Polyhedra of Finite State Machines and Their Use in the Identification Problem

Sergey Yu. Melnikov[1(✉)] and Konstantin E. Samouylov[1,2]

[1] Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
melnikov@linfotech.ru, ksam@sci.pfu.edu.ru
[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

Abstract. The possible sets of joint distribution of the word occurrence frequencies in the finite state machine input and output sequences are considered. A geometric description of such sets as convex polyhedra in a real unit cube of suitable dimension is proposed. A method has been developed for comparison of unknown and reference automata by the observed input and output sequence fragments. The method does not require installation to the fixed initial state.

Keywords: FSM identification · Statistical properties of automata · Word occurrence statistics

1 Introduction

We will consider the problem of testing the hypothesis that an unknown automaton A (which input and output sequences are observed) coincides with the known automaton A_0. It is necessary to check whether such an initial state of the automaton A_0 exists, starting from which it transforms the observed input sequence into the observed output sequence. We believe that the unknown automaton A is selected from some finite class containing the automaton A_0. We assume that all automata from this class have the same alphabets and are pairwise nonequivalent.

This task is relevant in the theory of technical device testing and diagnostics, as well as in a number of cryptographic applications, in particular, when testing the hypothesis that the analyzed device implements some encryption algorithm with unknown key.

The formulated problem can be solved by installation of the automaton A_0 in each of the possible initial states, and application of the observable input sequence to its input. If for any initial state the resulting output sequence does not coincide with the observed one, then the hypothesis that A and A_0 coincide is rejected. If at least one of the variants shows a coincidence of sequences, it is concluded that the observed data do not contradict the tested hypothesis. The

O. Galinina et al. (Eds.): NEW2AN 2020/ruSMART 2020, LNCS 12526, pp. 110–121, 2020.
https://doi.org/10.1007/978-3-030-65729-1_10

complexity of this method is proportional to the nonequivalent state number of the automaton and is extremely high for automata modeling of the information processing equipment nodes.

The proposed approach uses the construction of a special polyhedron corresponding to the automaton A_0 inside a real unit cube of suitable dimension. In the observed input and output sequences of automaton A, the relative frequencies of certain word occurrences are calculated. These relative frequencies determine the coordinates of the points in the cube. The distance between the polyhedron of the automaton A_0 and these points is calculated. In the case when this distance exceeds a threshold depending on the observed sequence lengths, the hypothesis that A and A_0 coincide is rejected.

2 Related Works

Geometric representations are traditionally used to identify non-obvious statistical dependencies in the output sequence when analyzing pseudorandom sequence generators [1,2].

In [3], an approach is described related to the construction of automata geometric images, in which the automaton behavior is displayed in geometric figures, in particular, in curves on a plane.

When all possible words are fed to the automaton input, some output words do not appear (these words are "prohibitions" of the automaton [4]), but some output words appear repeatedly. The word frequencies in the output sequences in [5] are studied using the so-called histogram automaton function, which associates the word in the output alphabet with its frequency. The geometric constructions associated with the convex hull construction of point sets in n-dimensional space were used in [6] in the study of quantitative languages that assign a real number to each word.

In [7], a method was proposed for detection of covert channels in information systems by checking for the presence of forbidden fragments ("prohibitions") in transmitted sequences. Since the covert channel organizers do not know about this, then if such a fragment is found in the observed sequence, the controller determines that the covert channel is functioning. An analogy can be drawn between such an approach and the one considered in this paper: hypotheses about the absence of a covert channel or about the coinciding of an automaton with a reference one are rejected when a certain inequality holds for the certain event frequency in the observed sequences. The rejection criterion in both cases is deterministic, it has a zero error of the second kind.

The proposed approach can also be useful for checking the quality of pseudorandom sequence generators, which are widely used in modern traffic control technologies, such as Random-Access Channel [8] and device-to-device (D2D) communications [9,10].

3 Definition of a Polyhedron of an Automaton

Let B be a finite set (alphabet). By B^* we denote the set of all words in the alphabet B. We denote by Ω the set of all infinite sequences over B:

$$\Omega = \{\omega = w_1 w_2 \dots | w_t \in B,\, t = 0, 1, \dots\}. \tag{1}$$

For each word $\alpha \in B^*$, $\alpha = a_0 a_1 \dots a_{m-1}$, where $a_i \in B$, $i = 0, 1, \dots, m-1$, $m = 1, 2, \dots$ we define a cylinder

$$[\alpha] = [a_0 a_1 \dots a_{m-1}] = \{\omega = w_0 w_1 \dots | w_0 = a_0, w_1 = a_1, w_{m-1} = a_{m-1}\} \subset \Omega. \tag{2}$$

The characteristic function of an arbitrary subset $F \subset \Omega$ will be denoted by I_F:

$$I_F = \begin{cases} 1, & \text{if} \quad \omega \in F \\ 0, & \text{if} \quad \omega \notin F \end{cases}. \tag{3}$$

Instead of $I_{[\alpha]}$ we will simply write I_α.

Define a mapping T ("sequence shift") $T : \Omega \to \Omega$ by

$$T : \omega = w_0 w_1 \dots \to \omega T = w_1 w_2 \dots. \tag{4}$$

The equality

$$I_\alpha\left(\omega T^t\right) = 1 \tag{5}$$

means in such a way that

$$w_t = a_0,\, w_{t+1} = a_1, \dots,\, w_{t+m-1} = a_{m-1}. \tag{6}$$

The number $\frac{1}{t}\sum_{j=0}^{t-1} I_\alpha\left(\omega T^{s+j}\right)$ is called the relative frequency of occurrence of the word α in the sequence ω on the segment from s to $s + t - 1$. We will use the notation

$$p_\alpha(\omega) = \lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} I_\alpha\left(\omega T^j\right), \tag{7}$$

if the limit on the right side exists. The value of $p_\alpha(\omega)$ can be interpreted as an average frequency of occurrence of the word α in the sequence ω [11].

Such limits exist, for example, for infinite periodic sequences (both purely periodic and periodic with an initial section), the set of which we denote by T_B. In this case, as can be seen from the formula (7), p_α is the ratio of the frequency of occurrence of the word α in the period (the number of places in the period from which the word α begins) to the length of the period. For example, in the case of sequence $010101\dots$ we have:

$$p_1 = 1/2,\, p_{01} = 1/2,\, p_{0101} = 1/2,\, p_{011} = 0. \tag{8}$$

Let $A = (X, Y, Q, h, f)$ be a strongly connected finite Moore machine with X and Y as input and output alphabets; Q as the set of states; $h : Q \times X \to Q$ as transition function; $f : Q \times X \to Y$ as output function.

Let us fix two sets of words

$$\{\alpha_i \in X^*, i = 1, 2, \ldots, t\} \text{ and } \{\beta_j \in Y^*, j = 1, 2, \ldots, k\}, \; t \geq 0, \; k \geq 1. \quad (9)$$

Let us suppose that an automaton A, starting to work from the state q_0, processes a sequence $\chi = (x_0, x_1, \ldots)$ into a sequence $\gamma = (y_0, y_1, \ldots)$. With sequence χ we associate the vector

$$z_{(A,q_0)}(\chi) = (p_{\alpha_1}(\chi), \ldots, p_{\alpha_t}(\chi), p_{\beta_1}(\gamma), \ldots, p_{\beta_k}(\gamma)), \quad (10)$$

if all quantities on the right-hand side exist.

The rule (10) defines a map

$$Z_{(A,q_0)} : T_X \to [0,1]^{t+k} \subset R^{t+k}. \quad (11)$$

The subject of our study is the closure (the set of all limit points) of the set $Z_{(A,q_0)}(T_X)$. This set will be denoted by R_A. The correctness of the accepted notation follows from the fact that if A is strongly connected, then $Z_{(A,q_0)}(T_X) = Z_{(A,q_0')}(T_X)$ for arbitrary two states q_0 and q_0'. It will be proved later (Theorem 1) that the set R_A is a convex polyhedron in the cube $[0,1]^{t+k}$. The set R_A will be called the polyhedron of the automaton A, corresponding to the sets of words $\{\alpha_i \in X^*, i = 1, 2, \ldots, t\}$ and $\{\beta_j \in Y^*, j = 1, 2, \ldots, k\}$.

The result of the Theorem 2 shows that if an automaton A processes a sufficiently long sequence χ with occurrences of words $\alpha_1, \ldots, \alpha_t$ close to $(p_{\alpha_1}(\chi), \ldots, p_{\alpha_t}(\chi))$ into a sequence γ with occurrences of words $\beta_1, \ldots, \beta_k$, close to $(p_{\beta_1}(\gamma), \ldots, p_{\beta_k}(\gamma))$, then point $(p_{\alpha_1}(\chi), \ldots, p_{\alpha_t}(\chi), p_{\beta_1}(\gamma), \ldots, p_{\beta_k}(\gamma))$ is located inside or near the automaton polyhedron.

4 The Automaton Polyhedron Structure

Let l be the maximum of the word lengths of the sets $\{\alpha_i \in X^*, i = 1, 2, \ldots, t\}$ and $\{\beta_j \in Y^*, i = 1, 2, \ldots, k\}$. We define the automaton $A^{(l)} = \left(X, Y, Q^{(l)}, h^{(l)}, f^{(l)}\right)$, by setting

$Q^{(l)} = \left\{\left(\left(q^{(1)}, x^{(1)}\right), \left(q^{(2)}, x^{(2)}\right), \ldots, \left(q^{(l-1)}, x^{(l-1)}\right), q^{(l)}\right)\right.$, where $h\left(q^{(i)}, x^{(i)}\right) = q^{(i+1)}$, $i = 1, 2, \ldots, l-1$; $q^{(j)} \in Q$, $j = 1, 2, \ldots, l$, $x^{(j)} \in X$, $\left.j = 1, 2, \ldots, l-1\right\}$ is a set of states;

$h^{(l)} : Q^{(l)} \times X \to Q^{(l)}$ is a transition function;
$h^{(l)}\left(\left(\left(q^{(1)}, x^{(1)}\right), \left(q^{(2)}, x^{(2)}\right), \ldots, \left(q^{(l-1)}, x^{(l-1)}\right), q^{(l)}\right), x\right)$
$= \left(\left(q^{(2)}, x^{(2)}\right), \ldots, \left(q^{(l-1)}, x^{(l-1)}\right), \left(q^{(l)}, x\right), h\left(q^{(l)}, x\right)\right)$;
$f^{(l)} : Q^{(l)} \times X \to Y$ is an output function;
$f^{(l)}\left(\left(\left(q^{(1)}, x^{(1)}\right), \left(q^{(2)}, x^{(2)}\right), \ldots, \left(q^{(l-1)}, x^{(l-1)}\right), q^{(l)}\right), x\right) = f\left(q^{(l)}, x\right)$.

By G_l we denote the transition graph of the automaton $A^{(l)}$, which arc $(q, h(l)(q, x))$ is labeled by the pair $(x, f(l)(q, x))$, $q \in Q(l)$. By an (oriented) cycle in a graph G_l we mean a cyclic sequence of pairwise distinct arcs in which the end of each arc coincides with the beginning of the next one. The set of all cycles in G_l is denoted by $C_l(A)$. With each cycle from $C_l(A)$ we associate the

cyclic sequences consisting of the first and second coordinates of this cycle arcs labels. These sequences will be called the input and output markups, respectively, taking the notation $c^{(x)}$ and $c^{(y)}$ for them.

For $\xi = (\xi_0, \xi_1, \ldots \xi_m) \in B^*$ by $\langle \xi \rangle$ we denote the periodic sequence $\xi_0, \xi_1, \ldots \xi_m, \xi_0, \xi_1, \ldots \xi_m, \ldots$ with the period ξ.

For $c \in C_l(A)$ we introduce the notation:

$l(c)$ – cycle length,

$\nu_\alpha(c) = \frac{1}{l} \sum_{j=0}^{l(c)-1} I_\alpha \left(\left\langle c^{(x)} \right\rangle T^j \right)$ – the word α occurrence relative frequency in the input markup $c^{(x)}$,

$\nu_\beta(c) = \frac{1}{l} \sum_{j=0}^{l(c)-1} I_\beta \left(\left\langle c^{(y)} \right\rangle T^j \right)$ – the word β occurrence relative frequency in the input markup $c^{(y)}$,

$z(c) = (\nu_{\alpha_1}(c), \ldots, \nu_{\alpha_k}(c), \nu_{\beta_1}(c), \ldots, \nu_{\beta_t}(c))$ – the relative frequencies vector.

If E is some set of points from R^n, then $ConvE$ denotes the convex hull of E.

Theorem 1. *The equality*

$$R_A = Conv\{z(c), c \in C_l(A)\} \tag{12}$$

holds.

Proof. Obviously, $z(c) \in R_A$ holds for $c \in C_l(A)$. Let $C_l(A) = \{c_1, c_2, \ldots, c_\theta\}$. Let us show that

$$\sum_{j=1}^{\theta} p_j z(c_j) \in R_A, \text{ if } \sum_{j=1}^{\theta} p_j = 1,\ p_1, p_2, \ldots, p_\theta \geq 0. \tag{13}$$

Let us choose $q_0 \in Q$. Let us fix arbitrarily $\varepsilon > 0$. The proof consists in construction of a periodic sequence χ, for which

$$\left| z_{(A,q_0)}(\chi) - \sum_{j=1}^{\theta} p_j z(c_j) \right| < \varepsilon. \tag{14}$$

Let $\tilde{q}^{(0)}$ be an arbitrary state from a set $Q^{(l)}$ of the form $((q', x'), \ldots, (q'', x''), q_0)$. Let $\tilde{q}^{(i)} \in Q^{(l)}$ be an arbitrary state through which the cycle c_i passes, $i = 1, 2, \ldots, \theta$. Let $\chi^{(i)} = (x_0^{(i)}, \ldots, x_{l_i-1}^{(i)})$ be the input sequence under which the automaton $A^{(l)}$ passes the cycle c_i, starting from state $\tilde{q}^{(i)}$, l_i being the length of the cycle c_i, $i = 1, 2, \ldots, \theta$. By $\eta(\tilde{q}, \tilde{q}')$ we denote the shortest input sequence that transfers the automaton $A^{(l)}$ from state $\tilde{q}$ to state $\tilde{q}'$. We denote $\xi_i = \eta\left(\tilde{q}^{(i)}, \tilde{q}^{(i+1)}\right)$, $i = 0, 1, \ldots, \theta - 1$, $\xi_\theta = \eta\left(\tilde{q}^{(\theta)}, \tilde{q}^{(0)}\right)$.

For a natural M by χ_M we denote a periodic sequence which period has the form

$$\xi_0 \wedge \left(\chi^{(1)}\right)^{[Mp_1]} \wedge \xi_1 \wedge \left(\chi^{(2)}\right)^{[Mp_2]} \wedge \ldots \wedge \xi_{\theta-1} \wedge \left(\chi^{(\theta)}\right)^{[Mp_\theta]} \wedge \xi_\theta, \tag{15}$$

where $[Mp_i]$ is an integer part Mp_i, and the symbol $\wedge$ means the concatenation of sequences.

It is easy to see that

$$Z_{(A,q_0)}(\chi_M) = \sum_{j=1}^{\theta} p_j z(c_j) + O\left(\frac{1}{M}\right). \tag{16}$$

Therefore, if M is sufficiently large, then the sequence χ_M satisfies the condition (14).

So, we have proved the inclusion

$$R_A \supseteq Conv\{z(c), c \in C_l(A)\}. \tag{17}$$

Reverse inclusion. If a periodic sequence $x^{(i)}$, $i = 1, 2, \ldots$ arrives at the automaton input, then the sequence of vectors

$$\left(\left(q^{(i)}, x^{(i)}\right), \left(q^{(i+1)}, x^{(i+1)}\right), \ldots, \left(q^{(i+l-1)}, x^{(i+l-1)}\right)\right), \tag{18}$$

where $h\left(q^{(i)}, x^{(i)}\right) = q^{(i+1)}$, is also periodic. We denote its period by L. Consider the vector of relative frequencies

$$z(L) = (\nu_{\alpha_1}(L), \ldots, \nu_{\alpha_k}(L), \nu_{\beta_1}(L), \ldots, \nu_{\beta_t}(L)). \tag{19}$$

We show that

$$z(L) \in Conv\{z(c), c \in C_l(A)\}. \tag{20}$$

Induction by $|L|$.

1^0. $|L| = 1$. This case corresponds to a loop in the graph G_l. Obviously, the set $C_l(A)$ contains all the loops of the graph. Therefore $z(L) \in Conv\{z(c), c \in C_l(A)\}$.

2^0. Suppose that for $|L'| < |L|$ vector $z(L')$, formed by the selected words relative frequencies in the input and output markups of period L', belongs to the set $Conv\{z(c), c \in C_l(A)\}$. Now let the length of the period be equal to $|L|$. Two cases are possible:

a) All sections of length l are different. Then the period L of the sequence in G_l question is a cycle in and therefore $z(L) \in Conv\{z(c), c \in C_l(A)\}$.
b) Period L contains a pair of matching sections of the length l. Without loss of generality, we can assume that L has the following form:

$$\begin{bmatrix} ((q_i, x_i), \ldots, (q_{i+l-1}, x_{i+l-1})), \ldots, (\ldots, (q', x')), \\ ((q_i, x_i), \ldots, (q_{i+l-1}, x_{i+l-1})), \ldots, (\ldots, (q'', x'')) \end{bmatrix}. \tag{21}$$

Let us consider two periodic sequences: ζ_1 with the period

$$L_1 = [((q_i, x_i), \ldots, (q_{i+l-1}, x_{i+l-1})), \ldots, (\ldots, (q', x'))] \tag{22}$$

of the length $|L_1|$ and ζ_2 with the period

$$L_2 = [((q_i, x_i), \ldots, (q_{i+l-1}, x_{i+l-1})), \ldots, (\ldots, (q'', x''))] \tag{23}$$

of the length $|L_2|$. The vectors of relative frequencies $z(L_1)$ and $z(L_2)$ correspond to these two sequences.

Whereas

$$|L_1| + |L_2| = |L|, \tag{24}$$

and the word lengths of the sets $\{\alpha_i \in X^*, i = 1, 2, \ldots, t\}$ and $\{\beta_j \in Y^*, j = 1, 2, \ldots, k\}$ are limited by l, it is easy to see that

$$z(L) = z(\zeta_1)\frac{|L_1|}{|L_1| + |L_2|} + z(\zeta_2)\frac{|L_2|}{|L_1| + |L_2|}. \tag{25}$$

By the induction hypothesis

$$z(\zeta_i) \in Conv\{z(c), c \in C_l(A)\}, \tag{26}$$

so

$$z(L) \in Conv\{z(c), c \in C_l(A)\}. \tag{27}$$

Thus we have established that

$$Z_{(A,q)}(T_X) \subseteq Conv\{z(c), c \in C_l(A)\}. \tag{28}$$

Since the set on the right-hand side is closed, we obtain the set inclusion

$$R_A \subseteq Conv\{z(c), c \in C_l(A)\}. \tag{29}$$

The theorem is proved.

5 The Example of the Automaton Polyhedron

We give an example of construction of a polyhedron of the automaton with two states. Let $A = (X = Y = Q = \{0, 1\}, h, f)$ be the finite automaton, where $h(q, x) = q \oplus x$ is XOR, $f(q, x) = q$.

Let us choose $k = t = 1$, $\alpha_1 = \beta_1 = 1$. The automaton graph is shown in the Fig. 1, it contains exactly three elementary cycles: a loop at the "0" vertex with a label $(0, 0)$, a loop at the "1" vertex with a label $(0, 1)$, and a cycle of length 2 between vertices which arc labels are $(1, 0)$ and $(1, 1)$.

The vectors $z(c)$ of these cycles are $(0, 0)$, $(0, 1)$ and $\left(1, \frac{1}{2}\right)$. Thus, $R_A = Conv\left\{(0, 0), (0, 1), \left(1, \frac{1}{2}\right)\right\}$. The automaton polyhedron in this case is a flat polygon; it is shown in the Fig. 2. The abscissa and ordinates correspond to the relative frequencies of occurrence of the characters "1" in the input and output sequences.

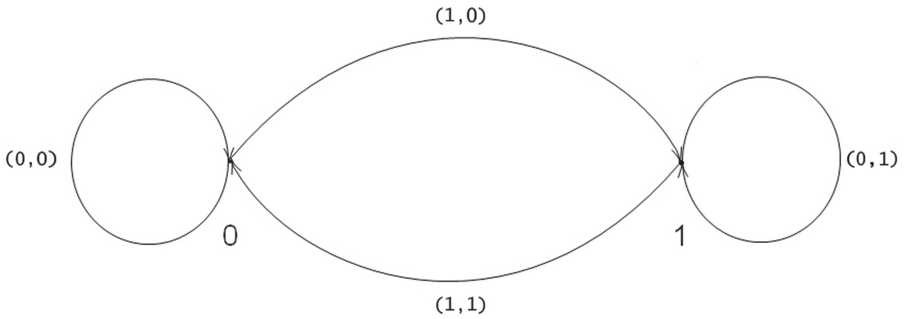

Fig. 1. The graph of the automaton A.

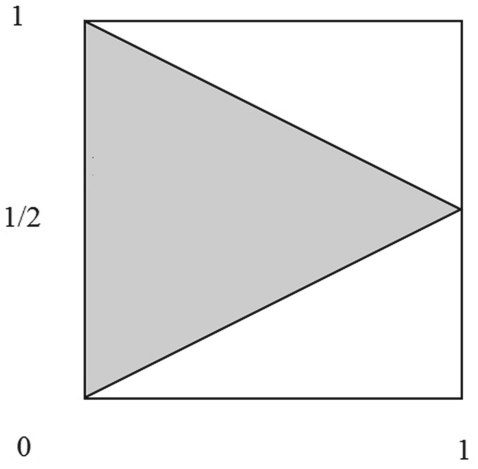

Fig. 2. The polygon of the automaton A.

6 The Case of Finite Sequences

The automaton A, starting to work from a certain initial state q_0, processes the input sequence $\chi^{(N)} = (x_0, x_1, \ldots, x_{N-1})$ into the output sequence $\gamma^{(N)} = (y_0, y_1, \ldots, y_{N-1})$. Let $p^{(N)}_{\alpha_1}, \ldots, p^{(N)}_{\alpha_t}$ be the relative frequencies of occurrence of words $\alpha_1, \ldots, \alpha_t$ in sequence $\chi^{(N)}$, and $p^{(N)}_{\beta_1}, \ldots, p^{(N)}_{\beta_k}$ be the relative frequencies of occurrence of words $\beta_1, \ldots, \beta_k$ in sequence $\gamma^{(N)}$:

$$\begin{aligned} p^{(N)}_{\alpha_s} &= \frac{\nu^{(N)}_{\alpha_s}}{N} = \frac{1}{N}\sum_{j=0}^{N-1} I_{\alpha_s}\left(\chi^{(N)}T^j\right),\ s = 1, 2, \ldots, t; \\ p^{(N)}_{\beta_r} &= \frac{\nu^{(N)}_{\beta_r}}{N} = \frac{1}{N}\sum_{j=0}^{N-1} I_{\beta_r}\left(\chi^{(N)}T^j\right),\ s = 1, 2, \ldots, k. \end{aligned} \tag{30}$$

Let us find out how far the point

$$z^{(N)} = \left(p_{\alpha_1}^{(N)}, \ldots, p_{\alpha_t}^{(N)}, p_{\beta_1}^{(N)}, \ldots, p_{\beta_k}^{(N)}\right) \tag{31}$$

can be located from polyhedron R_A. By the distance between two points $u, v \in R^{t+k}$ we mean the maximum modulus of the difference of coordinates:

$$\rho(u, v) = |u - v| = \max \left| u^{(i)} - v^{(i)} \right|. \tag{32}$$

The distance between the two sets F and G, F, $G \subset R^{t+k}$ is defined as the exact lower bound of the distances between these sets points: $\rho(F, G) = \inf |u - v|$, where the infimum is taken for all pairs $u \in F$, $v \in G$.

Processing by the automaton A of the sequence $\chi^{(N)}$ corresponds to movement in the graph G_l along arcs which first coordinates of the labels are elements of this sequence. Let $\tilde{q}^{(0)}$ be an arbitrary state from a set $Q^{(l)}$ of the form $((q', x'), \ldots, (q'', x''), q_0)$. The vertex $\tilde{q}^{(0)}$ is one of the possible beginnings of the path in question. Let $\tilde{q}$ be the graph vertex to which this path will come after processing the sequence $\chi^{(N)}$, starting from $\tilde{q}^{(0)}$. We supplement the sequence $\chi^{(N)}$ with the characters $x_N, \ldots, x_{N+m-1}$ selected in such a way as to go from state $\tilde{q}$ to the initial state $\tilde{q}^{(0)}$. We need $m \leq D_1$ characters, where D_1 is the diameter of the graph G_l. Denote the corresponding section of the output sequence $y_N, \ldots, y_{N+m-1}$.

Let $\chi = (x_0, x_1, \ldots, x_{N-1}, x_N, \ldots, x_{N+m-1})$. Let us estimate the distance between points $z(\chi)$ and $z^{(N)}$.

Let $\nu_{\alpha_s}^{(N)}$, $\nu_{\beta_r}^{(N)}$, $\nu_{\alpha_s}^{(N+m)}$, $\nu_{\beta_r}^{(N+m)}$ denote the frequencies of occurrence of the words α_s and β_r, $s = 1, 2, \ldots, t$, $r = 1, 2, \ldots, k$, in the sequences $\chi^{(N)}$ and χ and in the corresponding output sequences (in the periodic case – on the period).

Then, as it is easy to see,

$$0 \leq \nu_{\alpha_s}^{(N+m)} - \nu_{\alpha_s}^{(N)}, \nu_{\beta_r}^{(N+m)} - \nu_{\beta_r}^{(N)} \leq m + l - 1. \tag{33}$$

Therefore,

$$\begin{aligned}
&\left| z(\chi) - z^{(N)} \right| \\
&\quad \leq \max_{s,r} \left\{ \left| \frac{(N+M)\nu_{\alpha_s}^{(N)} - N\nu_{\alpha_s}^{(N+m)}}{N(N+M)} \right|, \left| \frac{(N+M)\nu_{\beta_r}^{(N)} - N\nu_{\beta_r}^{(N+m)}}{N(N+M)} \right| \right\} \\
&\leq \max_{s,r} \left\{ \left| \frac{m\nu_{\alpha_s}^{(N)} - N(\nu_{\alpha_s}^{(N+m)} - \nu_{\alpha_s}^{(N)})}{N(N+M)} \right|, \left| \frac{m\nu_{\beta_r}^{(N)} - N(\nu_{\beta_r}^{(N+m)} - \nu_{\beta_r}^{(N)})}{N(N+M)} \right| \right\} \\
&\qquad\qquad \leq \frac{m + l - 1}{N + m}.
\end{aligned} \tag{34}$$

Using the monotonicity of the function $\frac{x+A}{x+B}$ for $B > A > 0$, $x > 0$, we obtain

$$\left| z(\chi) - z^{(N)} \right| \leq \frac{D_l + l - 1}{N + D_l}. \tag{35}$$

Let D denote the diameter of the transition graph of the automaton A. Using the inequality $D_l \leq D + l - 1$, we obtain the following statement.

Theorem 2. *Suppose that an automaton A processes a sequence $\chi^{(N)} = (x_0, x_1, \ldots, x_{N-1})$ into a sequence $\gamma^{(N)} = (y_0, y_1, \ldots, y_{N-1})$, $N = 1, 2, \ldots$. Let $z^{(N)} = \left(p_{\alpha_1}^{(N)}, \ldots, p_{\alpha_t}^{(N)}, p_{\beta_1}^{(N)}, \ldots, p_{\beta_k}^{(N)}\right)$ be the vector of relative frequencies of occurrence of words $\alpha_1, \ldots, \alpha_t$ (for $\chi^{(N)}$) and $\beta_1, \ldots, \beta_k$ (for $\gamma^{(N)}$). Let D be the diameter of the transition graph of the automaton A.*

Then the inequality

$$\rho(z^{(N)}, R_A) \leq \frac{D + 2(l-1)}{N + D + l - 1}, \tag{36}$$

holds, where $l = \max\{|\alpha_i|, |\beta_j|\}$.

7 The Use of Automaton Polyhedra in the Identification Problem

By the task of identification we understand the task of testing of the hypothesis that an unknown automaton (which input and output sequences are observed) coincides with a reference automaton.

Theorem 2 allows us to construct the following procedure for verifying that an unknown automaton A is identical to a given automaton A_0.

1. Sets $\alpha_1, \ldots, \alpha_t$ of words (for the input sequence) and $\beta_1, \ldots, \beta_k$ (for the output sequence) are selected and the polyhedron R_{A_0} of the automaton A_0 is constructed.
2. The word occurrence relative frequencies in the observed sequences are calculated. The distance ρ between the automaton polyhedron and the relative frequency vector is calculated. If the frequencies vector belongs to the polyhedron, then $\rho = 0$.
3. If $\rho > \frac{D+2(l-1)}{N+D+l-1}$, then the observed output sequence could not be obtained from the input one using an automaton A_0. If $\rho \leq \frac{D+2(l-1)}{N+D+l-1}$, then the observed frequencies of the selected words do not contradict the hypothesis that the unknown automaton is identical to the reference automaton. In the latter case it is reasonable either to move to another segment of the available sequences, or to change the word sets which frequencies are analyzed.

The described procedure is valid for arbitrary word sets $\{\alpha_1, \ldots, \alpha_t\}$ and $\{\beta_1, \ldots, \beta_k\}$. In particular, the set $\{\alpha_1, \ldots, \alpha_t\}$ may be empty. In this case, the analysis is based on word frequencies only in the output automaton sequence.

We emphasize that the described procedure, firstly, does not depend on the analyzed automaton initial state, and secondly, despite the fact that the certain event frequencies in the observed sequences are analyzed, it does not use any assumptions about the input sequence probabilistic nature.

Let us analyze the proposed procedure computational complexity if the number $|Q|$ of states of the automaton A_0 is large. It is determined by the contribution of two terms. Firstly, the preliminary polyhedron construction complexity, and secondly, the complexity of checking inequality (36).

The construction of a polyhedron by the Theorem 1 requires finding of all the cycles of the graph G_l and constructing of the convex hull of the set $\{z(c)\}$. Both of these problems are well studied; see, for example, [12–14].

The complexity of finding of all the cycles in our case can be limited by the number $O\left(2^{(|Q|\times|X|)^l}\right)$ of all subgraphs of G_l. The convex hull constructing complexity, in the case of a flat polygon or a three-dimensional polyhedron, can be estimated [14] as $O\left(|\{z(c)\}|\, Log\, |\{z(c)\}|\right)$. Note that analytical methods for construction of polyhedra are possible for some automaton classes.

If the polyhedron R_{A_0} is already constructed, then the inequality (36) checking complexity, as is easy to see, is not more than 2^{t+k} times the complexity of checking whether a given point belongs to a convex polyhedron R_{A_0}. The computational complexity of the last problem in the two-dimensional case ($t = k = 1$) can be estimated [14] by the value $O\left(Log v\right)$, where v is the polygon vertex number. Note that even faster algorithms [15] are proposed. To estimate v, we use the fact that all the polygon R_{A_0} vertices have the form $\left(\frac{p_1}{q_1}, \frac{p_2}{q_2}\right)$, $0 \le p_i \le q_i \le \left|Q^{(l)}\right|$, $i = 1, 2$. Counting the possible different vertex abscissa number, due to the polygon convexity, we get $v \le |Q|^{2l}$. Therefore, the inequality (36) check complexity in the case of a preliminarily constructed polygon is estimated as $O\left(Log|Q|\right)$.

Generally speaking, we can select several shorter continuous fragments of the observed sequences, and perform the procedure for each of them separately. If inequality (36) is violated for at least one fragment, the hypothesis about the coincidence of automata is rejected.

8 Conclusion

A method for verifying that an automaton which input and output sequences are observed coincides with the reference one is proposed. The method uses word occurrence frequencies in the input and output sequences. Specially selected input sequences are not required. Information on the analyzed machine initial state is not required.

If the polyhedron of an unknown automaton coincides with the reference automaton polyhedron for given sets of words in the input and output sequences, then the proposed procedure cannot distinguish between these automata. Therefore, the problem arises of classifying automata by their polyhedra.

It is intuitively clear that two automata, the polyhedra of which have an insignificant common part, are easily distinguishable. The important thing here is how likely it is that the point corresponding to a sequence fragment falls into the both polyhedral common part. It depends on the probability distribution on the input sequence set.

Acknowledgments. The publication has been prepared with the support of the RUDN University Program "5-100" (recipient K. Samouylov). The reported study was funded by RFBR, project numbers 19-07-00933 and 18-00-01555 (18-00-01685).

References

1. Marsaglia, G.: Random numbers fall mainly in the planes. Proc. Natl. Acad. Sci. **61**(1968), 5–28 (1968)
2. Haramoto, H., Matsumoto, M.: Again, random numbers fall mainly in the planes: xorshift128+, arxiv.org/abs/1908.10020. Accessed 30 Jun 2020
3. Tverdokhlebov, V.A.: Geometrical approach to technical diagnosing of automatons. In: Proceedings of the IEEE East-West Design & Test Symposium, EWDTS 2011, National University of Radioelectronics, Kharkov, pp. 240–243 (2011)
4. Babash, A.V.: Automaton barriers. Math. Notes **91**(5–6), 625–629 (2012)
5. Parkhomenko, D.V.: Automata generated p-languages. Discrete Math. Appl. **24**(4), 207–212 (2014)
6. Chatterjee, K., Doyen, L., Edelsbrunner, H., Henzinger, T.A., Rannou, P.: Mean-payoff automaton expressions. In: Gastin, P., Laroussinie, F. (eds.) CONCUR 2010. LNCS, vol. 6269, pp. 269–283. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15375-4_19
7. Grusho, A.A., Timonina, E.E.: Prohibitions in discrete probabilistic statistical problems. Discrete Math. Appl. **21**(3), 275–281 (2011)
8. Gerasimenko, M., Petrov, V., Galinina, O., Andreev, S., Koucheryavy, Y.: Energy and delay analysis of LTE-Advanced RACH performance under MTC overload. In: 2012 IEEE Globecom Workshops, GC Wkshps 2012, pp. 1632–1637 (2012). Art. no. 6477830
9. Pyattaev, A., Johnsson, K., Surak, A., Florea, R., Andreev, S., Koucheryavy, Y.: Network-assisted D2D communications: implementing a technology prototype for cellular traffic offloading. In: IEEE Wireless Communications and Networking Conference, WCNC, pp. 3266–3271 (2014). Art. no. 6953070
10. Ometov, A., et al.: Toward trusted, social-aware D2D connectivity: bridging across the technology and sociality realms. IEEE Wirel. Commun. **23**(4), 103–111 (2016). Art. no. 7553033
11. Jacobs, K.: Turing-Maschinen und zufällige 0–1-Folgen. In: Jacobs, K. (ed.) Selecta Mathematica II Heidelberger Taschenbücher, vol. 67, pp. 141–167. Springer, Heidelberg (1970). https://doi.org/10.1007/978-3-642-88162-6_6
12. Johnson, D.B.: Finding all the elementary circuits of a directed graph. SIAM J. Comput. **4**(1), 77–84 (1975)
13. Liu, H., Wang, J.: A new way to enumerate cycles in graph. In: Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, AICT-ICIW 2006, Washington, DC, USA, pp. 57–59 (2006)
14. Preparata, F.P., Shamos, M.I.: Computational Geometry - An Introduction. Monographs in Computer Science. Springer, New York (1988). https://doi.org/10.1007/978-1-4612-1098-6
15. Skala, V.: Point-in-convex polygon and point-in-convex polyhedron algorithms with O(1) complexity using space subdivision. In: ICNAAM 2015, pp. 22–28. AIP Publishing LLC. (2015)