Decision trees for regular factorial languages

Array 15 (2022) 100203

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array

Decision trees for regular factorial languages

Mikhail Moshkov

Computer, Electrical and Mathematical Sciences and Engineering Division and Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

ARTICLE INFO

Keywords:
Regular factorial language
Recognition problem
Membership problem
Deterministic decision tree
Nondeterministic decision tree

ABSTRACT

In this paper, we study arbitrary regular factorial languages over a finite alphabet Σ. For the set of words $L(n)$ of the length n belonging to a regular factorial language L, we investigate the depth of decision trees solving the recognition and the membership problems deterministically and nondeterministically. In the case of recognition problem, for a given word from $L(n)$, we should recognize it using queries each of which, for some $i \in \{1, \dots, n\}$, returns the ith letter of the word. In the case of membership problem, for a given word over the alphabet Σ of the length n, we should recognize if it belongs to the set $L(n)$ using the same queries. For a given problem and type of trees, instead of the minimum depth $h(n)$ of a decision tree of the considered type solving the problem for $L(n)$, we study the smoothed minimum depth $H(n) = \max\{h(m) : m \leq n\}$. With the growth of n, the smoothed minimum depth of decision trees solving the problem of recognition deterministically is either bounded from above by a constant, or grows as a logarithm, or linearly. For other cases (decision trees solving the problem of recognition nondeterministically, and decision trees solving the membership problem deterministically and nondeterministically), with the growth of n, the smoothed minimum depth of decision trees is either bounded from above by a constant or grows linearly. As corollaries of the obtained results, we study joint behavior of smoothed minimum depths of decision trees for the considered four cases and describe five complexity classes of regular factorial languages. We also investigate the class of regular factorial languages over the alphabet $\{0, 1\}$ each of which is given by one forbidden word.

1. Introduction

In this paper, we study arbitrary regular factorial languages over a finite alphabet Σ. A factorial language satisfies the following condition: if a word w_1uw_2 belongs to the language, then the word u also belongs to it. For the set of words $L(n)$ of the length n belonging to a regular factorial language L, we investigate the depth of decision trees solving the recognition and the membership problems deterministically and nondeterministically. In the case of recognition problem, for a given word from $L(n)$, we should recognize it using queries each of which, for some $i \in \{1, \dots, n\}$, returns the ith letter of the word. In the case of membership problem, for a given word over the alphabet Σ of the length n, we should recognize if it belongs to $L(n)$ using the same queries.

For a given problem (problem of recognition or membership problem) and type of trees (solving the problem deterministically or nondeterministically), instead of the minimum depth $h(n)$ of a decision tree of the considered type solving the problem for $L(n)$, we study the smoothed minimum depth $H(n) = \max\{h(m) : m \leq n\}$.

For an arbitrary regular factorial language, with the growth of n, the smoothed minimum depth of decision trees solving the problem of recognition deterministically is either bounded from above by a constant, or grows as a logarithm, or linearly. These results follow immediately from more general, obtained in [1] for arbitrary regular languages.

For other cases (decision trees solving the problem of recognition nondeterministically, and decision trees solving the membership problem deterministically and nondeterministically), with the growth of n, the smoothed minimum depth of decision trees is either bounded from above by a constant, or grows linearly. In the conference paper [2], a classification of arbitrary regular languages depending on the smoothed minimum depth of decision trees solving the problem of recognition nondeterministically was announced without proofs. In the present paper, we consider simpler classification for regular factorial languages with full proof. Results related to the decision trees solving the membership problem are new.

As corollaries of the obtained results, we study joint behavior of smoothed minimum depths of decision trees for the considered four cases and describe five complexity classes of regular factorial languages. We also investigate the class of regular factorial languages over the alphabet $E = \{0, 1\}$ each of which is given by one forbidden word.

A well-known approach to evaluate complexity of an infinite language L over a finite alphabet Σ is to study its so-called combinatorial

E-mail address: mikhail.moshkov@kaust.edu.sa.

https://doi.org/10.1016/j.array.2022.100203
Received 7 January 2022; Received in revised form 2 June 2022; Accepted 3 June 2022
Available online 8 June 2022

complexity (known also as counting function) $f_L(n)$ that is the number of words of the length n in L [3,4]. The present paper proposes additional ways to evaluate the complexity of the language L based on the study how the depth of decision trees solving the recognition and the membership problems deterministically and nondeterministically depends on the length of words. This way is more complicated, but can give more detailed classification of languages. To show this, we compare languages generated by diagrams I_3 and I_4 depicted in Figs. 5 and 6. For both languages, the counting function grows linearly. For the first language, the minimum depth of decision trees solving the problem of recognition deterministically grows as a logarithm, but for the second language, the minimum depth of decision trees solving the problem of recognition deterministically grows linearly.

We should mention a recent paper [5] in which similar results were obtained for languages over the alphabet E that are subword-closed: if a word $w_1u_1w_2 \cdots w_mu_mw_{m+1}$ belongs to the language, then the word $u_1 \cdots u_m$ also belongs to it.

It is clear that each subword-closed language is a factorial language. Moreover, each subword-closed language over a finite alphabet is a regular language [6]. One can show that the language $L(00)$ over the alphabet E given by one forbidden word 00 is a regular factorial language, which is not subword-closed. Therefore the class of subword-closed languages over the alphabet E is a proper subclass of the class of regular factorial languages over the alphabet E.

The main difference between the present paper and [5] is that, in the latter paper, we do not assume that the subword-closed languages are given by deterministic finite automata. Instead of this, we describe simple criteria (based on the presence in the language of words of special types) for the behavior of the minimum depths of decision trees solving the problem of recognition deterministically and nondeterministically. Differently formulated criteria for the behavior of the minimum depth of decision trees solving the recognition problem require very different proofs. One more difference is that in [5] we directly consider the minimum depth of decision trees.

The rest of the paper is organized as follows. In Section 2, we consider main notions, in Section 3 – main results, and in Section 4 – two corollaries of these results.

2. Main notions

In this section, we discuss the notions related to regular factorial languages and decision trees solving problems of recognition and membership for these languages.

2.1. Regular factorial languages

Let $\omega = \{0, 1, 2, \dots\}$ be the set of nonnegative integers and Σ be a finite alphabet with at least two letters. By Σ^*, we denote the set of all finite words over the alphabet Σ, including the empty word λ. A word $w \in \Sigma^*$ is called a factor of a word $u \in \Sigma^*$ if $u = v_1wv_2$ and $v_1, v_2 \in \Sigma^*$. A language $L \subseteq \Sigma^*$ is called factorial if it contains all factors of its words. A word $w \in \Sigma^*$ is called a minimal forbidden word for L if $w \notin L$ and all proper factors of w belong to L. We denote by $MF(L)$ the language of minimal forbidden words for L. It is known [7] that a factorial language L is regular if and only if the language $MF(L)$ is regular. In particular, a factorial language L with a finite set of minimal forbidden words $MF(L)$ is regular. In this paper, we study arbitrary nonempty regular factorial languages.

It is well known that each regular language can be represented by a deterministic finite automaton (DFA) [8]. As in [8], we will consider not only complete DFA with total transition function but also partial DFA with partial transition function. Such DFA can be represented by its transition diagram (diagram for short) [9].

A diagram over the alphabet Σ is a triple $I = (G, q_0, Q)$, where G is a finite directed graph, possibly with multiple edges and loops, in which each edge is labeled with a letter from Σ and edges leaving each node are labeled with pairwise different letters, q_0 is a node of G called starting, and Q is a nonempty set of the graph G nodes called final.

A path of the diagram I is an arbitrary sequence $\xi = v_1, d_1, \dots, v_m, d_m, v_{m+1}$ of nodes and edges of G such that the edge d_i leaves the node v_i and enters the node v_{i+1} for $i = 1, \dots, m$. We now define a word $w(\xi)$ from Σ^* in the following way: if $m = 0$, then $w(\xi) = \lambda$. Let $m > 0$ and let δ_j be the letter attached to the edge d_j, $j = 1, \dots, m$. Then $w(\xi) = \delta_1 \cdots \delta_m$. We say that the path ξ generates the word $w(\xi)$. Note that different paths which start in the same node generate different words.

We denote by $\Xi(I)$ the set of all paths of the diagram I each of which starts in the node q_0 and finishes in a node from Q. Let

$$L_I = \{w(\xi) : \xi \in \Xi(I)\}.$$

We say that the diagram I generates the language L_I. It is well known that L_I is a regular language.

The diagram I is called complete over the alphabet Σ if exactly $|\Sigma|$ edges leave each node of G. Note that these edges are labeled with pairwise different letters from Σ. Such diagram corresponds to a complete DFA [8]. The diagram I is called reduced if, for each node of G, there exists a path from $\Xi(I)$, which contains this node. Such diagram corresponds to a reduced DFA [8]. It is known [8] that, for each regular language over the alphabet Σ, there exists a complete over the alphabet Σ diagram, which generates this language. Therefore, for each nonempty regular language, there exists a reduced diagram, which generates this language.

Let L be a regular factorial language and $I = (G, q_0, Q)$ be a reduced diagram that generates the language L. Since the language L is factorial, we can assume additionally that each node of the graph G is final — it will not change the language generated by I since with each word the language L contains each prefix of this word. The diagram I will be called f-reduced if it is reduced and each node of the graph G is final. Further we will assume that a considered regular factorial language L is nonempty and it is given by an f-reduced diagram, which generates this language.

We will not consider nondeterministic finite automata (NFA) to represent regular factorial languages since the study of NFA is essentially more complicated task.

2.2. Decision trees for recognition and membership problems

Let L be a regular factorial language over the alphabet Σ. For any natural n, denote $L(n) = L \cap \Sigma^n$, where Σ^n is the set of words over the alphabet Σ, which length is equal to n. We consider two problems related to the set $L(n)$. The problem of recognition: for a given word from $L(n)$, we should recognize it using attributes (queries) $l_1^n, \dots, l_n^n$, where l_i^n, $i \in \{1, \dots, n\}$, is a function from Σ^n to Σ such that $l_i^n(a_1 \cdots a_n) = a_i$ for any word $a_1 \cdots a_n \in \Sigma^n$. The problem of membership: for a given word from Σ^n, we should recognize if this word belongs to the set $L(n)$ using the same attributes. To solve these problems, we use decision trees over $L(n)$.

A decision tree over $L(n)$ is a marked finite directed tree with root, which has the following properties:

- The root and the edges leaving the root are not labeled.
- Each node, which is not the root nor terminal node, is labeled with an attribute from the set $\{l_1^n, \dots, l_n^n\}$.
- Each edge leaving a node, which is not a root, is labeled with a letter from the alphabet Σ.

A decision tree over $L(n)$ is called deterministic if it satisfies the following conditions:

- Exactly one edge leaves the root.
- For any node, which is not the root nor terminal node, the edges leaving this node are labeled with pairwise different letters.

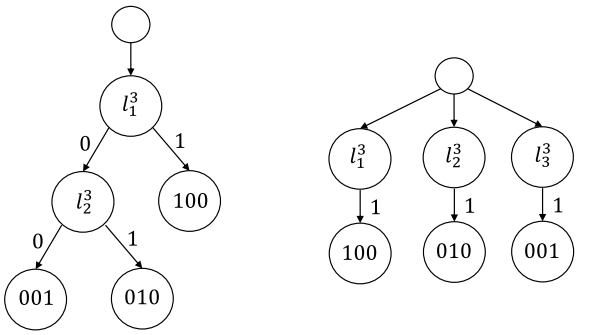

Fig. 1. Decision trees that solve the problem of recognition for the set of words $\{100, 010, 001\}$ deterministically and nondeterministically.

Let Γ be a decision tree over $L(n)$. A complete path in Γ is any sequence $\xi = v_0, e_0, \dots, v_m, e_m, v_{m+1}$ of nodes and edges of Γ such that v_0 is the root, v_{m+1} is a terminal node, and v_i is the initial and v_{i+1} is the terminal node of the edge e_i for $i = 0, \dots, m$. We define a subset $\Sigma(n, \xi)$ of the set Σ^n in the following way: if $m = 0$, then $\Sigma(n, \xi) = \Sigma^n$. Let $m > 0$, the attribute $l^n_{i_j}$ be attached to the node v_j, and b_j be the letter attached to the edge e_j, $j = 1, \dots, m$. Then

$$\Sigma(n, \xi) = \{a_1 \cdots a_n \in \Sigma^n : a_{i_1} = b_1, \dots, a_{i_m} = b_m\}.$$

Let $L(n) \neq \emptyset$. We say that a decision tree Γ over $L(n)$ solves the problem of recognition for $L(n)$ nondeterministically if Γ satisfies the following conditions:

- Each terminal node of Γ is labeled with a word from $L(n)$.
- For any word $w \in L(n)$, there exists a complete path ξ in the tree Γ such that $w \in \Sigma(n, \xi)$.
- For any word $w \in L(n)$ and for any complete path ξ in the tree Γ such that $w \in \Sigma(n, \xi)$, the terminal node of the path ξ is labeled with the word w.

We say that a decision tree Γ over $L(n)$ solves the problem of recognition for $L(n)$ deterministically if Γ is a deterministic decision tree, which solves the problem of recognition for $L(n)$ nondeterministically.

Examples of decision trees illustrating the considered notions are presented in Fig. 1.

We say that a decision tree Γ over $L(n)$ solves the problem of membership for $L(n)$ nondeterministically if Γ satisfies the following conditions:

- Each terminal node of Γ is labeled with a number from the set $\{0, 1\}$.
- For any word $w \in \Sigma^n$, there exists a complete path ξ in the tree Γ such that $w \in \Sigma(n, \xi)$.
- For any word $w \in \Sigma^n$ and for any complete path ξ in the tree Γ such that $w \in \Sigma(n, \xi)$, the terminal node of the path ξ is labeled with the number 1 if $w \in L(n)$ and with the number 0, otherwise.

We say that a decision tree Γ over $L(n)$ solves the problem of membership for $L(n)$ deterministically if Γ is a deterministic decision tree which solves the problem of membership for $L(n)$ nondeterministically.

Let Γ be a decision tree over $L(n)$. We denote by $h(\Gamma)$ the maximum number of nodes in a complete path in Γ that are not the root nor terminal node. The value $h(\Gamma)$ is called the depth of the decision tree Γ.

We denote by $h^{ra}_L(n)$ ($h^{rd}_L(n)$) the minimum depth of a decision tree over $L(n)$, which solves the problem of recognition for $L(n)$ nondeterministically (deterministically). If $L(n) = \emptyset$, then $h^{ra}_L(n) = h^{rd}_L(n) = 0$.

We denote by $h^{ma}_L(n)$ ($h^{md}_L(n)$) the minimum depth of a decision tree over $L(n)$, which solves the problem of membership for $L(n)$ nondeterministically (deterministically). If $L(n) = \emptyset$, then $h^{ma}_L(n) = h^{md}_L(n) = 0$.

3. Bounds on decision tree depth

Let L be a nonempty factorial regular language. In this section, we consider the behavior of four functions H^{ra}_L, H^{rd}_L, H^{ma}_L, and H^{md}_L defined on the set $\omega \setminus \{0\}$ and with values from ω. For any natural n,

$$H^{ra}_L(n) = \max\{h^{ra}_L(m) : 1 \le m \le n\},$$
$$H^{rd}_L(n) = \max\{h^{rd}_L(m) : 1 \le m \le n\},$$
$$H^{ma}_L(n) = \max\{h^{ma}_L(m) : 1 \le m \le n\},$$
$$H^{md}_L(n) = \max\{h^{md}_L(m) : 1 \le m \le n\}.$$

For any pair $bc \in \{ra, rd, ma, md\}$, the function $H^{bc}_L(n)$ is a smoothed analog of the function $h^{bc}_L(n)$.

3.1. Decision trees solving recognition problem deterministically

Let $I = (G, q_0, Q)$ be a f-reduced diagram over the alphabet Σ. A path of the diagram I is called a cycle of the diagram I if there is at least one edge in this path, and the first node of this path is equal to the last node of this path. A cycle of the diagram I is called elementary if nodes of this cycle, with the exception of the last node, are pairwise different.

The diagram I is called simple if every two different elementary cycles of the diagram I do not have common nodes. Let I be a simple diagram and ξ be a path of the diagram I. The number of different elementary cycles of the diagram I, which have common nodes with ξ, is denoted by $cl(\xi)$ and is called the cyclic length of the path ξ. The value

$$cl(I) = \max\{cl(\xi) : \xi \in \Xi(I)\}$$

is called the cyclic length of the diagram I.

Let I be a simple diagram, C be an elementary cycle of the diagram I, and v be a node of the cycle C. Beginning with the node v, the cycle C generates an infinite periodic word over the alphabet Σ. This word will be denoted by $W(I, C, v)$. We denote by $r(I, C, v)$ the minimum period of the word $W(I, C, v)$. The diagram I is called dependent if there exist two different elementary cycles C_1 and C_2 of the diagram I, nodes v_1 and v_2 of the cycles C_1 and C_2, respectively, and a path π of the diagram I from v_1 to v_2, which satisfy the following conditions: $W(I, C_1, v_1) = W(I, C_2, v_2)$ and the length of the path π is a number divisible by $r(I, C_1, v_1)$. If the diagram I is not dependent, then it is called independent. Next theorem follows immediately from Theorem 2.1 [1], which is a similar statement that holds for all regular languages.

Theorem 1. *Let L be a nonempty regular factorial language over the alphabet Σ and I be a f-reduced diagram, which generates the language L. Then the following statements hold:*

(a) If I is an independent simple diagram and $cl(I) \le 1$, then $H^{rd}_L(n) = O(1)$.

(b) If I is an independent simple diagram and $cl(I) \ge 2$, then $H^{rd}_L(n) = \Theta(\log n)$.

(c) If I is not independent simple diagram, then $H^{rd}_L(n) = \Theta(n)$.

3.2. Decision trees solving recognition problem nondeterministically

Let L be a nonempty regular factorial language over the alphabet Σ. For any natural n, we define a parameter $T_L(n)$ of the language L. If $L(n) = \emptyset$, then $T_L(n) = 0$. Let $L(n) \neq \emptyset$, $w = a_1 \cdots a_n \in L(n)$, and $J \subseteq \{1, \dots, n\}$. Denote $L(w, J) = \{b_1 \cdots b_n \in L(n) : b_j = a_j, j \in J\}$ (if $J = \emptyset$, then $L(w, J) = L(n)$) and $M_L(n, w) = \min\{|J| : J \subseteq \{1, \dots, n\}, |L(w, J)| = 1\}$. Then

$$T_L(n) = \max\{M_L(n, w) : w \in L(n)\}.$$

Note that, for any word $w \in L(n)$, $M_L(n, w)$ is the minimum number of letters of the word w, which allow us to distinguish it from all other words belonging to $L(n)$.

Lemma 2. *Let L be a nonempty regular factorial language over the alphabet Σ. Then $h_L^{ra}(n) = T_L(n)$ for any natural n.*

Proof. First, we prove that $h_L^{ra}(n) \geq T_L(n)$. Let Γ be a decision tree over $L(n)$, which solves the problem of recognition for $L(n)$ nondeterministically and for which $h(\Gamma) = h_L^{ra}(n)$. Let w be a word from $L(n)$ for which $T_L(n) = M_L(n, w)$. Then the decision tree Γ contains a complete path ξ such that $w \in \Sigma(n, \xi)$ and the terminal node of the path ξ is labeled with the word w. It is clear that $\Sigma(n, \xi) \cap L(n) = \{w\}$. Let ξ contain m nodes that are not the root nor terminal node and $l_{i_1}^n, \ldots, l_{i_m}^n$ be attributes attached to these nodes. Denote $J = \{i_1, \ldots, i_m\}$. Then $L(w, J) = \{w\}$. Therefore $m \geq M_L(n, w) = T_L(n)$. It is clear that $h(\Gamma) \geq m$. Thus, $h_L^{ra}(n) = h(\Gamma) \geq m \geq M_L(n, w) = T_L(n)$.

We now prove that $h_L^{ra}(n) \leq T_L(n)$. One can show that, for each $w \in L(n)$, we can construct a complete path ξ_w, which satisfies the following conditions: the number of nodes in ξ_w that are not the root nor terminal node is equal to $M_L(n, w)$, $\Sigma(n, \xi_w) \cap L(n) = \{w\}$, and the terminal node of ξ_w is labeled with the word w. If we merge roots of all paths ξ_w, $w \in L(n)$, we obtain a decision tree, which solves the problem of recognition for $L(n)$ nondeterministically and which depth is equal to $T_L(n)$. Thus, $h_L^{ra}(n) \leq T_L(n)$ and $h_L^{ra}(n) = T_L(n)$. □

Theorem 3. *Let L be a nonempty regular factorial language over the alphabet Σ and $I = (G, q_0, Q)$ be a f-reduced diagram, which generates the language L. Then the following statements hold:*

(a) If I is an independent simple diagram, then $H_L^{ra}(n) = O(1)$.

(b) If I is not independent simple diagram, then $H_L^{ra}(n) = \Theta(n)$.

Proof. (a) Let I be an independent simple diagram and $cl(I) \leq 1$. By Theorem 1, $H_L^{rd}(n) = O(1)$. It is clear that $H_L^{ra}(n) \leq H_L^{rd}(n)$. Therefore $H_L^{ra}(n) = O(1)$.

Let I be an independent simple diagram and $cl(I) \geq 2$. Let n be a natural number. If $L(n) = \emptyset$, then $T_L(n) = 0$. Let $L(n) \neq \emptyset$. Denote by d the number of nodes in the graph G. In the proof of Lemma 4.5 [1], it was proved that $M_L(n, w) \leq d(4d + 1)$ for any word $w \in L(n)$. Therefore $T_L(n) \leq d(4d + 1)$. Thus, by Lemma 2, $h_L^{ra}(n) \leq d(4d + 1)$ for any natural n and $H_L^{ra}(n) = O(1)$.

(b) Let I be not simple diagram and C_1, C_2 be different elementary cycles of the diagram I, which have a common node v. Since I is a f-reduced diagram, it contains a path ξ from the node q_0 to the node v, and v is a final node. Let the length of the path ξ be equal to a, the length of the cycle C_1 be equal to b, and the length of the cycle C_2 be equal to c. Let α be the word generated by the path ξ, β be the word generated by a path from v to v obtained by the passage c times along the cycle C_1, and γ be the word generated by a path from v to v obtained by the passage b times along the cycle C_2. The words β and γ are different and they have the same length bc.

Consider the sequence of numbers $n_i = a + ibc$, $i = 1, 2, \ldots$. Let $i \in \omega \setminus \{0\}$. The set $L(n_i)$ contains the word $\alpha\gamma^i$ and the words $\alpha\gamma^j\beta\gamma^{i-j-1}$ for $j = 0, \ldots, i-1$. It is easy to show that $M_L(n_i, \alpha\gamma^i) \geq i$: to distinguish the word $\alpha\gamma^i$ from the words $\alpha\gamma^j\beta\gamma^{i-j-1}$, $j = 0, \ldots, i-1$, we need to use at least one letter from each of i words γ appearing in $\alpha\gamma^i$. Therefore $T_L(n_i) \geq i$ and, by Lemma 2, $h_L^{ra}(n_i) \geq i = (n_i - a)/(bc)$. Let $n \geq n_1$ and let i be the maximum natural number such that $n \geq n_i$. Evidently, $n - n_i \leq bc$. Hence $H_L^{ra}(n) \geq h_L^{ra}(n_i) \geq (n - bc - a)/(bc)$. Therefore $H_L^{ra}(n) \geq n/(2bc)$ for large enough n. The inequality $H_L^{ra}(n) \leq n$ is obvious. Thus, $H_L^{ra}(n) = \Theta(n)$.

Let I be a dependent simple diagram. Then there exist two different elementary cycles C_1 and C_2 of the diagram I, nodes v_1 and v_2 of the cycles C_1 and C_2, respectively, and a path π of the diagram I from v_1 to v_2, which satisfy the following conditions: $W(I, C_1, v_1) = W(I, C_2, v_2)$ and the length of the path π is a number divisible by $r(I, C_1, v_1)$. Let us remind that, for $i = 1, 2$, $W(I, C_i, v_i)$ is the infinite periodic word over the alphabet Σ generated by the cycle C_i beginning with the node v_i, and $r(I, C_1, v_1)$ is the minimum period of the word $W(I, C_1, v_1)$. Since I is a f-reduced diagram, it contains a path ξ from the node q_0 to the node v_1, and all nodes of the graph G are final. Let the path ξ generate the word α of the length a. Denote $r = r(I, C_1, v_1)$. Let the length of the cycle C_1 be equal to br, the length of the path π be equal to cr, and the path π generate the word β. Denote by γ the prefix of the length r of the word $W(I, C_1, v_1)$. We now define two words of the length rbc: $u = \gamma^{bc}$ and $w = \beta\gamma^{c(b-1)}$. It is clear that $u \neq w$.

Consider the sequence of numbers $n_i = a + irbc$, $i = 1, 2, \ldots$. Let $i \in \omega \setminus \{0\}$. The set $L(n_i)$ contains the word αu^i and the words $\alpha u^j w u^{i-j-1}$ for $j = 0, \ldots, i-1$. It is easy to show that $M_L(n, \alpha u^i) \geq i$: to distinguish the word αu^i from the words $\alpha u^j w u^{i-j-1}$, $j = 0, \ldots, i-1$, we need to use at least one letter from each of i words u appearing in αu^i. Therefore $T_L(n_i) \geq i$ and, by Lemma 2, $h_L^{ra}(n_i) \geq i = (n_i - a)/(rbc)$. Let $n \geq n_1$ and let i be the maximum natural number such that $n \geq n_i$. Evidently, $n - n_i \leq rbc$. Hence $H_L^{ra}(n) \geq h_L^{ra}(n_i) \geq (n - rbc - a)/(rbc)$. Therefore $H_L^{ra}(n) \geq n/(2rbc)$ for large enough n. The inequality $H_L^{ra}(n) \leq n$ is obvious. Thus, $H_L^{ra}(n) = \Theta(n)$. □

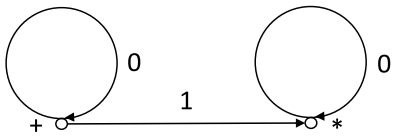

Fig. 2. Diagram I_0.

Note that in general case (when we consider not only factorial languages) the classification of reduced diagrams depending on the minimum depth of decision trees solving the problem of recognition nondeterministically is more complicated [2]. In particular, there exists a dependent simple reduced diagram I_0 (see Fig. 2) with the starting node labeled with the symbol + and the unique final node labeled with the symbol $*$ that generates the regular language $L_0 = \{0^i 10^j : i, j \in \omega\}$ over the alphabet $\{0, 1\}$, which is not factorial and for which $H_{L_0}^{ra}(n) = O(1)$.

3.3. Decision trees solving membership problem

For a regular factorial language L, the notation $|L| = \infty$ means that L is an infinite language, and the notation $|L| < \infty$ means that L is a finite language.

Theorem 4. *Let L be a regular factorial language over the alphabet Σ.*

(a) If $|L| = \infty$ and $L \neq \Sigma^*$, then $H_L^{md}(n) = \Theta(n)$ and $H_L^{ma}(n) = \Theta(n)$.

(b) If $|L| < \infty$ or $L = \Sigma^*$, then $H_L^{md}(n) = O(1)$ and $H_L^{ma}(n) = O(1)$.

Proof. It is clear that $h_L^{ma}(n) \leq h_L^{md}(n)$ for any natural n.

(a) Let $|L| = \infty$, $L \neq \Sigma^*$, and w_0 be a word with the minimum length from $\Sigma^* \setminus L$. Denote by t the length of w_0. Since $|L| = \infty$, $L(n) \neq \emptyset$ for any natural n. Let n be a natural number such that $n > t$ and Γ be a decision tree over $L(n)$ that solves the problem of membership for $L(n)$ nondeterministically and has the minimum depth. Let $w \in L(n)$ and ξ be a complete path in Γ such that $w \in \Sigma(n, \xi)$. Then the terminal node of ξ is labeled with the number 1. Beginning with the first letter, we divide the word w into $\lfloor n/t \rfloor$ blocks with t letters in each and the suffix of the length $n - t\lfloor n/t \rfloor$. Let us assume that the number of nodes labeled with attributes in ξ is less than $\lfloor n/t \rfloor$. Then there is a block such that queries (attributes) attached to nodes of ξ does not ask about letters from the block. We replace this block in the word w with the word w_0 and denote by w' the obtained word. It is clear that $w' \notin L$ and $w' \in \Sigma(n, \xi)$, but this is impossible since the terminal node of the path ξ is labeled with the number 1. Therefore the depth of Γ is greater than or equal to $\lfloor n/t \rfloor$. Thus, $h_L^{ma}(n) \geq \lfloor n/t \rfloor$. It is easy to construct a decision tree over $L(n)$ that solves the problem of membership for $L(n)$ deterministically and has the depth equals to n. Therefore $h_L^{md}(n) \leq n$. Thus, $H_L^{md}(n) = \Theta(n)$ and $H_L^{ma}(n) = \Theta(n)$.

Table 1
Complexity classes $\mathcal{F}_1, \ldots, \mathcal{F}_5$.

	I is independent simple diagram	$cl(I)$	L_I	$H^{rd}_{L_I}$	$H^{ra}_{L_I}$	$H^{md}_{L_I}$	$H^{ma}_{L_I}$
$\mathcal{F}_1$	Yes	$= 0$		$O(1)$	$O(1)$	$O(1)$	$O(1)$
$\mathcal{F}_2$	Yes	$= 1$		$O(1)$	$O(1)$	$\Theta(n)$	$\Theta(n)$
$\mathcal{F}_3$	Yes	≥ 2		$\Theta(\log n)$	$O(1)$	$\Theta(n)$	$\Theta(n)$
$\mathcal{F}_4$	No		$\neq \Sigma^*$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$	$\Theta(n)$
$\mathcal{F}_5$	No		$= \Sigma^*$	$\Theta(n)$	$\Theta(n)$	$O(1)$	$O(1)$

(b) Let $|L| < \infty$. Then there exists natural m such that $L(n) = \emptyset$ for any natural $n \geq m$. Therefore, for each natural $n \geq m$, $h^{md}_L(n) = 0$ and $h^{ma}_L(n) = 0$. Thus, $H^{md}_L(n) = O(1)$ and $H^{ma}_L(n) = O(1)$.

Let $L = \Sigma^*$, n be a natural number, and Γ be a decision tree over $L(n)$, which consists of the root, a terminal node labeled with 1, and an edge that leaves the root and enters the terminal node. One can show that Γ solves the problem of membership for $L(n)$ deterministically and has the depth equals to 0. Therefore $h^{md}_L(n) = 0$ and $h^{ma}_L(n) = 0$. Thus, $H^{md}_L(n) = O(1)$ and $H^{ma}_L(n) = O(1)$. □

4. Corollaries

In this section, we consider two corollaries of Theorems 1, 3, and 4.

4.1. Joint behavior of functions H^{ra}_L, H^{rd}_L, H^{ma}_L, and H^{md}_L

In this section, we assume that each regular factorial language over the alphabet Σ is given by a f-reduced diagram I, which generates the considered language denoted by L_I. To study all possible types of joint behavior of functions $H^{rd}_{L_I}$, $H^{ra}_{L_I}$, $H^{md}_{L_I}$, and $H^{ma}_{L_I}$, we consider five classes of regular factorial languages $\mathcal{F}_1, \ldots, \mathcal{F}_5$ described in the columns 2–4 of Table 1. In particular, $\mathcal{F}_1$ consists of all regular factorial languages L_I for which the diagram I is an independent simple diagram and $cl(I) = 0$. It is easy to show that the complexity classes $\mathcal{F}_1, \ldots, \mathcal{F}_5$ are pairwise disjoint, and each regular factorial language L_I belongs to one of these classes. The behavior of functions $H^{rd}_{L_I}$, $H^{ra}_{L_I}$, $H^{md}_{L_I}$, and $H^{ma}_{L_I}$ for languages from these classes is described in the last four columns of Table 1. For each class, the results considered in Table 1 for the functions $H^{rd}_{L_I}$ and $H^{ra}_{L_I}$ follow directly from Theorems 1 and 3.

We now consider the behavior of the functions $H^{md}_{L_I}$ and $H^{ma}_{L_I}$ for each of the classes $\mathcal{F}_1, \ldots, \mathcal{F}_5$. Let $I = (G, q_0, Q)$ be a f-reduced diagram over the alphabet Σ, which generates a regular factorial language.

Let $L_I \in \mathcal{F}_1$. Since $cl(I) = 0$, G is a directed acyclic graph, and the language L_I is finite. Using Theorem 4 we obtain $H^{md}_{L_I}(n) = O(1)$ and $H^{ma}_{L_I}(n) = O(1)$.

Let $L_I \in \mathcal{F}_2$. Since $cl(I) = 1$, G is a graph containing a cycle, and the language L_I is infinite. By Lemma 4.2 [1], $|L_I(n)| = O(1)$. Therefore $L_I \neq \Sigma^*$. Using Theorem 4 we obtain $H^{md}_{L_I}(n) = \Theta(n)$ and $H^{ma}_{L_I}(n) = \Theta(n)$.

Let $L_I \in \mathcal{F}_3$. Since $cl(I) \geq 2$, G is a graph containing a cycle, and the language L_I is infinite. By Lemma 4.2 [1], $|L_I(n)| = O(n^{cl(I)})$. Therefore $L_I \neq \Sigma^*$. Using Theorem 4 we obtain $H^{md}_{L_I}(n) = \Theta(n)$ and $H^{ma}_{L_I}(n) = \Theta(n)$.

Let $L_I \in \mathcal{F}_4$. Since I is not an independent simple diagram, G is a graph containing a cycle, and the language L_I is infinite. We know that $L_I \neq \Sigma^*$. Using Theorem 4 we obtain $H^{md}_{L_I}(n) = \Theta(n)$ and $H^{ma}_{L_I}(n) = \Theta(n)$.

Let $L_I \in \mathcal{F}_5$. Then $L_I = \Sigma^*$. Using Theorem 4 we obtain $H^{md}_{L_I}(n) = O(1)$ and $H^{ma}_{L_I}(n) = O(1)$.

We now show that the classes $\mathcal{F}_1, \ldots, \mathcal{F}_5$ are nonempty. For simplicity, we assume that $\Sigma = E$, where $E = \{0, 1\}$. It is easy to generalize the considered examples to the case of an arbitrary finite alphabet Σ with at least two letters. In the examples of diagrams, the starting node is labeled with the symbol +, and all nodes are final.

Denote by I_1 the diagram over the alphabet E depicted in Fig. 3. One can show that I_1 is an independent simple f-reduced diagram and $cl(I_1) = 0$. This diagram generates the language $L_{I_1} = \{\lambda, 0\}$, which is factorial. Therefore $L_{I_1} \in \mathcal{F}_1$.

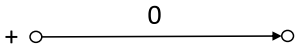

Fig. 3. Diagram I_1.

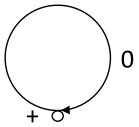

Fig. 4. Diagram I_2.

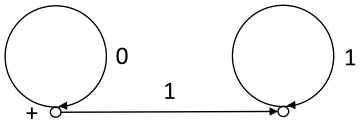

Fig. 5. Diagram I_3.

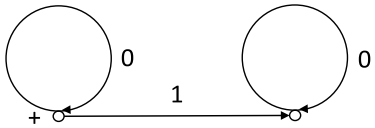

Fig. 6. Diagram I_4.

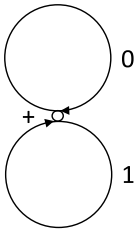

Fig. 7. Diagram I_5.

Denote by I_2 the diagram over the alphabet E depicted in Fig. 4. One can show that I_2 is an independent simple f-reduced diagram and $cl(I_2) = 1$. This diagram generates the language $L_{I_2} = \{0^i : i \in \omega\}$, which is factorial. Therefore $L_{I_2} \in \mathcal{F}_2$.

Denote by I_3 the diagram over the alphabet E depicted in Fig. 5. One can show that I_3 is an independent simple f-reduced diagram and $cl(I_1) = 2$. This diagram generates the language $L_{I_3} = \{0^i 1^j : i, j \in \omega\}$, which is factorial. Therefore $L_{I_3} \in \mathcal{F}_3$.

Denote by I_4 the diagram over the alphabet E depicted in Fig. 6. One can show that I_4 is a dependent simple f-reduced diagram generating the language $L_{I_4} = \{0^i 1^j 0^k : i, k \in \omega, j \in \{0, 1\}\}$, which is factorial. It is clear that $L_{I_4} \neq E^*$. Therefore $L_{I_4} \in \mathcal{F}_4$.

Denote by I_5 the diagram over the alphabet E depicted in Fig. 7. One can show that I_5 is a f-reduced diagram that is not simple. This diagram generates the language $L_{I_5} = E^*$, which is factorial. It is clear that $L_{I_5} = E^*$. Therefore $L_{I_5} \in \mathcal{F}_5$.

A regular factorial language L can have different f-reduced diagrams, which generate it. However, for each of such diagrams I, the language $L_I = L$ will belong to the same complexity class. Let us assume the contrary: there exist a regular factorial language L and two f-reduced diagrams I_1 and I_2, which generate it and for which languages L_{I_1} and L_{I_2} belong to different complexity classes. Then, for some pair $bc \in \{rd, ra, md, ma\}$, the functions $H^{bc}_{L_{I_1}}$ and $H^{bc}_{L_{I_2}}$ have different behavior, but this is impossible since $H^{bc}_{L_{I_1}}(n) = H^{bc}_{L_{I_2}}(n)$ for any natural n.

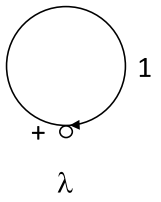

Fig. 8. Diagram $I(0)$.

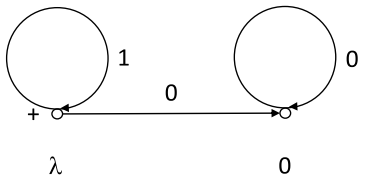

Fig. 9. Diagram $I(01)$.

4.2. Languages over alphabet $\{0,1\}$ given by one forbidden word

Let $E = \{0,1\}$, $\alpha \in E^*$, and $\alpha \neq \lambda$. We denote by $L(\alpha)$ the language over the alphabet E, which consists of all words from E^* that do not contain α as a factor. This is a regular factorial language with $MF(L(\alpha)) = \{\alpha\}$. The following theorem indicates for each nonempty word $\alpha \in E^*$ the complexity class $\mathcal{F}_i$ to which the language $L(\alpha)$ belongs.

Theorem 5. *Let $\alpha \in E^*$ and $\alpha \neq \lambda$.*
(a) If $\alpha \in \{0,1\}$, then $L(\alpha) \in \mathcal{F}_2$.
(b) If $\alpha \in \{01,10\}$, then $L(\alpha) \in \mathcal{F}_3$.
(c) If $\alpha \notin \{0,1,01,10\}$, then $L(\alpha) \in \mathcal{F}_4$.

We now describe a f-reduced diagram $I(\alpha)$ that generates the language $L(\alpha)$ for a nonempty word $\alpha \in E^*$. Let $\alpha = a_1 \cdots a_n$, $\alpha_0 = \lambda$, and $\alpha_i = a_1 \cdots a_i$ for $i = 1, \ldots, n-1$. The set $P(\alpha) = \{\alpha_0, \alpha_1, \ldots, \alpha_{n-1}\}$ is the set of all proper prefixes of the word α. Then $I(\alpha) = (G, q_0, Q)$, where the set of nodes of the graph G is equal to $P(\alpha)$, $q_0 = \alpha_0$, and $Q = P(\alpha)$. For $i = 0, \ldots, n-2$, an edge leaves the node α_i and enters the node α_{i+1}. This edge is labeled with the letter a_{i+1}. For $i = 0, \ldots, n-1$, an edge leaves the node α_i and enters the node $\alpha_j \in P(\alpha)$ such that α_j is the longest suffix of the word $\alpha_i \bar{a}_{i+1}$, where $\bar{a}_{i+1} = 0$ if $a_{i+1} = 1$ and $\bar{a}_{i+1} = 1$ if $a_{i+1} = 0$. This edge is labeled with the letter $\bar{a}_{i+1}$. It is easy to show that $I(\alpha)$ is a f-reduced diagram over the alphabet E. From Theorem 10 [7] it follows that the diagram $I(\alpha)$ generates the language $L(\alpha)$.

Let $\alpha \in E^* \setminus \{\lambda\}$ and $\alpha = a_1 \cdots a_n$. We denote by $\bar{\alpha}$ the word $\bar{a}_1 \cdots \bar{a}_n$. It is easy to prove the following statement.

Lemma 6. *Let $\alpha \in E^*$ and $\alpha \neq \lambda$. Then $H^{bc}_{L(\bar{\alpha})}(n) = H^{bc}_{L(\alpha)}(n)$ for any pair $bc \in \{rd, ra, md, ma\}$ and any natural n.*

Lemma 7. *Let $\alpha \in E^* \setminus \{\lambda\}$, $\beta \in E^*$, and $L(\alpha) \in \mathcal{F}_4$. Then $L(\alpha\beta) \in \mathcal{F}_4$.*

Proof. Since $L(\alpha) \in \mathcal{F}_4$, $H^{rd}_{L(\alpha)}(n) = \Theta(n)$ and $H^{ra}_{L(\alpha)}(n) = \Theta(n)$. One can show that $L(\alpha) \subseteq L(\alpha\beta)$. Using this fact it is not difficult to prove that $H^{rd}_{L(\alpha)}(n) \leq H^{rd}_{L(\alpha\beta)}(n)$ and $H^{ra}_{L(\alpha)}(n) \leq H^{ra}_{L(\alpha\beta)}(n)$ for any natural n. From here and from Theorems 1 and 3 it follows that $H^{rd}_{L(\alpha\beta)}(n) = \Theta(n)$ and $H^{ra}_{L(\alpha\beta)}(n) = \Theta(n)$.

Since $\alpha\beta \notin L(\alpha\beta)$, $L(\alpha\beta) \neq E^*$. The diagram $I(\alpha\beta)$ contains at least one circle formed by the edge that leaves and enters the node λ and is labeled with the letter $\bar{a}_1$, where a_1 is the first letter of the word α. Therefore the language $L(\alpha\beta)$ is infinite. By Theorem 4, $H^{md}_{L(\alpha\beta)}(n) = \Theta(n)$ and $H^{ma}_{L(\alpha\beta)}(n) = \Theta(n)$. Thus, $L(\alpha\beta) \in \mathcal{F}_4$. □

Proof of Theorem 5. In each figure depicting a diagram $I(\alpha)$, $\alpha \in E^* \setminus \{\lambda\}$, we label each node with a corresponding prefix of the word α.

(a) The diagram $I(0)$ is depicted in Fig. 8. This is an independent simple f-reduced diagram with $cl(I(0)) = 1$. Therefore $L(0) \in \mathcal{F}_2$. By Lemma 6, $L(1) \in \mathcal{F}_2$.

(b) The diagram $I(01)$ is depicted in Fig. 9. This is an independent simple f-reduced diagram with $cl(I(01)) = 2$. Therefore $L(01) \in \mathcal{F}_3$. By Lemma 6, $L(10) \in \mathcal{F}_3$.

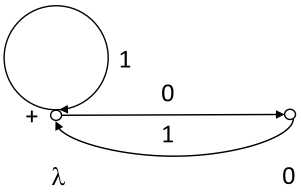

Fig. 10. Diagram $I(00)$.

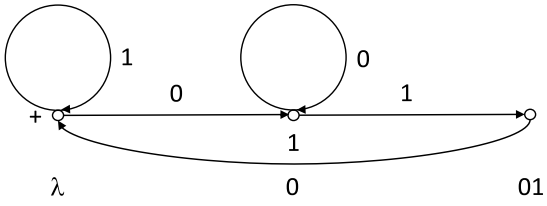

Fig. 11. Diagram $I(010)$.

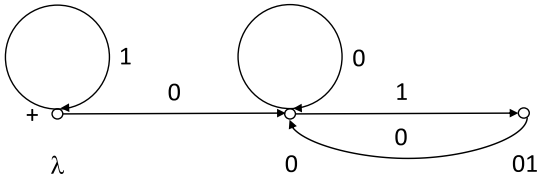

Fig. 12. Diagram $I(011)$.

(c) The diagram $I(00)$ is depicted in Fig. 10. This is not a simple diagram. It is clear that $L(00) \neq E^*$. Therefore $L(00) \in \mathcal{F}_4$. By Lemma 6, $L(11) \in \mathcal{F}_4$. Using Lemma 7 we obtain $L(000), L(001), L(110), L(111) \in \mathcal{F}_4$.

The diagram $I(010)$ is depicted in Fig. 11. This is not a simple diagram. It is clear that $L(010) \neq E^*$. Therefore $L(010) \in \mathcal{F}_4$. By Lemma 6, $L(101) \in \mathcal{F}_4$.

The diagram $I(011)$ is depicted in Fig. 12. This is not a simple diagram. It is clear that $L(011) \neq E^*$. Therefore $L(011) \in \mathcal{F}_4$. By Lemma 6, $L(100) \in \mathcal{F}_4$.

We proved that, for any word $\alpha \in E^*$ of the length three, $L(\alpha) \in \mathcal{F}_4$. Using Lemma 7 we obtain that, for any word $\alpha \in E^*$ of the length greater than or equal to four, $L(\alpha) \in \mathcal{F}_4$. □

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Research reported in this publication was supported by King Abdullah University of Science and Technology (KAUST), Saudi Arabia. The

author is grateful to the anonymous reviewers for useful remarks and suggestions.

References

[1] Moshkov M. Decision trees for regular language word recognition. Fund Inform 2000;41(4):449–61.
[2] Moshkov M. Complexity of deterministic and nondeterministic decision trees for regular language word recognition. In: Bozapalidis S, editor. Proceedings of the 3rd international conference developments in language theory, DLT 1997, Thessaloniki, Greece, July 20–23, 1997. Aristotle University of Thessaloniki; 1997, p. 343–9.
[3] D'Alessandro F, Intrigila B, Varricchio S. On the structure of the counting function of sparse context-free languages. Theoret Comput Sci 2006;356(1–2):104–17.
[4] Shur AM. Combinatorial complexity of regular languages. In: Hirsch EA, Razborov AA, Semenov AL, Slissenko A, editors. Computer science - theory and applications, third international computer science symposium in Russia, CSR 2008, Moscow, Russia, June 7-12, 2008, proceedings. Lecture notes in computer science, vol. 5010, Springer; 2008, p. 289–301.
[5] Moshkov M. Decision trees for binary subword-closed languages. 2022, CoRR, arXiv:2201.01493.
[6] Haines LH. On free monoids partially ordered by embedding. J Combin Theory 1969;6:94–8.
[7] Crochemore M, Mignosi F, Restivo A. Automata and forbidden words. Inf Process Lett 1998;67(3):111–7.
[8] Yu S. Regular languages. In: Rozenberg G, Salomaa A, editors. Handbook of formal languages, Volume 1: word, language, grammar. Springer; 1997, p. 41–110.
[9] Hopcroft JE, Motwani R, Ullman JD. Introduction to automata theory, languages, and computation. Pearson international edition, third ed.. Addison-Wesley; 2007.