Completeness and Complexity of Some Extremal Problems for Sets of Prohibited Words.

Sergey Kitaev

Let $\mathbf{A} = \{a_1, \ldots, a_n\}$ be an alphabet of n letters. Any i consecutive letters of a word X generate a *subword* of length i.

$\mathbf{A}^*$ is the set of all the words in the alphabet $\mathbf{A}$.

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

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

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

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

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

Then $\mathbf{S}$ is incomplete, since the word $\underbrace{124124\ldots124}_{3k}$ is in $\widehat{\mathbf{S}}$ for any k.

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

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

It is easy to check that $\mathbf{S}$ is complete.

The Basic Problem:

Given: A set of prohibitions $\mathbf{S}$.

The question: Is $\mathbf{S}$ complete?

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

$\mathbf{2121}2 \notin \widehat{\mathbf{S}}_1^2$;

$124123 \in \widehat{\mathbf{S}}_1^4$.

$\mathbf{S}_1^n$ is incomplete for $n \geq 3$.

The first example of an infinite sequence that is free from $\mathbf{S}_1^3$ was given by Thue in 1906. He suggested the following iterated morphism:

$$a \to abc, \;\; b \to ac, \;\; c \to b;$$

a, abc, $abcacb$, $abcacbabcbac$, ...

Let $\nu_i(X)$ be the number of occurrences of the letter a_i in X.

The $\overline{\nu}(X) = (\nu_1(X), \ldots, \nu_n(X))$ is the *content* vector of X.

$\overline{\nu}(1311231) = (4, 1, 2)$;

Let $\mathbf{S}_2^n = \{XY \mid \overline{\nu}(X) = \overline{\nu}(Y)\}$. That is, we prohibit the repetition of two consecutive subwords of the same content.

In each word from $\widehat{\mathbf{S}}_2^n$, no two adjacent subwords are permutations of each other.

$1\mathbf{312321}23 \notin \widehat{\mathbf{S}}_2^3$;

$1232413 \in \widehat{\mathbf{S}}_2^4$.

In 1963, Erdös asked whether $\mathbf{S}_2^4$ is incomplete.

The following sets of prohibitions were shown to be incomplete:

$\mathbf{S}_2^{25}$ – Evdokimov (1968)

$\mathbf{S}_2^7$

$\mathbf{S}_2^5$ – Pleasants (1970)

Erdös's question was answered in the affirmative in 1992 by Keränen.

$|f(a)| = |f(b)| = |f(c)| = |f(d)| = 85$

$f(a) = abcacdcbcdcadcdbdabacabadbabcbdbcbacbcdcacbabdabacadcb$ -
$cdcacdbcbacbcdcacdcbdcdadbdcbca$

$f(b) = bcdbdadcdadbadacabcbdbcbacbcdcacdcbdcdadbdcbcabcbdbadc$ -
$dadbdacdcbdcdadbdadcadabacadcdb$

$f(c) = cdacabadabacbabdbcdcacdcbdcdadbdadcadabacadcdbcdcacbad$ -
$abacabdadcadabacabadbabcbdbadac$

$f(d) = dabdbcbabcbdcbcacdadbdadcadabacabadbabcbdbadacdadbdcba$ -
$bcbdbcabadbabcbdbcbacbcdcacbabd$

Now apply f to a and iterate.

Let $d(X,Y)$ be the number of letters in which the words X and Y differ (Hamming metric).

$d(10212, 21213) = 3$;

$\mathbf{S}_3^{n,k} = \{XY \mid d(X,Y) \leq k, |X| = |Y| \geq k+1, k \in \mathbf{N}\}$, where $|X|$ is the length of the word X. That is, we prohibit any two consecutive subwords X, Y of length greater then k such that $d(X,Y) \leq k$.

$\mathbf{3132212}23 \notin \widehat{\mathbf{S}}_3^{3,2}$;

$131121231 \in \widehat{\mathbf{S}}_3^{3,3}$.

Theorem*[Kitaev]* $\mathbf{S}_3^{n,k}$ is incomplete for $n \geq 3$.

Other Questions about $\widehat{\mathbf{S}}$:

If $\mathbf{S}$ is complete then:

- Find or estimate $L(\widehat{\mathbf{S}}) = \max\limits_{X \in \widehat{\mathbf{S}}} |X|$;
- Construct a word of length $L(\widehat{\mathbf{S}})$ that is free from $\mathbf{S}$;
- $|\widehat{\mathbf{S}}| = ?$

If $\mathbf{S}$ is incomplete then:

- Find a sequence that is free from $\mathbf{S}$;
- Describe all such sequences;
- Find the cardinality of the set of these sequences.

$\mathbf{A} = \{0, 1\}$;

$\mathbf{S_1} = \{000, 001, 101\underline{1}, 0101, 1111\}$;

$\mathbf{S_2} = \{000, 001, 101\underline{0}, 0101, 1111\}$.

$\mathbf{S_1}$ is complete; $L(\widehat{\mathbf{S}_1}) = 8$; 01110100.

$\mathbf{S_2}$ is incomplete since

$\underbrace{011}\underbrace{011}\ldots$ is free from $\mathbf{S_2}$;

$\underbrace{0111}\underbrace{0111}\ldots$ is free from $\mathbf{S_2}$;

$\underbrace{011}\underbrace{0111}$ and $\underbrace{0111}\underbrace{011}$ are free from $\mathbf{S_2}$;

So taking *any* sequence in the alphabet A and substituting all 0 with 011 and all 1 with 0111 we get a sequence that is free from $\mathbf{S_2}$ whence the cardinality of $\widehat{\mathbf{S}_2}$ is the $\mathbf{continuum}$!

In general it is important to know criteria of completeness for different classes of sets of prohibited words. An example of such a class can be the following:

$$\{\mathbf{S} : \mathbf{S} \subseteq \mathbf{A}^n\};$$

that is, we are interested in words of length n. It is interesting to know numerical characteristics that describe the bound between complete and incomplete sets of prohibitions. For instance, let

$$M(n) = \min |\mathbf{S}| \quad \text{and} \quad L(n) = \max L(\widehat{\mathbf{S}}),$$

where the extremum is taken with respect to all complete $\mathbf{S} \subseteq \mathbf{A}^n$ and $L(\widehat{\mathbf{S}}) = \max\limits_{X \in \widehat{\mathbf{S}}} |X|$.

De Brujin Graph

We define a directed graph $\vec{G}_n(V, E)$ by the following.

$V = \mathbf{A}^n$; There is an arch from $v_i = (v_{i_1}, \ldots, v_{i_n})$ to $v_j = (v_{j_1}, \ldots, v_{j_n})$ if

$$v_{i_2} = v_{j_1},\, v_{i_3} = v_{j_2}, \ldots, v_{i_n} = v_{j_{n-1}},$$

that is, when the words v_i and v_j overlap in $(n-1)$ letters.

A *chord* of a directed simple path $\vec{P}$ in $\vec{G}_n$ is an arch that does not belong to $\vec{P}$ but connects two of its vertices in a such way that there is a directed circuit generated by this arch and the part of the path between the ends of the arch.

$C(n)$ is the greatest length (the number of vertices) of a simple path in $\vec{G}_n$ that does not have chords and does not go through any vertex that has a loop.

Theorem 1. We have

$$L(n) = C(n) + n - 1 = |\mathbf{A}|^{n-1} + n - 2.$$

Theorem 2. We have

$$M(n) = \frac{1}{n} \sum_{d|n} \varphi(n/d) |\mathbf{A}|^d,$$

where $\varphi(n)$ is the number of integers among $1, 2, \ldots n-1$ that are relatively prime to n (Euler's φ-function).

Since any set of prohibitions $\mathbf{S}$ for $|\mathbf{S}| \leq M(n)$ is incomplete, it is helpful to have a table for $M(n)$. If $|\mathbf{A}| = 2$ and $1 < n \leq 10$ then we have

n	2	3	4	5	6	7	8	9	10
M(n)	3	4	6	8	14	20	36	60	108

In particular, any set of binary words of length 10 that has less than 108 words is incomplete.

$M(n) \sim |\mathbf{A}|^n/n$, when $n \to \infty$.

Proposition. There exist at least $2^{|\mathbf{A}|^n(1-\varepsilon_n)}$ complete sets $\mathbf{S}$ such that $\mathbf{S} \subseteq \mathbf{A}^n$. Here $\varepsilon_n \to 0$ when $n \to \infty$.

The definitions of completeness and the function $L(n)$ immediately lead to the following.

Proposition. A set of prohibitions $\mathbf{S} \subseteq \mathbf{A}^n$ is complete iff any word of length $L(n)+1$ is not free from $\mathbf{S}$.

So in order to decide if a given set of prohibitions $\mathbf{S}$ is complete or not, we can consider all words of length $L(n)+1 = |\mathbf{A}|^{n-1} + n - 1$ and check if there is a word that is free from $\mathbf{S}$.

Theorem. The complexity of deciding whether or not an arbitrary set $\mathbf{S} \subseteq \mathbf{A}^n$ is complete is $O(|\mathbf{S}| \cdot n)$.

Problem A:

Given: An arbitrary set of words $\mathbf{S}$ and a natural number ℓ.

The question: Does there exist a word of length at least ℓ that is free from $\mathbf{S}$?

Problem B that is equivalent to the Basic Problem:

Given: An arbitrary set of words $\mathbf{S}$.

The question: Does there exist $\ell \in \mathbf{N}$ such that $|X| \le \ell$ for any word X that is free from $\mathbf{S}$?

$\mathbf{A} = \{a_1, \ldots, a_n\}$ is an alphabet;

$\mathbf{A}_\ell$ is the set of all the words of the alphabet $\mathbf{A}$ whose length is at most ℓ. We assume that the empty word belongs to $\mathbf{A}_\ell$.

$\mathbf{S}_1 \subseteq \mathbf{A}_2 \setminus \mathbf{A}_1$;
$\mathbf{S}_2 = \{xXx \mid x \in \mathbf{A}, X \in \mathbf{A}_{n-1}\}$.

So $\mathbf{S}_2$ contains all possible words of length at most $n+1$ whose first letter coincides with their last letter.

Suppose $\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2$.

Problem $\mathbf{A'}$:

Given: An arbitrary set $\mathbf{S}$ of the type described above and a natural number $\ell \leq n$.

The question: Does there exist a word of length at least ℓ that is free from $\mathbf{S}$?

Problem "path":

Given: Directed graph $\vec{G}(V, E)$ and a natural number $\ell \leq |V| = n$.

The question: Does there exist a simple directed path of length at least ℓ?

Correspondence between problem $\mathbf{A}'$ and problem "path":

We compare vertices $v_1, \ldots, v_n$ from $V(\vec{G})$ to the letters $a_1, \ldots, a_n$ in the alphabet $\mathbf{A}$.

We compare each arch $v_i v_j$ from $E(\vec{G})$ to the word $a_i a_j$.

We form the set $\mathbf{S}_1$ from all such words of $\mathbf{A}_2$ that correspond to the arch of the complement of $\vec{G}$ with respect to the complete directed graph.

$v_{i_1}, \ldots, v_{i_\ell}$ is a simple directed path iff $a_{i_1} \ldots a_{i_\ell}$ is free from $\mathbf{S}$.

Problem “path” is NP-complete $\Rightarrow$

Problem A' is NP-complete $\Rightarrow$

Problem A is NP-complete!