,\ldots,k_n}) = \prod_{i=1}^{n} k_i - 1.$$

Proof:

We give an upper bound.

Any word $z_1 \ldots z_\ell$ free from $\mathbf{S}_3^{n,k_1,\ldots,k_n}$ has the property that all vectors $\overline{\nu}(z_1 \ldots z_i)$ are different modulo $\mathbf{k} = (k_1, \ldots, k_n)$ for $i = 1, 2, \ldots, \ell$, since else, if for example $\overline{\nu}(z_1 \ldots z_j) = \overline{\nu}(z_1 \ldots z_i)$ and $j > i$, we have

$$\overline{\nu}(z_{i+1} \ldots z_j) = \overline{\nu}(z_1 \ldots z_j) - \overline{\nu}(z_1 \ldots z_i) \equiv \overline{0} \pmod{\mathbf{k}},$$

which contradicts the word $z_1 \ldots z_\ell$ is free from $\mathbf{S}_3^{n,k_1,\ldots,k_n}$.

The number of different words whose content vector is not congruent to the zero vector modulo $\mathbf{k}$ is $\prod_{i=1}^{n} k_i - 1$, hence

$$L_{max}(\mathbf{S}_3^{n,k_1,\ldots,k_n}) \leq \prod_{i=1}^{n} k_i - 1.$$

A lower bound is given by the construction

$$A_1 = a_1 \ldots a_1 \ (k_1 - 1 \text{ letters } a_1)$$

$$A_n = A_{n-1} a_n A_{n-1} \ldots A_{n-1} a_n A_{n-1} \ (k_n \text{ blocks } A_{n-1}).$$

This word is obviously crucial and its length is $\prod_{i=1}^{n} k_i - 1$.

Remark.

The crucial word with respect to $\mathbf{S}_4^{1,k}$ is unique and its length is $2k + 1$.

Theorem 6. *We have*

$$L_{max}(\mathbf{S}_4^{2,k}) = 3k + 3.$$

Proof:

Let

$$\bar{a} = \begin{cases} 1, & \text{if } a = 2, \\ 2, & \text{if } a = 1. \end{cases}$$

Moreover, let us consider an arbitrary crucial word A, with respect to $\mathbf{S}_4^{2,k}$, of length greater then $3k+3$. It is easy to see that if $a_1a_2\dots a_{k+1}$ are the first $k+1$ letters of A then the next $k+1$ letters of A must be $\bar{a}_1\bar{a}_2\dots\bar{a}_{k+1}$, because otherwise the first $2k+2$ letters of A will form a prohibited subword. By the same argument, we can show that

$$A = a_1a_2\dots a_{k+1}\bar{a}_1\bar{a}_2\dots\bar{a}_{k+1}a_1a_2\dots a_{k+1}\bar{a}_1\dots.$$

Let us consider the subwords A_i of A of the length $2k+4$ which start from the ith letter, where $1 \le i \le k$:

$$A_i = \underbrace{a_ia_{i+1}\dots a_{k+1}\bar{a}_1\dots\bar{a}_i}_{k+2}\underbrace{\bar{a}_{i+1}\dots\bar{a}_{k+1}a_1\dots a_{i+1}}_{k+2}$$

If $a_i = \bar{a}_{i+1}$ then the underbraced subwords of A_i are the same in the first and in the last positions, so they differ in at most k positions, hence A_i is prohibited. So we must have $a_i = a_{i+1}$ for $i = 1,\dots,k$.

Without loss of generality we can assume that $a_1 = 1$, so

$$A = \underbrace{11\dots1}_{k+1}\underbrace{22\dots2}_{k+1}\underbrace{11\dots1}_{k+1}2\dots.$$

It is easy to see that if the length of A is greater then $3k+3$ then A has a prohibited subword of length $2k+4$:

$$A = \underbrace{11\dots1}_{k}\overbrace{1\underbrace{22\dots2}_{k+1}}\overbrace{\underbrace{11\dots1}_{k+1}2}2\dots.$$

(here and then two braces above an word show us a disposition of a prohibited subword and, in particular, a disposition of parts of this subword that correspond to X and Y from the definition of the set of prohibitions $\mathbf{S}_4^{n,k}$).

So $L_{max}(\mathbf{S}_4^{2,k}) \le 3k+3$.

To prove the theorem it is sufficient to check that there are no prohibited subwords in the word $A = \underbrace{11\dots1}_{k+1}\underbrace{22\dots2}_{k+1}\underbrace{11\dots1}_{k+1}$.

Obviously the left end of a possible prohibition can be only in the left block $\underbrace{1\ldots1}_{k+1}$:

$$\overbrace{\underbrace{1\ldots1}_{j}\underbrace{2\ldots2}_{i}}\overbrace{\underbrace{2\ldots2}_{k-i+1}\ \underbrace{1\ldots1}_{2i+j-k-1}}$$

with

$$j+i\geq k+1 \tag{1.1}$$

Two cases are possible:

1. $j\geq k-i+1$
2. $j<k-i+1$

In the first case there is non-coincidence between the left and right parts of the prohibition in the first $k-i+1$ letters and in the last i letters that is non-coincidence in $k+1$ letters. So this case is impossible.

In the second case we have non-coincidence in the first j letters and in the last $2i+j-k-1$ letters. Hence we have non-coincidence in $2(i+j)-k-1$ letters, that according to (1) is greater than or equal to $k+1$.

It follows that the word $\underbrace{1\ldots1}_{k+1}\underbrace{2\ldots2}_{k+1}\underbrace{1\ldots1}_{k+1}$ does not contain a prohibition and thus the theorem is proved.

Theorem 7. *[Incompleteness] The set of prohibitions* $\mathbf{S}_4^{n,k}$ *for* $n\geq 3$ *is incomplete.*

Proof:

Since the alphabet $\mathbf{A}$ is finite, there is no trivial solution of the problem (such as taking all letters of $\mathbf{A}$ and obtaining an infinite sequence with the properties needed). So to prove the incompleteness of the set $\mathbf{S}_4^{n,k}$ we have to show the existence of an infinite word which is free from the set of prohibitions $\mathbf{S}_4^{n,k}$.

We consider the case $n=3$ and the alphabet $\mathbf{A}=\{1,2,3\}$, since the incompleteness of the set of prohibitions $\mathbf{S}_4^{n,k}$ for the case $n>3$ will follow from the incompleteness of the set of prohibitions for the case $n=3$.

Let $\mathbf{B}=\{a,b,c\}$ be an alphabet. $\mathbf{B}^*$ is the set of all words of the alphabet $\mathbf{B}$.

We define the mapping f as follows:

$$\underbrace{1\ldots1}_{k+1} \to a,\ \underbrace{2\ldots2}_{k+1} \to b,\ \underbrace{3\ldots3}_{k+1} \to c.$$

The domain of the mapping f is the set of words of the alphabet

$$\mathbf{C} = \{\ \underbrace{1\ldots1}_{k+1},\ \underbrace{2\ldots2}_{k+1},\ \underbrace{3\ldots3}_{k+1}\ \}.$$

The image of the mapping f is the set $\mathbf{B}^*$.

Let the set of prohibitions $\mathbf{S}' = \{XX | X \in \mathbf{B}^*\}$. Obviously, the set $\mathbf{S}'$ coincides with the set $\mathbf{S}_2^n$ whenever $\mathbf{A} = \mathbf{B}$.

It is known [1] that for the alphabet $\mathbf{B}$ there exists the infinite sequence L' which is free from the set of prohibitions $\mathbf{S}'$. L' is built by iteration of morphisms:

$$a \to abc$$

$$b \to ac$$

$$c \to b$$

The morphism iteration procedure is as follows.

We start from the letter a. Then we substitute this letter with abc. Then we substitute each letter in abc by the rule above. We obtain after this step $abcacb$. And so on. Executing this procedure an infinite number of times gives us the sequence L'.

Let us prove that the sequence $L = f^{-1}(L')$ does not contain words prohibited by $\mathbf{S}_4^{3,k}$.

We are going to prove the statement by considering L and all possible dispositions of words prohibited by $\mathbf{S}_4^{3,k}$.

The sequence L is built up from the letters of the alphabet $\mathbf{C}$ or in other words from the *blocks* $\underbrace{x\ldots x}_{k+1}$, where $x \in \{1, 2, 3\}$. It means that there are only three different cases for a disposition of a possible prohibition in L.

Case 1. $\overbrace{\underbrace{x\ldots x}_{k+1}\ldots\underbrace{y\ldots y}_{k+1}}\overbrace{\underbrace{z\ldots z}_{k+1}\ldots\underbrace{t\ldots t}_{k+1}}$;

Case 2. $\underbrace{x\ldots x}_{i}\overbrace{\underbrace{x\ldots x}_{k-i+1}\ldots\underbrace{y\ldots y}_{k+1}}\overbrace{\underbrace{z\ldots z}_{k+1}\ldots\underbrace{t\ldots t}_{k-i+1}}\underbrace{t\ldots t}_{i}$, where $0 < i < k+1$;

Case 3. $\underbrace{x\dots x}_{i}\overbrace{\underbrace{x\dots x}_{k-i+1}\dots\underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1}\dots\underbrace{t\dots t}_{k-j+1}}\underbrace{t\dots t}_{j}$,
where $0 \le i, j, l \le k+1$.

Now we will consider these cases and show that each of them is impossible.

Case 1. Let $\mathbf{P}$ denote the prohibited subword (prohibition) under consideration, $\mathbf{R}$ and $\mathbf{L}$ denote the right and the left parts of $\mathbf{P}$ respectively.

It is obvious that $\mathbf{L}$ and $\mathbf{R}$ have the same number of blocks. Moreover, the ith block of $\mathbf{L}$ (from the left to the right) is equal to the ith block of $\mathbf{R}$, because otherwise we have non-coincidence of $\mathbf{L}$ and $\mathbf{R}$ in at least $k+1$ letters which contradicts the fact that $\mathbf{P} \in \mathbf{S}_4^{3,k}$. So we have that $\mathbf{P} = WW$ for some $W \in \mathbf{C}^*$.

Now, $f(\mathbf{P}) = f(W)f(W)$ is a subword of L'. But $f(W)f(W) \in \mathbf{S}'$ which is impossible by the properties of L'. So Case 1 is impossible.

Note. As an important consequence of Case 1 we have the following. If $\underbrace{x\dots x}_{k+1}\underbrace{y\dots y}_{k+1}$ is a subword of L then $x \ne y$.

Case 2. If there are no letters between $\underbrace{x\dots x}_{k+1}$ and $\underbrace{y\dots y}_{k+1}$, that is

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1}\underbrace{y\dots y}_{k+1}}\overbrace{\underbrace{z\dots z}_{k+1}\underbrace{t\dots t}_{k-i+1}} ,$$

then we must have $x = z$, because otherwise we have $x \ne z$ and $y \ne z$ which gives us that $\mathbf{L}$ and $\mathbf{R}$ differ in the first $k+1$ positions, but this contradicts $\mathbf{P} \in \mathbf{S}_4^{3,k}$.

By the same argument we have $y = t$, so

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1}\underbrace{y\dots y}_{k+1}}\overbrace{\underbrace{x\dots x}_{k+1}\underbrace{y\dots y}_{k-i+1}} .$$

But if we consider now $f(L) = L'$ then it has

$$\mathbf{P}' = \overbrace{f(\underbrace{x\dots x}_{k+1})f(\underbrace{y\dots y}_{k+1})}\overbrace{f(\underbrace{x\dots x}_{k+1})f(\underbrace{y\dots y}_{k+1})} .$$

as a subword, which is impossible since $\mathbf{P}' \in \mathbf{S}'$.

So there is some non-empty subword in $\mathbf{L}$ between $\underbrace{x \ldots x}_{k+1}$ and $\underbrace{y \ldots y}_{k+1}$, and $\mathbf{P}$ can be written as

$$\mathbf{P} = \overbrace{\underbrace{x \ldots x}_{k-i+1} \underbrace{x_1 \ldots x_1}_{k+1} \ldots \underbrace{x_p \ldots x_p}_{k+1} \underbrace{y \ldots y}_{k+1}} \overbrace{\underbrace{z \ldots z}_{k+1} \underbrace{z_1 \ldots z_1}_{k+1} \ldots \underbrace{z_p \ldots z_p}_{k+1} \underbrace{t \ldots t}_{k-i+1}} \; .$$

There are two possible subcases here.

1. $x = z$.

Since $x \neq x_1$ we have $x_1 \neq z$. If $x_1 \neq z_1$ then $\mathbf{L}$ and $\mathbf{R}$ differ in $k+1$ position starting from the $(k-i+2)$th position, which is impossible since $\mathbf{P} \in \mathbf{S}_4^{3,k}$. So $x_1 = z_1$.

In the same way, for each of x_2, x_3, $\ldots x_p$, y, we can obtain that

$$\mathbf{P} = \overbrace{\underbrace{z \ldots z}_{k-i+1} \underbrace{z_1 \ldots z_1}_{k+1} \ldots \underbrace{z_p \ldots z_p}_{k+1} \underbrace{t \ldots t}_{k+1}} \overbrace{\underbrace{z \ldots z}_{k+1} \underbrace{z_1 \ldots z_1}_{k+1} \ldots \underbrace{z_p \ldots z_p}_{k+1} \underbrace{t \ldots t}_{k-i+1}}$$

which leads us to the fact that $\mathbf{L}$ has a subword WW for some $W \in \mathbf{C}^*$, hence $\mathbf{L}'$ has a subword $f(W)f(W)$ which is impossible.

So the subcase 1 is impossible.

2. $x \neq z$.

If $x_1 \neq z$ then $\mathbf{L}$ and $\mathbf{R}$ differ in $k+1$ position starting from the first position, which is impossible since $\mathbf{P} \in \mathbf{S}_4^{3,k}$. So $x_1 = z$.

If $x_2 \neq z_1$ then $\mathbf{L}$ and $\mathbf{R}$ differ in $k+1$ position starting from the $(k+2)$th position, what is impossible by the same arguments as above. So $x_2 = z_1$. And so on.

We have

$$\mathbf{P} = \overbrace{\underbrace{x \ldots x}_{k-i+1} \underbrace{z \ldots z}_{k+1} \underbrace{z_1 \ldots z_1}_{k+1} \ldots \underbrace{z_p \ldots z_p}_{k+1}} \overbrace{\underbrace{z \ldots z}_{k+1} \underbrace{z_1 \ldots z_1}_{k+1} \ldots \underbrace{z_p \ldots z_p}_{k+1} \underbrace{t \ldots t}_{k-i+1}} \; .$$

Applying f to L gives us a subword $\mathbf{P}'$ of L',

$$\mathbf{P}' = \overbrace{f(\underbrace{z \ldots z}_{k+1}) f(\underbrace{z_1 \ldots z_1}_{k+1}) \ldots f(\underbrace{z_p \ldots z_p}_{k+1})} \overbrace{f(\underbrace{z \ldots z}_{k+1}) f(\underbrace{z_1 \ldots z_1}_{k+1}) \ldots f(\underbrace{z_p \ldots z_p}_{k+1})},$$

which is prohibited in L' by $\mathbf{S}'$.

We have got that subcase 2 is impossible and hence Case 2 is impossible.

Case 3. We can assume that $\ell \neq 0$ and $\ell \neq k+1$, because otherwise we deal with either Case 1 or Case 2 which are impossible.

We suppose that $i \geq \ell$ (the case $i < \ell$ can be considered in the same way).

If there are no letters between $\underbrace{y\dots y}_{k-\ell+1}$ and $\underbrace{t\dots t}_{k-j+1}$, then we have either

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1}\underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1}\underbrace{t\dots t}_{k-j+1}}$$

or

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1}\underbrace{z\dots z}_{k+1}\underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1}\underbrace{t\dots t}_{k-j+1}}\ .$$

In the first of these cases we have that $x \neq y$ and $y \neq t$ which gives us that $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in at least $k+1$ letters, but this contradicts $\mathbf{P} \in \mathbf{S}_4^{3,k}$.

In the second case we must have $z = t$, because otherwise since $z \neq y$ and $t \neq y$, $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the last $k+1$ letters which is impossible. So in the second case we have

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1}\underbrace{t\dots t}_{k+1}\underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1}\underbrace{t\dots t}_{k-j+1}}\ .$$

If $x \neq y$ then $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the first $k-\ell+1$ positions and in the last ℓ positions, that is they have non-coincidence in at least $k+1$ positions which is impossible. So $x = y$.

Now applying f to L gives us that L' has a subword

$$\mathbf{P}' = \overbrace{f(\underbrace{x\dots x}_{k+1})f(\underbrace{t\dots t}_{k+1})}\overbrace{f(\underbrace{x\dots x}_{k+1})f(\underbrace{t\dots t}_{k+1})}$$

which is impossible.

So there is some non-empty subword in $\mathbf{R}$ between $\underbrace{y\dots y}_{k-\ell+1}$ and $\underbrace{t\dots t}_{k-j+1}$, and $\mathbf{P}$ can be written in the form

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1}L_1\dots L_p\underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1}R_1\dots R_{p'}\underbrace{t\dots t}_{k-j+1}}\ ,$$

where $L_s,\ R_m \in \mathbf{C}$, for $1 \le s \le p$, $1 \le m \le p'$, and either $p = p'$ or $p = p'+1$.

We define $\triangle(L_s) = x_s$ if $L_s = \underbrace{x_s \dots x_s}_{k+1}$. In the same way we define $\triangle(R_m)$.

Let us consider two cases:

1. $p = p'$. There are two subcases here:

a) $x = y$; Since $\triangle(L_1) \neq x$ we must have $L_1 = R_1$, because otherwise $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the $k+1$ letters starting from the $(k-i+2)$th position, which is impossible.

Since $\triangle(L_2) \neq \triangle(L_1)$, that is $\triangle(L_2) \neq \triangle(R_1)$, we must have $L_2 = R_2$, because otherwise $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the $k+1$ letters starting from the $(2k-i+3)$th position, which is impossible. And so on. For each of of $L_3, \dots, L_p$ we have that

$$\mathbf{P} = \overbrace{\underbrace{y \dots y}_{k-i+1} R_1 \dots R_p \underbrace{y \dots y}_{\ell}}\overbrace{\underbrace{y \dots y}_{k-\ell+1} R_1 \dots R_p \underbrace{t \dots t}_{k-j+1}}\ .$$

So we have got that L has WW as a subword, where $W = \underbrace{y \dots y}_{k+1} R_1 \dots R_p$, but it means that L' has $f(W)f(W)$ as a subword which is impossible.

b) $x \neq y$; There are two special subcases here, namely either $\triangle(L_1) = y$ or $L_1 = R_1$.

When $\triangle(L_1) = y$ it must be that $L_2 = R_1$, because otherwise, since $\triangle(R_1) \neq y$, $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the $k+1$ letters starting from the $(k-l+2)$th position, which is impossible.

By similar reasoning for $L_3, \dots, L_p$ we have that

$$\mathbf{P} = \overbrace{\underbrace{x \dots x}_{k-i+1}\underbrace{y \dots y}_{k+1} R_1 \dots R_{p-1} \underbrace{y \dots y}_{\ell}}\overbrace{\underbrace{y \dots y}_{k-\ell+1} R_1 \dots R_p \underbrace{t \dots t}_{k-j+1}}\ .$$

Again, L has WW as a subword, where $W = \underbrace{y \dots y}_{k+1} R_1 \dots R_{p-1}$, which is impossible by the same reasons as above.

So $L_1 = R_1$.

Using the same technique of as above we can easily obtain that in this case

$$\mathbf{P} = \overbrace{\underbrace{x \dots x}_{k-i+1} R_1 \dots R_p \underbrace{y \dots y}_{\ell}}\overbrace{\underbrace{y \dots y}_{k-\ell+1} R_1 \dots R_p \underbrace{t \dots t}_{k-j+1}}\ .$$

If $y \neq t$ then $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the first $k-\ell+1$ positions and in the last ℓ positions, so they have non-coincidence in $k+1$ positions which contradicts $\mathbf{P} \in \mathbf{S}_4^{3,k}$.

If $y = t$ then

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1} R_1 \dots R_p \underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1} R_1 \dots R_p \underbrace{y\dots y}_{k-j+1}}$$

and L has WW as a subword, where $W = R_1 \dots R_p \underbrace{y\dots y}_{k+1}$ which is impossible.

2. $p = p' + 1$. There are two subcases here:

a) $x = y$; It must be that $L_1 = R_1$, because otherwise $\mathbf{L}$ and $\mathbf{R}$ differ in $k+1$ positions starting from the $(k-i+2)$th position. Then we consider L_2, L_3, $\dots$,L_p.

We can see that in this subcase

$$\mathbf{P} = \overbrace{\underbrace{y\dots y}_{k-i+1} R_1 \dots R_{p'} \underbrace{t\dots t}_{k+1} \underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1} R_1 \dots R_{p'} \underbrace{t\dots t}_{k-j+1}} ,$$

and L has WW as a subword, where $W = \underbrace{y\dots y}_{k+1} R_1 \dots R_{p'} \underbrace{t\dots t}_{k+1}$ which is impossible.

b) $x \neq y$; There are two special subcases here, namely either $\triangle(L_1) = y$ or $L_1 = R_1$.

If $\triangle(L_1) = y$ then

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1} \underbrace{y\dots y}_{k+1} R_1 \dots R_{p'} \underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1} R_1 \dots R_{p'} \underbrace{t\dots t}_{k-j+1}} ,$$

and L has WW as a subword, where $W = \underbrace{y\dots y}_{k+1} R_1 \dots R_{p'}$ which is impossible.

So $L_1 = R_1$. In this case we have

$$\mathbf{P} = \overbrace{\underbrace{x\dots x}_{k-i+1} R_1 \dots R_{p'} \underbrace{t\dots t}_{k+1} \underbrace{y\dots y}_{\ell}}\overbrace{\underbrace{y\dots y}_{k-\ell+1} R_1 \dots R_{p'} \underbrace{t\dots t}_{k-j+1}} .$$

Since $y \neq x$, $y \neq \triangle(R_1)$ and $y \neq t$, $\mathbf{L}$ and $\mathbf{R}$ have non-coincidence in the first $k-l+1$ positions and in the last l positions, so they have non-coincidence in $k+1$ positions which contradicts $\mathbf{P} \in \mathbf{S}_4^{3,k}$.

We have got that Case 3 is impossible.

We have proved that the infinite word L contains no word from the set $\mathbf{S}_4^{3,k}$ as a subword, therefore $\mathbf{S}_4^{n,k}$ is incomplete for $n \geq 3$.

References

1 Evdokimov A. Complete Sets of Words and their Numeral Characteristics, Methods of discrete analysis in research of extremal structures, Novosibirsk, IM SB RAS, Issue **39**, (1983), 7-32.

2 Salomaa A. Jewels of Formal Language, Theory Computer Science Press, (1981).

3 Choffrutc, Karhumäk Handbook of formal languages. Vol. **1**: Word, language, grammar, Berlin, Springer, (1997), 329-438.

4 Lothaire M. Combinatorics on Words, Encyclopedia of Mathematics, Vol. **17**, Addison-Wesley (1986).

Chapter 2

On Non-Existence of an Iterative Morphism that Defines the Arshon Sequence

Abstract

In [2], Berstel proved that the Arshon sequence cannot be obtained by iteration of a morphism. An alternative proof of this fact is given here.

2.1 Introduction and Background

In 1937, Arshon gave a construction of a symbolic sequence w, which in the alphabet $\{1, 2, 3\}$ is built as follows: Let $w_1 = 1$. For $k \geq 1$, w_{k+1} is obtained by replacing the letters of w_k in odd positions thus:

$$1 \to 123,\ 2 \to 231,\ 3 \to 312$$

and in even positions thus:

$$1 \to 321,\ 2 \to 132,\ 3 \to 213.$$

Then

$$w_2 = 123, \quad w_3 = 123132312,$$

and each w_i is the initial subword of w_{i+1}, so the infinite symbolic sequence $w = \lim_{n\to\infty} w_n$ is well defined. It is called the *Arshon sequence*.

This method of constructing w is called the *Arshon Method* (**AM**), and ψ will denote the indicated map of the letters 1, 2, 3, according to position as described above.

We will denote the natural decomposition of w in 3-blocks by lower braces:

$$w = \underbrace{123}\underbrace{132}\underbrace{312}\ldots.$$

The paper by Arshon [1] was published in connection with the problem of building a nonrepetitive sequence in a 3-letter alphabet, that is, a sequence that does not contain any subwords of the type $XX = X^2$, where X is any word of a 3-letter alphabet. The sequence w has that property. The question of the existence of such a sequence was studied in algebra, discrete analysis and in dynamical systems.

Let Σ be an alphabet and Σ^* be the set of all words of Σ. A map $\varphi : \Sigma^* \to \Sigma^*$ is called a *morphism*, if we have $\varphi(uv) = \varphi(u)\varphi(v)$ for any $u, v \in \Sigma^*$. It easy to see that a morphism φ can be defined by defining $\varphi(i)$ for each $i \in \Sigma$.

Suppose a word $\varphi(a)$ begins with a for some $a \in \Sigma$, and that the length of $\varphi^k(a)$ increases without bounds. The symbolic sequence $\lim\limits_{k\to\infty} \varphi^k(a)$ is called a *fixed point* of the morphism φ.

We now study classes of sequences, that are defined by iterative schemes. There are many techniques to study sequences generated by morphisms. So it is reasonable to try to determine if a sequence under consideration can be obtained by iteration of a morphism.

Let us consider the following classical sequence as an example of such a sequence. This sequence has two different definitions (actually it has at least 4 different definitions).

The *Thue-Morse sequence* $\mathbf{t}$:

$$t_0, t_1, t_2, t_3, \ldots$$

is defined by the conditions:

$$\begin{aligned} &t_0 = 0, \\ &t_{2n+1} \equiv t_n + 1 \pmod 2, \\ &t_{2n} = t_n. \end{aligned}$$

So the initial letters of $\mathbf{t}$ are

$$0110100110010110\ldots.$$

This sequence is defined by the morphism σ:

$$\sigma(0) = 01 \ ,$$

$$\sigma(1) = 10 \ .$$

Since the construction of the Arshon sequence w is similar to the iteration morphism scheme, and because w is constructed by two morphisms f_1 and f_2, applied depending on whether the letter position is even or odd, we might expect that there exists a morphism f which generates w.

But this turns out not to be true, due to the following theorem.

2.2 The Theorem

Theorem. There does not exist a morphism, whose fixed point is the Arshon sequence.

Note. A corollary of this theorem is the non-existence of a morphism which defines the Arshon sequence. In fact, if such a morphism exists, it must have the property that is 1 mapped to $1X$ by the action of the morphism, where X is some word, and from this it follows that the Arshon sequence is a fixed point of this morphism.

Proof (*of the theorem*)**:**

It is enough to prove the non-existence of a morphism f with the property $w = f(w)$, since from the definition of a fixed point we have that if w is a fixed point of the morphism f then $w = f(w)$. Suppose there exists a morphism f such that

$f(1) = X,\ f(2) = Y,\ f(3) = Z$ and $w = f(w)$.

From all such morphisms we choose a morphism with minimal length of X.

The morphism f is not an erasing morphism, that is $|X| \geq 1$, $|Y| \geq 1$, $|Z| \geq 1$, since otherwise $w = f(w)$ contains a subword of the type PP (where P is some word) which cannot belong to w. Now $|X| + |Y| + |Z| \neq 3$, since otherwise $|f^l(1)| = 1$ for $l = 1, 2...$, and w is not a fixed point of the morphism f.

$$f(w) = w = XYZXZYZXY...$$

Hence X consists of $|X|$ of the first letters of w, Y is $|Y|$ of the following letters, and Z is $|Z|$ of the letters following that.

We will use upper braces to show the decomposition of w into f-blocks (that is, to show the disposition of the words X, Y and Z in w). We have

$$w = \overbrace{\underbrace{123}\underbrace{132}\dots a_{|X|}}^{X}\overbrace{a_{|X|+1}\dots a_{|X|+|Y|}}^{Y}\overbrace{a_{|X|+|Y|+1}\dots a_{|X|+|Y|+|Z|}}^{Z}\overbrace{a_{|X|+|Y|+|Z|+1}\dots\dots}^{X}\dots,$$

where all a_i are letters of the alphabet $\{1,2,3\}$.

Lemma 1. We have $|X|+|Y|+|Z| \equiv 0 \pmod 3$.

Proof: From the structure of w, the frequencies of 1, 2, 3 in w coincide, hence the frequencies of these letters in $f(w) = w$ coincide as well. But this is only possible when $|X|+|Y|+|Z| \equiv 0 \pmod 3$, since otherwise there are two letters, whose frequencies in $f(w) = w$ do not coincide.

Lemma 2. The situation $|X| \equiv |Y| \equiv |Z| \equiv 0 \pmod 3$ is impossible.

Proof: Suppose $|X| \equiv |Y| \equiv |Z| \equiv 0 \pmod 3$. Then X, Y and Z consist of a whole number of 3-blocks. Hence we can consider the words $X' = \psi^{-1}(X)$, $Y' = \psi^{-1}(Y)$, $Z' = \psi^{-1}(Z)$. The properties of ψ give

$$w = \psi^{-1}(w) = X'Y'Z'X'Z'Y'Z'X'Y'...$$

so there exists a morphism f' which maps 1 to X', 2 to Y', 3 to Z' and $w = f'(w)$. Since $|X'| = |X|/3$, we have $|X'| < |X|$. This contradicts the choice of the morphism f.

Lemma 3. With the assumption of the existence of the morphism f, $|X| \leq 5$.

Proof: Suppose $|X| \geq 6$, that is, $X = 123132\dots$. If $|X| \equiv 2 \pmod 3$ ($|X| \equiv 1 \pmod 3$), then $|X| \geq 7$ and using Lemma 1 we consider the 4th f-block $X = 12\underbrace{313}\dots$ ($X = 1\underbrace{231}\underbrace{323}\dots$). This contradicts the **AM**. Hence $|X| \equiv 0 \pmod 3$.

It follows from Lemma 2 that the situation $|Y| \equiv 0 \pmod 3$ is impossible. If $|Y| \equiv 1 \pmod 3$ ($|Y| \equiv 2 \pmod 3$), then we consider the 10th (3rd) f-block $X = 12\underbrace{313}2\dots$ and it brings us to a contradiction with the **AM**. Hence if $|X| \geq 6$ then the morphism f can not exist.

Lemma 4. With the assumption of the existence of the morphism f, $|X| \neq 1$.

Proof: If $|X| = 1$, then $X = 1$ and the length of the words $f^k(1)$ for $k = 1, 2, \dots$ does not increase, whence w is not a fixed point of the morphism f. This is a contradiction.

Lemma 5. With the assumption of the existence of the morphism f, $|X| \neq 2$.

Proof: Suppose $|X| = 2$, that is $X = 12$.

We have $|X| \equiv 2 \pmod 3$, hence, using Lemma 1, we have $|Y| + |Z| \equiv 1 \pmod 3$.

We consider the 2nd f-block X and the f-block Z next after it. It can be seen that Z begins with 3. We consider the 4th f-block X and Y preceding it and find that Y ends with 3. But then, considering YZ, which is a subword of w, we see, that 33 is a subword of w, which is impossible. That is for $|X| = 2$ the morphism f cannot exist.

The 3-blocks 123, 231, 312 are said to be *odd* 3-blocks. All other 3-blocks are said to be *even*.

Lemma 6. With the assumption of the existence of the morphism f, $|X| \neq 3$.

Proof: Suppose $|X| = 3$, that is $X = 123$.

We have $|X| \equiv 0 \pmod 3$, hence, using Lemma 1 we have $|Y| + |Z| \equiv 0 \pmod 3$. Considering the **AM**, the 2nd f-block X must be an odd 3-block, hence $|Y| + |Z| \equiv 1 \pmod 2$.

Let $|Z| \geq 2$. Then the 2nd f-block Z begins with an even 3-block, and the 3rd Z begins with an odd 3-block. This is impossible since 2 letters define the evenness of the 3-block unambiguously. Thus $|Z| = 1$.

Let $|Y| \geq 2$. In XYZ (or in an arbitrary permutation of these letters) there is an even number of 3-blocks, so the 9th f-block Y begins with an odd 3-block, but the 1st Y begins with an even 3-block. Hence $|Y| = 1$.

This is a contradiction with $|Y| + |Z| \equiv 0 \pmod 3$ (and also a contradiction with $|Y| + |Z| \equiv 1 \pmod 2$). That is for $|X| = 3$ the morphism f cannot exist.

Lemma 7. With the assumption of the existence of the morphism f, $|X| \neq 4$.

Proof: Suppose $|X| = 4$, that is $X = 1231$.

We have $|X| \equiv 1 \pmod 3$, hence, using Lemma 1, we have $|Y| + |Z| \equiv 2 \pmod 3$.

We have $|Y| \geq 2$, since otherwise $Y = 3$ and hence XYX which is a subword of w, contains 3131, which is impossible. Hence $Y = 32\ldots$. We consider ZX and ZY and see that Z ends with 2. Now $|Z| \geq 2$, since otherwise $Z = 2$ and XZX which is a subword of w, contains 1212, which is impossible. Hence $Z = \ldots 32$, or $Z = \ldots 12$. The former is impossible since 3232 is contained in ZY, and hence in w. The latter is impossible too, since considering the 9th f-block Z and the f-block X following it, we obtain $ZX = \ldots\underbrace{121}\underbrace{231}$, which contradicts the **AM**. That is for $|X| = 4$ the morphism f cannot exist.

Lemma 8. With the assumption of the existence of the morphism f, $|X| \neq 5$.

Proof: Suppose $|X| = 5$, that is $X = 12313$.

We have $|X| \equiv 2 \pmod 3$, hence, using Lemma 1, we have $|Y| + |Z| \equiv 1 \pmod 3$. Then the 4th f-block is $X = 12\underbrace{313}$, which is a contradiction with the **AM**. That is if $|X| = 5$ then the morphism f cannot exist.

From Lemmas 3 - 8 we have a contradiction with the assumption of the existence of the morphism f. This proves the Theorem.

References

1 Arshon S. E. The Proof of the Existence of n - letter Non-Repeated Asymmetric Sequences, Math. collected papers, New series, Publ. H. AS USSR, Vol. **2**, Issue **3**, Moscow (1937), 769-779.

2 Berstel J. Mots sans carré et morphismes itérés, Discrete Math. **29** (1979), 235-244.

3 Cobham A. Uniform tag sequences, Math. Systems Theory **6** (1972), 164-192.

4 Lothaire M. Combinatorics on Words, Encyclopedia of Mathematics, Vol. **17**, Addison-Wesley (1986).

5 Salomaa A. Jewels of Formal Language, Theory Computer Science Press, (1981).

Chapter 3

There is no Iterative Morphism which Defines the Sequence of σ

Abstract

The sequence of σ was constructed by Evdokimov in order to build chains of maximal length in the n-dimensional unit cube. We prove that the sequence can not be defined by iteration of a morphism.

3.1 Introduction and Background

Any natural number n can be presented unambiguously as $n = 2^t(4s + \sigma)$, where $\sigma < 4$, and t is the greatest natural number such that 2^t divides n. If n runs through the natural numbers then σ runs through the sequence that we will call the *sequence of* σ. We let w denote that sequence. Obviously, w consists of 1s and 3s. The initial letters of w are $11311331113313\ldots$.

Defining a morphism we will follow [4].

Let Σ be an alphabet and Σ^* be the set of all words of Σ. A map $\varphi : \Sigma^* \to \Sigma^*$ is called a *morphism*, if we have $\varphi(uv) = \varphi(u)\varphi(v)$ for any $u, v \in \Sigma^*$. It easy to see that a morphism φ can be defined by defining $\varphi(i)$ for each $i \in \Sigma$.

A sequence α is *defined by the iteration of a morphism* φ, if $\alpha = \lim_{k\to\infty} \varphi^k(a)$, where a is the first letter of α.

In [1], Evdokimov built chains of maximal length in the n-dimensional unit cube using the sequence of σ. Naturally a question arises as to the possibility of constructing w using the iteration of a morphism, since the possibility of such a constructing could help us with studying of w.

We now give an alternative definition of the sequence w, by the following inductive scheme:

$$C_1 = 1, \qquad D_1 = 3$$

$$C_{k+1} = C_k 1 D_k, \qquad D_{k+1} = C_k 3 D_k$$
$$k = 1, 2, \ldots$$

and $w = \lim_{k\to\infty} C_k$.

3.2 The Theorem

Theorem. There does not exist a morphism whose iteration defines the sequence w.

Proof (*of the theorem*)**:** Suppose there exists a morphism f, such that $f(1) = X$, $f(3) = Y$ and $w = \lim_{k\to\infty} f^k(1)$. Obviously, X consists of the first $|X|$ letters of w, where $|X|$ is the length of X.

Lemma 1. The subsequence of w consisting of the letters in odd positions is the alternating sequence of 1s and 3s: $1313131\ldots$.

Proof: The odd positions of w correspond to the odd numbers $n = 2^0(4s + \sigma) = 4s + \sigma$, so cleary σ alternates between 1 and 3.

Lemma 2. If there exists a morphism f whose iteration gives w then $|X| \equiv 0 \pmod 4$.

Proof: It is easy to see that $f(1) = 1X^{(1)}$, where $|X^{(1)}| \geq 1$, since otherwise $|f^k(1)| = 1$, for $k = 1, 2, 3\ldots$, so w cannot be obtained by iterating f.

Suppose $|X^{(1)}| = 1$, that is $f(1) = 11$. But then w consists of 1s only, which is impossible, hence $f(1) = 11X^{(2)}$, where $|X^{(2)}| \geq 1$.

Suppose $|X^{(2)}| = 1$, that is $f(1) = 113$. Since w has the subword 111, then w has a subword $f(111) = 113113113$. If $f(111)$ begins with a letter in an odd position, then the marked letters **1**1**3**1**1**3**1**1**3**, read from left to right will make up consecutive letters of w in odd positions. This contradicts Lemma 1. If $f(111)$ begins with a letter in an even position, then marking letters in odd positions will lead to the same contradiction with Lemma 1, hence $f(1) = 113X^{(3)}$, where $|X^{(3)}| \geq 1$.

Suppose $|X^{(3)}| = 1$, that is $f(1) = 1131$. Then $f^2(1) = 11311\mathbf{1}31Y1131$ and the marked letter does not coincide with the letter of w standing in the same place, hence $f(1) = 1131X^{(4)}$, where $|X^{(4)}| \geq 1$.

If $|X|$ is odd, then the marked letters in $f^2(1) = 1131X^{(4)}1\mathbf{131}X^{(4)}\ldots$ are two consecutive letters in odd places. This contradicts Lemma 1. Hence $|X|$ is even.

We have $f^2(1) = XX\ldots = X1131X^{(4)}\ldots$, whence the next-to-last letter of X is in an odd position and is equal to 3, since otherwise two consequent 1 in w stand at odd places, which contradicts Lemma 1. The natural number which corresponds to the next-to-last letter of X is written as $2^0(4s+3)$, the next number is equal to $|X|$ and to $2^0(4s+3)+1 = 4(s+1) \equiv 0 \pmod 4$.

The following Lemma is straightforword to prove.

Lemma 3. If $n_1 = 2^{t_1}(4s_1+1)$, $n_2 = 2^{t_2}(4s_2+1)$, $n_3 = 2^{t_3}(4s_3+3)$ and $n_4 = 2^{t_4}(4s_4+3)$ then n_1n_2, n_3n_4 can be written as $2^t(4s+1)$, and n_1n_3 as $2^t(4s+3)$.

It follows from Lemma 2 that $|X| = 4t$.

Suppose X ends with 1 (the case when X ends with 3 is similar), that is at the $(4t)$th position in X we have 1. According to the multiplication by 2 does not change σ, so at the $(2t)$th position in X we have 1.

Consider $f^2(1) = X\mathbf{X}\ldots$. The letters of the marked X occupy the positions of $f^2(1)$ from $(4t+1)$th to $(8t)$th. Since $X = \mathbf{X}$, then at the $(6t)$th place we have 1. But $6t = 3(2t)$, whence, by Lemma 3, at the $(2t)$th and the $(6t)$th places there must stand different letters. This is a contradiction and the Theorem is proved.

References

1 Berstel J. Mots sans carré et morphismes itérés, Discrete Math. **29** (1979), 235-244.

2 Evdokimov A. On the Maximal Chain Length of an Unit n-dimensional Cube, Maths Notes **6**, No. **3** (1969), 309-319.

3 Lothaire M. Combinatorics on Words, Encyclopedia of Mathematics, Vol. **17**, Addison-Wesley (1986).

4 Salomaa A. Jewels of Formal Language, Theory Computer Science Press, (1981).

5 Discrete Mathematics and Mathematical Questions of Cybernetics, Vol. **1**, Moscow (1974), 112-116.