State Complexity of Projected Languages

Galina Jirásková [1,⋆] and Tomáš Masopust [2,3,⋆⋆]

[1] Mathematical Institute, Slovak Academy of Sciences
Grešákova 6, 040 01 Košice, Slovak Republic
jiraskov@saske.sk
[2] CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
[3] Institute of Mathematics, Czech Academy of Sciences
Žižkova 22, 616 62 Brno, Czech Republic
masopust@math.cas.cz

Abstract. This paper discusses the state complexity of projected regular languages represented by incomplete deterministic finite automata. It is shown that the known upper bound is reachable only by automata with one unobservable transition, that is, a transition labeled with a symbol removed by the projection. The present paper improves this upper bound by considering the structure of the automaton. It also proves that the new bounds are tight, considers the case of finite languages, and presents several open problems.

1 Introduction

Projections, also called natural projections since they can be seen as natural transformations of category theory, or abstractions, play an important role in many fields of computer science and engineering, such as verification, diagnoses, and supervisory control [1, 16–18, 30]. Given a regular language L and a projection P, it is well-known that the minimal deterministic finite automaton (dfa) accepting language $P(L)$ can be of exponential size in comparison with the dfa accepting language L. The known upper bound for projection is $3 \cdot 2^{n-2} - 1$ [29]. On the other hand, however, this result does not consider the structure of the automaton, which is of interest because, as shown in this paper, this upper bound is reachable only for automata with one *unobservable transition*, that is, a transition that is labeled with a symbol removed by the projection. Note that several unobservable transitions connecting the same two states in the same direction (called unobservable multi-transitions) are considered as only one unobservable transition, that is, we disregard unobservable multi-transitions.

In this paper, we improve the upper bound by considering the structure of the automaton. Specifically, we study the state complexity with respect to the structure of unobservable transitions. This parameter turns out to be more

⋆ Research supported by the Slovak Research and Development Agency under contract APVV-0035-10 "Algorithms, Automata, and Discrete Data Structures".
⋆⋆ Research supported by the European Community's 7^{th} Framework Programme grant no. INFSO-ICT-224498, and by the GAČR grant no. 202/11/P028.

convenient than the number of unobservable transitions. We show that, given a projection and a minimal incomplete dfa with n states, the minimal incomplete dfa accepting the projected language has no more than $2^{n-1}+2^{n-m}-1$ states, where m is the number of states incident with unobservable transitions. This bound is reachable if the number of unobservable transitions is $m-1$. However, any additional unobservable transition can introduce a new unreachable subset, which means that the bound is not tight if there are more than $m-1$ unobservable transitions. Therefore, we also discuss the case the automaton has at least m unobservable transitions, and show that in this case the tight upper bound is $3 \cdot 2^{n-3}+2^{n-m}-1$.

The paper also discusses the case of projected finite languages, and shows that the upper bounds on the number of states correspond to the upper bounds on the nfa to dfa conversion [26].

For several operations, $op(\cdot)$, such as the determinization of nfa's, it has been shown that for all integers n and α with $f(n) \leq \alpha \leq g(n)$, where $f(n)$ and $g(n)$ are the tight lower and upper bounds for $op(\cdot)$, there exists a regular language L represented by a minimal dfa of size n such that the minimal dfa for $op(L)$ is of size α. A number α for which no such language exists is called *magic* for n with respect to $op(\cdot)$. For instance, there are no magic numbers for the determinization of nfa's with the input alphabet of cardinality at least three, where $f(n)=n$ and $g(n)=2^n$. During the last few years, this topic has widely been discussed in the literature. The reader is referred to [6, 8, 10–13, 15, 28] for more information. Our last theorem solves the magic number problem for projections using the result on magic numbers for stars of regular languages [14].

We conclude the paper with a short overview of open problems concerning projected regular languages.

2 Preliminaries and Definitions

We assume that the reader is familiar with automata theory, and for all unexplained notions, we refer the reader to [27, 31].

For an alphabet (finite nonempty set) Σ, denote by Σ^* the set of all finite strings over the alphabet Σ including the empty string ε. A *language* over Σ is any subset of Σ^*. A language L is *finite* if L is a finite set.

Let $\Sigma_o \subseteq \Sigma$. A homomorphism $P : \Sigma^* \to \Sigma_o^*$ is called the *(natural) projection* if it is defined so that $P(a)=\varepsilon$ if $a \in \Sigma \setminus \Sigma_o$, and $P(a)=a$ if $a \in \Sigma_o$.

An *(incomplete) dfa* is a quintuple $A=(Q,\Sigma,\delta,s,F)$, where Q is a finite set of *states*, Σ is an *input alphabet*, $\delta : Q \times \Sigma \to Q$ is a *(partial) transition function*, $s \in Q$ is the *initial state*, and $F \subseteq Q$ is the set of *final states*. In the usual way, transition function δ can be extended to the domain $Q \times \Sigma^*$. The language *accepted* by A is defined as the set $L(A)=\{w \in \Sigma^* \mid \delta(s,w) \in F\}$. A transition $\delta(p,a)=q$ is said to be *unobservable* with respect to P if $a \in \Sigma \setminus \Sigma_o$, that is, if $P(a)=\varepsilon$.

For a regular language L, we denote by $\|L\|$ the smallest number of states in any incomplete dfa accepting L.

In comparison with complete dfa's, each incomplete dfa represents two languages. The language accepted by the dfa as defined above, also called a marked language, and the language of all strings that the dfa can read called a generated language, that is, the strings for which the corresponding transitions are defined. For complete dfa's, the latter language is equal to Σ^*.

Considering complete automata, the corresponding upper bounds can be derived from the results for incomplete automata by considering only those unobservable transitions that are not incident with the *dead* or *sink* state. For this reason, we only discuss the case of incomplete dfa's in this paper.

3 DFAs as Graphs

Here we concentrate our attention on the number of states potentially reachable in the subset automaton constructed from a given dfa after applying a projection. For simplification, we consider the important parts of automata as graphs.

A *directed graph* is a pair $G = (V, E)$, where V is a finite set of nodes, and $E \subseteq V \times V$ is a set of edges. An edge $(u, v) \in E$ is called a *loop* if $u = v$. Let $u \in V$ be a node, then *in-degree* and *out-degree* of v are the sizes of sets $\{u \in V \mid (u, v) \in E\}$ and $\{w \in V \mid (v, w) \in E\}$, respectively. A node with in-degree 0 and out-degree 1, or with in-degree 1 and out-degree 0 is called a *leaf*. This definition requires that the node is incident to an edge. Thus, a node incident to no edge is not considered to be a leaf.

A *path* is a sequence of nodes $v_0, v_1, \ldots, v_k$ such that $v_i \neq v_j$ if $i \neq j$, and (v_i, v_{i+1}) is an edge in E for $i = 0, 1, \ldots, k-1$. A *non-oriented path* is a sequence $v_0, v_1, \ldots, v_k$ such that $v_i \neq v_j$ if $i \neq j$, and either (v_i, v_{i+1}) or (v_{i+1}, v_i) is an edge in E for $i = 0, 1, \ldots, k-1$. A graph G is *connected* if for all nodes u, v in V, there is a non-oriented path from u to v. For a node v in V, let $G \setminus \{v\}$ denote the graph constructed from G by removing node v and all edges incident to v.

A subset X of V is said to be *bad* in graph $G = (V, E)$ if there exists an edge (u, v) in E such that $u \in X$ and $v \notin X$. A set is said to be *good* if it is not bad; thus a good subset of V is closed under outgoing transitions. Let $b(G)$ denote the number of bad subsets in G, and $g(G)$ the number of good subsets in G. We first study the number of bad subsets in a graph.

Lemma 1. *Let $m, n \geq 2$ and let $G = (V, E)$ be a directed graph without loops with n nodes. Let $U = \{u, v \in V \mid (u, v) \in E\}$ and assume that U is of size m. Then $b(G) \geq (2^{m-1} - 1)\, 2^{n-m}$.*

Proof. Let G and U be as assumed in the theorem, and consider a special case where the edges involved in nodes of U go only from $m - 1$ different nodes to the last m-th node. This means that there exists a node v in V such that for each node u in $U \setminus \{v\}$, the edge (u, v) is in E, while for each node z in V, the edge (z, u) is not in E. Then there are $2^{m-1} - 1$ nonempty subsets of U which do not contain node v, and so are bad. This gives $b(G) \geq (2^{m-1} - 1)\, 2^{n-m}$.

Now, we will show the theorem to be true in general, and not just under the assumption that the edges in U go only from $m - 1$ different nodes to the last m-th node as was done in the paragraph above. The proof is by induction on m.

If $m = 2$, then U involves either one or two edges. Note first that if X is a bad subset in G, then X is bad after addition of any number of edges to G. Thus, we can consider that there is only one edge because the other one cannot decrease the number of bad subsets. Then, if we have one edge, say (a, b), we can have a along with any combination of elements of $V \setminus \{a, b\}$ in a bad subset, and thus we have $b(G) \geq 2^{n-2} = (2^{2-1} - 1)\, 2^{n-2}$. Assume that the statement holds for all sets U of size less than m, and consider the case U is of size m. There are two possibilities. Either the number of edges is strictly less than m, or it is greater then or equal to m. In the former case, consider the number of edges and denote it by t, and in the latter case, consider the subset of edges of size t forming the minimal spanning tree (forest). Thus $t < m$ and there is a leaf v in U such that v is connected with a node u in $U \setminus \{v\}$. Then, either (i) all nodes in $U \setminus \{v\}$ are incident with some of the t edges, or (ii) node u was connected only with v and now it is not incident with any other node in $U \setminus \{v\}$.

In case (i), the set $U \setminus \{v\}$ is of size $m - 1$, and by the induction hypothesis, there are at least $2^{m-2} - 1$ bad subsets of $U \setminus \{v\}$. If $(v, u) \in E$, then for each subset A of $U \setminus \{v\}$ that is bad in $U \setminus \{v\}$, the sets A and $A \cup \{v\}$ are bad in U, and $\{v\}$ is a new bad set. This gives $b(G) \geq (2^{m-2} - 1 + 2^{m-2} - 1 + 1)\, 2^{n-m}$. Similarly, if $(u, v) \in E$, then for each subset A of $U \setminus \{v\}$ that is bad in $U \setminus \{v\}$, the sets A, $A \cup \{v\}$ are bad in U, and the set $U \setminus \{v\}$ is a new bad set.

In case (ii), the set $U \setminus \{u, v\}$ is of size $m-2$, and so, there are at least $2^{m-3}-1$ bad subsets of $U \setminus \{u, v\}$. We now have $m \geq 4$. The sets $\emptyset$ and $U \setminus \{u, v\}$ are not bad. Thus $\{v\}$ or $\{u\}$, and $U \setminus \{u\}$ or $U \setminus \{v\}$, depending on the direction of the edge connecting u and v, are two new bad subsets. Moreover, all bad subsets of $U \setminus \{u, v\}$ are also bad in U. If there is at least one more proper non-empty good subset B of $U \setminus \{u, v\}$, then $B \cup \{u\}$ or $B \cup \{v\}$ is the third new bad subset of U. Summarized, this gives $b(G) \geq (2^2\,(2^{m-3} - 1) + 3)\, 2^{n-m} = (2^{m-1} - 1)\, 2^{n-m}$. If there are only two good subsets of $U \setminus \{u, v\}$, namely $\emptyset$ and $U \setminus \{u, v\}$, then the number of bad subsets of $U \setminus \{u, v\}$ is $2^{m-2} - 2$, which, since $m \geq 4$, gives $b(G) \geq 2^2(2^{m-2}-2)\, 2^{n-m} = (2^{m-1}-1+2^{m-1}-7)\, 2^{n-m} \geq (2^{m-1}-1)\, 2^{n-m}$. □

Consider the statement of Lemma 1. Then the number of all the subsets of $V \setminus U$ is 2^{n-m} while the number of bad subsets of U is $2^{m-1} - 1$. Moreover, there is a graph $G = (V, E)$ with U of size $|E| - 1$, for which the equality holds. However, if $m \leq |E|$, each additional transition can introduce a new bad subset. This problem is discussed in the following result.

Lemma 2. *Let $m, n \geq 2$ and let $G = (V, E)$ be a directed graph without loops with n nodes. Let $U = \{u, v \in V \mid (u, v) \in E\}$ and assume that $|U| = m \leq |E|$. Then $b(G) \geq (5 \cdot 2^{m-3} - 1)\, 2^{n-m}$.*

Proof. The proof is by induction on m. If $m = 2$, then the graph consists of two nodes connected by two edges. This gives two bad subsets of U, which results in $b(G) = 2 \cdot 2^{n-m} \geq 3/2 \cdot 2^{n-m}$. Assume that the statement holds for all sets U of cardinality less then m, and consider the case U is of cardinality m. Recall that $m \leq |E|$. Consider a subset of m edges forming a minimal spanning tree (forest). Then there is a leaf v in U. If $|U \setminus \{v\}| \leq |E(G \setminus \{v\})|$ then by the induction

hypothesis, the set $U \setminus \{v\}$ has at least $5 \cdot 2^{m-4} - 1$ bad subsets. Otherwise, by Lemma 1, the set $U \setminus \{v\}$ has at least $2^{m-2} - 1$ bad subsets.

In the former case, if $(v, u) \in E$, then for each bad subset A of $U \setminus \{v\}$, the set $A \cup \{v\}$ is a new bad subset of U and, in addition, $\{v\}$ is a new bad subset of U. If $(u, v) \in E$, then for each bad subset A of $U \setminus \{v\}$, the set $A \cup \{v\}$ is a new bad subset of U and, in addition, the set $U \setminus \{v\}$ is a new bad subset of set U. Thus $b(G) \geq (5 \cdot 2^{m-4} - 1 + 5 \cdot 2^{m-4})\, 2^{n-m} = (5 \cdot 2^{m-3} - 1)\, 2^{n-m}$.

In the latter case, notice that there are at least two edges connecting v and $U \setminus \{v\}$ in G. We have three possibilities:

(i) Node v is connected with $U \setminus \{v\}$ by edges (v, u_1) and (v, u_2) with $u_1 \neq u_2$. Then the sets $A \cup \{v\}$, $A \cup \{v, u_1\}$, and $A \cup \{v, u_2\}$ are bad in U for every subset A of $U \setminus \{v, u_1, u_2\}$. Hence we have at least $3 \cdot 2^{m-3}$ new bad subsets in U.

(ii) Node v is connected with $U \setminus \{v\}$ by edges (u_1, v) and (u_2, v). Then for each subset A of $U \setminus \{u_1, u_2, v\}$, if $A \cup \{u_1\}$ is bad in $U \setminus \{v\}$, then $A \cup \{v, u_1\}$ is bad in U, otherwise $A \cup \{u_1\}$ is bad in U; if $A \cup \{u_2\}$ is bad in $U \setminus \{v\}$, then $A \cup \{v, u_2\}$ is bad in U, otherwise $A \cup \{u_2\}$ is bad in U; if $A \cup \{u_1, u_2\}$ is bad in $U \setminus \{v\}$, then $A \cup \{u_1, u_2, v\}$ is bad in U, otherwise $A \cup \{u_1, u_2\}$ is bad in U. Summarized, there are $3 \cdot 2^{m-3}$ new bad subsets in U.

(iii) Node v is connected with $U \setminus \{v\}$ by edges (u_1, v) and (v, u_2). Then the sets $A \cup \{v\}$ and $A \cup \{u_1, v\}$ are bad in U for each subset A of $U \setminus \{u_1, u_2, v\}$. In addition, if $A \cup \{u_1, u_2\}$ is bad in $U \setminus \{v\}$, then the set $A \cup \{u_1, u_2, v\}$ is a new bad subset of U. Otherwise, the set $A \cup \{u_1, u_2\}$ is a new bad subset of U. Thus there are at least $3 \cdot 2^{m-3}$ new bad subsets of U.

This gives $b(G) \geq (2^{m-2} - 1 + 3 \cdot 2^{m-3})\, 2^{n-m} = (5 \cdot 2^{m-3} - 1)\, 2^{n-m}$. □

4 State Complexity of Projected Regular Languages

Recall that it is shown in [29] that the worst-case tight upper bound on projected regular languages is $2^{n-1} + 2^{n-2} - 1$, where n is the number of states of the minimal incomplete dfa recognizing the given language.

Theorem 1 ([29]). *Let $n \geq 2$ and L be a regular language over Σ with $\|L\| = n$. Let $\Sigma_o \subseteq \Sigma$ and P be the projection of Σ^* onto Σ_o^*. The tight upper bound on the size of the minimal incomplete dfa for projected language $P(L)$ is $3 \cdot 2^{n-2} - 1$.*

In what follows, we improve the upper bound by taking into account the structure of nonloop unobservable transitions. More specifically, we consider the number of states that are incident with nonloop unobservable transitions. Note that it follows from the results that the previous bound is reachable only by dfa's with one unobservable transition, up to unobservable multi-transitions.

Theorem 2. *Let $m, n \geq 2$, $\Sigma_o \subseteq \Sigma$, and P be the projection of Σ^* onto Σ_o^*. Let L be a regular language over alphabet Σ with $\|L\| = n$, and $(Q, \Sigma, \delta, s, F)$ be the minimal incomplete dfa recognizing language L, in which*

$$|\{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}| = m.$$

Then $\|P(L)\| \leq 2^{n-1} + 2^{n-m} - 1$.

Proof. Consider the minimal incomplete dfa $(Q, \Sigma, \delta, s, F)$ accepting L, and construct a directed graph $G = (Q, E)$ without loops so that E contains an edge (p, q) in $Q \times Q$ if and only if $p \neq q$ and there is a transition $\delta(p, a) = q$ for some unobservable symbol a in $\Sigma \setminus \Sigma_o$. Construct an nfa for language $P(L)$ from dfa A by replacing all the unobservable transitions with ε-transitions. Observe that each subset of Q that contains p, but not q, is not reachable in the corresponding subset automaton because every string leading the nfa to state p also leads the automaton to state q. This means that no subset of Q that is bad in graph G is reachable. By Lemma 1, for the number $g(G)$ of good subsets (that is, subsets closed under outgoing transitions) we have $g(G) = 2^n - b(G) \leq 2^n - (2^{m-1} - 1)\, 2^{n-m} = 2^{n-1} + 2^{n-m}$. Good subsets of Q in graph G correspond to potentially reachable states in the subset automaton. This number is decreased by one because the empty set (the dead state) is potentially reachable but it is not present in the minimal incomplete dfa. □

Notice that Theorem 1 is a consequence of Theorem 2 since $\|P(L)\|$ is maximal if $m = 2$. The next result shows that the bound $2^{n-1} + 2^{n-m} - 1$ is tight.

Theorem 3. *Let $m, n \geq 2$ and P be the projection of $\{a, b, c\}^*$ onto $\{a, b\}^*$. There exists a regular language L over $\{a, b, c\}$ with $\|L\| = n$, such that the minimal incomplete dfa accepting L has $m - 1$ unobservable nonloop transitions connecting m states, and $\|P(L)\| = 2^{n-1} + 2^{n-m} - 1$.*

Proof. Let L be the language over $\{a, b, c\}$ accepted by the incomplete dfa shown in Fig. 1. After applying the projection onto $\{a, b\}$ and removing ε-transitions, we get the n-state nfa shown in Fig. 2. The nfa accepts the string b^n only from state $n - 1$, and the string $a^i b^n$ only from state $n - 1 - i$ $(0 \leq i \leq n - 1)$. It follows that the states in the corresponding subset automaton are pairwise distinguishable. To prove the theorem, we only need to show that the subset automaton has $2^{n-1} + 2^{n-m} - 1$ reachable non-empty states.

We first prove by induction that every subset of $\{0, 1, \ldots, n - 1\}$ containing state 0 is reachable. The initial state $\{0\}$ goes to state $\{n - m\}$ by a^{n-m}, then by a string in b^* to states $\{0, i\}$ with $n - m + 1 \leq i \leq n - 2$. State $\{0, n - 2\}$ goes to state $\{0, 1, n - 1\}$ by a, and then by a string in b^* to states $\{0, i, n - 1\}$ with $1 \leq i \leq n - 2$. State $\{0, n - 2, n - 1\}$ goes to $\{0, n - 1\}$ by b, and then to $\{0, 1\}$ by a. By a string in b^*, state $\{0, 1\}$ goes to states $\{0, i\}$ with $1 \leq i \leq n - m$. Thus each

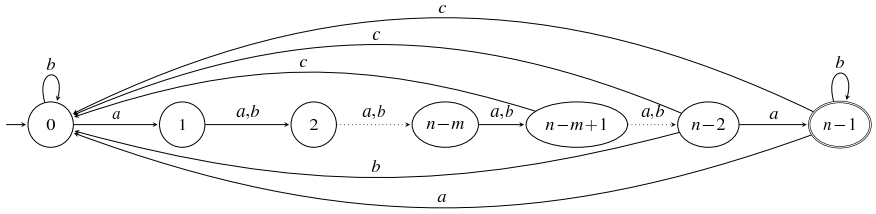

Fig. 1. The minimal incomplete dfa for a language L with $\|P(L)\| = 2^{n-1} + 2^{n-m} - 1$.

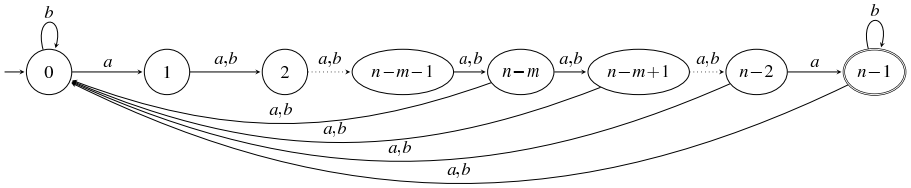

Fig. 2. An nfa accepting the projection of the language from Fig. 1.

subset of size 2 containing state 0 is reachable. Now let $X = \{0, i_1, i_2, \ldots, i_t\}$ be a set of size $t+1$, where $2 \leq t \leq n-1$ and $1 \leq i_1 < i_2 < \cdots < i_t \leq n-1$. Consider two cases:

(*i*) $i_t = n-1$. Then X is reached from $\{0, i_2 - i_1, \ldots, i_{t-1} - i_1, n-2\}$ by ab^{i_1-1}, and the latter set of size t is reachable by the induction hypothesis.
(*ii*) $i_t < n-1$. Then X is reached from $\{0, i_2 - i_1, \ldots, i_t - i_1, n-1\}$ by ab^{i_1-1}, and the latter set of size $t+1$ contains state $n-1$, and is reachable by (*i*).

This proves reachability of all subsets containing state 0. Next, if $\{i_1, i_2, \ldots, i_t\}$ is a non-empty subset of the set $\{1, 2, \ldots, n-m\}$, then it is reached from the set $\{0, i_2 - i_1, i_3 - i_1, \ldots, i_t - i_1\}$ containing state 0 by a^{i_1}. This gives $2^{n-1} + 2^{n-m} - 1$ reachable non-empty states, and completes our proof. □

In the theorems above, the number of unobservable transitions is considered to be less than the size of the set $\{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}$. However, an additional unobservable transition may introduce a new unreachable subset. The following example shows that if the size of this set is less than or equal to the number of unobservable nonloop transitions, then the upper bound is not tight. The precise upper bound for this case is open.

Example 1. Let $m, n \geq 2$. Consider a minimal incomplete dfa $(Q, \Sigma, \delta, s, F)$ of n states. Let the incomplete automaton have at least m unobservable transitions. Let $U = \{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}$ and assume that $|U| = m$. Construct a directed graph $G = (Q, E)$ without loops so that the set E contains an edge (p, q) in $Q \times Q$ if and only if $p \neq q$ and there is a transition $\delta(p, a) = q$ for some unobservable symbol a in $\Sigma \setminus \Sigma_o$.

In the case of $m = 2$, there must be a cycle of length two in G. In this case, however, we have $g(G) = 2^n - 2 \cdot 2^{n-2} = 2^{n-1}$.

In the case of $m = 3$, there are three possibilities: (i) if U contains a cycle of length three, then there are at least 6 subsets that are bad for U because all but the empty set and the whole set U are bad; (ii) if U contains a cycle with one transition reversed, then there are at least 4 bad subsets of U; (iii) if U contains a cycle of length two and an edge to (or from) the third node, then there are at least 5 bad subsets of U. In all three cases, we get $g(G) \leq 2^n - 4 \cdot 2^{n-3} = 2^{n-1}$.

Since only non-empty good subsets for G can be reached in the incomplete dfa for the projected language, we get the bound $2^{n-1} - 1$ on the size of this dfa in both cases. This is strictly less than $2^{n-1} + 2^{n-m} - 1$ given by Theorem 2. □

Finally, the situation is significantly different for projections of regular languages with one-letter co-domains.

Theorem 4. *Let a be a symbol in an alphabet Σ and P be the projection of strings in Σ^* to strings in a^*. Let L be a regular language over Σ with $\|L\| = n$. Then $\|P(L)\| \leq e^{(1+o(1))\sqrt{n \ln n}}$.*

Proof. Replace all the transitions unobservable for projection P in the minimal incomplete dfa recognizing language L with ε-transitions to get an n-state unary nfa for language $P(L)$. This unary nfa can be simulated by a dfa with no more than $e^{(1+o(1))\sqrt{n \ln n}}$ states [2, 6, 20], and the upper bound follows. □

The following theorem discusses a special case that gives an idea how to treat the cases with more and more unobservable transitions.

Theorem 5. *Let $m, n \geq 2$ and $\Sigma_o \subseteq \Sigma$. Let P be the projection of strings in Σ^* to strings in Σ_o^*. Let L be a regular language over alphabet Σ with $\|L\| = n$, and $(Q, \Sigma, \delta, s, F)$ be the minimal incomplete dfa recognizing language L, in which $|\{p, q \in Q \mid p \neq q \text{ and } q \in \delta(p, \Sigma \setminus \Sigma_o)\}| = m$. If at least m transitions in the dfa are unobservable for the projection, then $\|P(L)\| \leq 2^{n-2} + 2^{n-3} + 2^{n-m} - 1$.*

Proof. Consider the minimal incomplete dfa $(Q, \Sigma, \delta, s, F)$ for L, and construct a directed graph $G = (Q, E)$ without loops so that E contains an arc (p, q) if and only if $p \neq q$ and there is a transition $\delta(p, a) = q$ for some unobservable symbol a in $\Sigma \setminus \Sigma_o$. Construct an nfa for language $P(L)$ from the dfa for L by replacing all the unobservable transitions with ε-transitions. Then every subset that is reachable in the corresponding subset automaton must be good for G. By Lemma 2, we have $g(G) \leq 2^n - (5 \cdot 2^{m-3} - 1)\, 2^{n-m} = 2^{n-2} + 2^{n-3} + 2^{n-m}$. This number is decreased by one because of the empty set (the dead state). □

The next result proves the tightness of the bound $2^{n-2} + 2^{n-3} + 2^{n-m} - 1$ in the case of a four-letter domain alphabet.

Theorem 6. *Let $n \geq 2$ and P be the projection of $\{a, b, c, d\}^*$ onto $\{a, b, c\}^*$. There exists a regular language L over $\{a, b, c, d\}$ with $\|L\| = n$ such that the minimal incomplete dfa accepting L has m unobservable nonloop transitions on no more than m states, and $\|P(L)\| = 2^{n-2} + 2^{n-3} + 2^{n-m} - 1$.*

Proof. Consider the language L over the alphabet $\{a, b, c, d\}$ accepted by the incomplete n-state dfa shown in Fig. 3. Construct an nfa for language $P(L)$ from the dfa for L by replacing all the unobservable transitions with ε-transitions. After removing the ε-transitions, we get the n-state nfa for $P(L)$ shown in Fig. 4.

Notice that this nfa accepts string a^{n-i} with $2 \leq i \leq n$ only from state i, and string ca^{n-2} only from state 1. It follows that all the states in the corresponding subset automaton are pairwise distinguishable. Thus it is enough to show that the subset automaton has $2^{n-2} + 2^{n-3} + 2^{n-m}$ reachable states including the empty set.

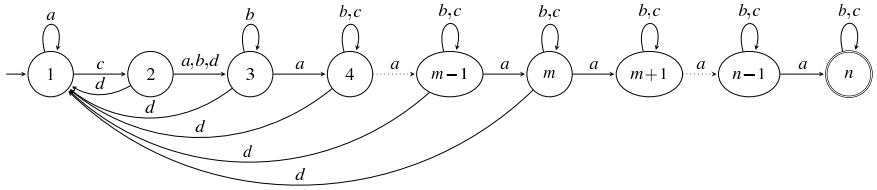

Fig. 3. The incomplete dfa over $\{a, b, c, d\}$ with m unobservable transitions on m states meeting the bound $2^{n-2} + 2^{n-3} + 2^{n-m} - 1$ on the projection onto $\{a, b, c\}$.

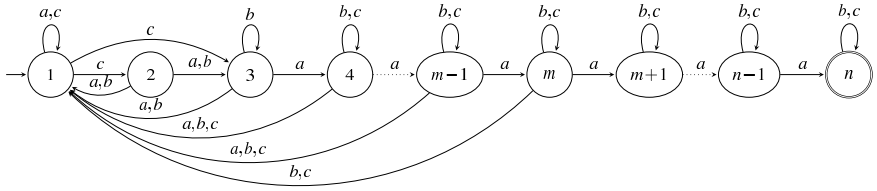

Fig. 4. The nfa for the projection of the language from Fig. 3.

State $\{1\}$ is the start state of the subset automaton. Each set $\{1, i_1, i_2, \ldots, i_t\}$ of size $t + 1$, where $3 \leq i_1 < i_2 < \cdots < i_t \leq n$ and $1 \leq t \leq n - 1$, is reached from the set $\{1, i_2 - (i_1 - 3), \ldots, i_t - (i_1 - 3)\}$ of size t by string cba^{i_1-3}. Thus, by induction, each state $\{1\} \cup X$ with $X \subseteq \{3, 4, \ldots, n\}$ is reachable. Next, such a state $\{1\} \cup X$ goes to state $\{1, 2, 3\} \cup X$ by c. Finally, if X is a subset of $\{m + 1, m + 2, \ldots, n\}$, then state $\{1\} \cup X$ goes to state X by b. This proves the reachability of the desired number of states, and concludes our proof. □

5 State Complexity of Projected Finite Languages

In this section, we consider the state complexity of projected finite languages. First, let us consider the case of projections with co-domains of size one.

Proposition 1. *Let a be a symbol in an alphabet Σ and let P be the projection of Σ^* onto a^*. If L is a finite regular language over Σ, then $\|P(L)\| \leq \|L\|$.*

Proof. Consider the minimal complete dfa with n states accepting language L. Since L is finite, there must exist a string that leads the dfa to the dead state. Hence the minimal incomplete dfa accepting L has $n-1$ states. After replacing all the unobservable transitions with ε-transitions and eliminating ε-transitions, the resulting nfa with $n - 1$ states accepts finite language $P(L)$. Therefore, this nfa can be simulated by an n-state complete dfa [26]. Again, some string must lead this complete dfa to the dead state, which implies that the minimal incomplete dfa accepting $P(L)$ has at most $n - 1$ states. Thus $\|P(L)\| \leq \|L\|$. □

The following theorem deals with finite languages and binary co-domain alphabets.

Theorem 7. *Let a and b be symbols in an alphabet Σ and P be the projection of Σ^* onto $\{a,b\}^*$. Let L be a finite language over Σ with $\|L\| = n$. Then*

$$\|P(L)\| \leq \begin{cases} 2 \cdot 2^{\lfloor n/2 \rfloor} - 2 & \text{if } n \text{ is even,} \\ 3 \cdot 2^{\lfloor n/2 \rfloor} - 2 & \text{if } n \text{ is odd.} \end{cases}$$

In addition, the bound is tight in the case of a ternary domain alphabet.

Proof. We first prove the upper bound. Consider an incomplete dfa accepting language L, and construct an n-state nfa for $P(L)$ by replacing all the unobservable transitions with ε-transitions, and eliminating the ε-transitions. The n-state nfa for finite language $P(L)$ can be simulated by a complete dfa of $2^{n/2+1} - 1$ states if n is even, or of $3 \cdot 2^{\lfloor n/2 \rfloor} - 1$ states if n is odd [26]. Since some string must lead this complete dfa to the dead state, this state is removed from the minimal incomplete dfa representation of $P(L)$.

For tightness, consider the ternary finite regular language recognized by the incomplete dfa shown in Fig. 5, where $k = \lceil n/2 \rceil - 1$. The application of the projection P results in the language

$$P(L) = \bigcup_{i=0}^{\lceil n/2 \rceil - 1} (a+b)^i a (a+b)^{\lfloor n/2 \rfloor - 1}$$

that can be written as $P(L) = \{uav \in \{a,b\}^* \mid |uav| < n \text{ and } |v| = \lfloor n/2 \rfloor - 1\}$. However, the minimal complete dfa accepting $P(L)$ has $2^{n/2+1} - 1$ states if n is even, or $3 \cdot 2^{\lfloor n/2 \rfloor} - 1$ states if n is odd, as shown in [26]. Since $P(L)$ is finite, the minimal incomplete dfa for $P(L)$ has one less state than the complete dfa. Hence the bounds are tight. □

In the next theorem, we consider the case of projections of finite languages with co-domains of size k with $k \geq 2$. In comparison with the previous result, where the sizes of the domain and co-domain differ by one, note that the size of the domain of the projection is required to be of linear size with respect to the number of states. It remains open if it can be limited by a constant.

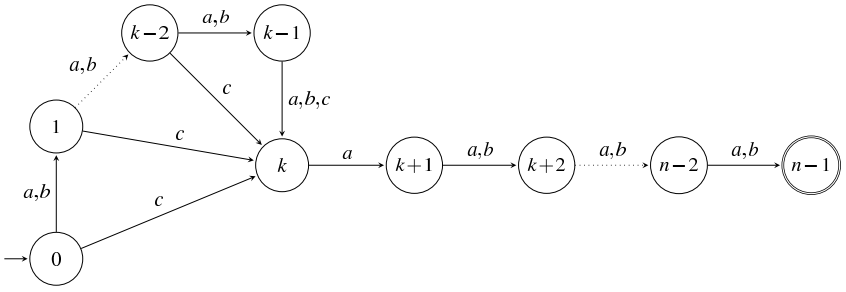

Fig. 5. The minimal incomplete dfa over $\{a,b,c\}$ accepting a finite language meeting the upper bound on the projection onto $\{a,b\}$; $k = \lceil n/2 \rceil - 1$.

Theorem 8. *Let $k, n \geq 2$. There exist alphabets Σ and Σ_o with $\Sigma_o \subseteq \Sigma$ and $|\Sigma_o| = k$, and a finite language L over Σ with $\|L\| = n$ such that*

$$\|P(L)\| = (k^{\lfloor n/(\log k+1)\rfloor+1} - 1)/(k-1) - 1\,,$$

where P is the projection of strings in Σ^* onto strings in Σ_o^*. In addition, the upper bound is $(k^{\lceil (n/(\log k+1)\rceil+1} - 1)/(k-1) - 1$.

Proof. The upper bound follows from [26, Theorem 5] in a similar way as shown in the proof of Theorem 7. To prove the lower bound, let $t = \lceil \log k \rceil$ and let $m = \lfloor n/(t+1) \rfloor$. Let $\Sigma_o = \{0, 1, \ldots, k-1\}$, let $\Sigma = \{a_1, a_2, \ldots, a_{n-m-1}\} \cup \Sigma_o$, and let P be the projection of Σ^* onto Σ_o^*.

Set $S_i = \{j \in \Sigma_o \mid j \bmod 2^i \geq 2^{i-1}\}$ for $i = 1, 2, \ldots, t$. Notice that a symbol j is in S_i if and only if the i-th digit from the end in the binary notation of j is 1.

Now let L' be the language over Σ_o consisting of all strings of length $n-1$ that have a symbol from S_i in position $i\,m$ from the end ($i = 1, 2, \ldots, t$). Language L' is accepted by an n-state incomplete dfa A' over Σ_o with states $0, 1, \ldots, n-1$, of which 0 is the initial state, and $n-1$ is the sole final state.

Construct an incomplete dfa A over Σ from dfa A' by adding an unobservable transition on a_ℓ from the initial state 0 to state ℓ for $\ell = 1, 2, \ldots, n-m-1$. Let L be the language over Σ recognized by A. The projected language $P(L)$ consists of all suffixes of length at least m of strings in L'. As shown in [25, 26], every incomplete dfa for $P(L)$ needs at least $(k^{\lfloor n/(\log k+1)\rfloor+1} - 1)/(k-1)$ states. □

Our last result shows that the size of the minimal dfa for a projected language may reach an arbitrary value from 1 up to the upper bound $2^{n-1} + 2^{n-2} - 1$. Hence there are no magic numbers for projections of regular languages.

Theorem 9. *Let $n \geq 2$ and $1 \leq \alpha \leq 2^{n-1} + 2^{n-2} - 1$. There exist an alphabet Σ, a projection P of strings in $(\Sigma \cup \{\#\})^*$ onto strings in Σ^* with $\# \notin \Sigma$, and a regular language L over $\Sigma \cup \{\#\}$ with $\|L\| = n$ such that $\|P(L)\| = \alpha$.*

Proof. If $1 \leq \alpha \leq n-2$, then take the minimal incomplete dfa of Fig. 6 with $\Sigma = \{a\}$. The projected language is $\{a^i \mid i \geq \alpha - 1\}$, for which the minimal incomplete dfa has α states.

If $\alpha = n-1$, then take the incomplete dfa of Fig. 7 with $\Sigma = \{a\}$. The projected language is $(a^{n-1})^*$, for which the minimal incomplete dfa has $n-1$ states.

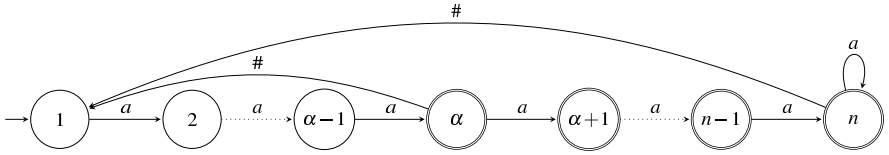

Fig. 6. The incomplete n-state dfa A over $\{a, \#\}$ with $\|P(L(A))\| = \alpha$; $1 \leq \alpha \leq n-2$.

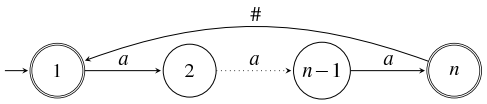

Fig. 7. The incomplete n-state dfa A over $\{a,\#\}$ with $\|P(L(A))\| = n-1$.

Now let $n \leq \alpha \leq 2^{n-1} + 2^{n-2} - 1$. Then $n+1 \leq \alpha+1 \leq \frac{3}{4} \cdot 2^n$, and so $\alpha+1$ can be expressed as $\alpha+1 = n-k+\frac{3}{4} \cdot 2^k + m$, where $2 \leqslant k \leqslant n$ and $0 \leqslant m \leqslant 2^{k-1}+2^{k-2}-2$. It is shown in [14, Lemma 9 and Lemma 10] that there exists a minimal n-state dfa $M_{n,k,m}$ over an alphabet Σ with states $1,2,\ldots,n$, of which 1 is the initial state, and k is the sole final state (and no state is dead) such that the minimal dfa for the star of language $L(M_{n,k,m})$ has $\alpha+1$ states.

Let us modify the dfa $M_{n,k,m}$ by adding an unobservable transition by symbol # from the final state k to the initial state 1. Then in the subset automaton for the projected language, all the states that were reachable in the subset automaton for star will be again reachable, except for the initial state $\{q_0\}$ that was added in the construction of an nfa for star in [14] . All the reachable states will be pairwise distinguishable. Therefore, the minimal incomplete dfa for the projected language has exactly α states. □

6 Conclusion

The dfa accepting a projected language is obtained from the dfa accepting an input language by replacing unobservable transitions with ε-transitions and by applying the subset construction to the resulting nfa. The minimal dfa for the projected language, however, may be of exponential size in comparison with the input automaton [9, 19, 21, 22]. This observation gives rise to a challenging open problem. How to characterize classes of dfa's, for which the minimal dfa for the projections is of a linear (polynomial, logarithmic) size?

Problem 1. Let P be a projection, and let $\mathbb{A}_P^f$ denote the class of all minimal dfa's such that $A \in \mathbb{A}_P^f$ if and only if the minimal dfa accepting $P(L(A))$ has no more than $f(n)$ states, where f is a (recursive) upper bound state-space function. Given a projection P and a function f, characterize the class $\mathbb{A}_P^f$.

It follows from the results of this paper that the class $\mathbb{A}_P^f$ does not include all minimal acyclic dfa's for any reasonable upper bound f (such as linear or polynomial). Note that there exists a property called an *observer property* [29] ensuring that the minimal automaton for the projected language has no more states than the minimal automaton for the input language, see also [23]. This property is well known and widely used in supervisory control of hierarchical and distributed discrete-event systems, and, as mentioned in [24], also in compositional verification [5] and modular synthesis [3, 7]. If the projection does not satisfies the property, the co-domain of the projection can be extended so that it satisfies it. However, the computation of such a minimal extension is NP-hard.

Nevertheless, there exists a polynomial-time algorithm that finds an acceptable extension [4]. A different approach with further references can be found in [24]. Although we know that the result is of polynomial size, the problem is how to compute it in polynomial time. Consider the determinization procedure of an nfa. This procedure can produce an exponential number of states where most of the states are equivalent. In [29], a polynomial-time algorithm running in $O(n^7m^2)$, where n is the number of states and m is the cardinality of the co-domain of the projection satisfying the observer property, has been proposed. However, the precise time complexity of this problem is open.

Problem 2. How to compute the minimal dfa accepting the projected language when the projection satisfies the observer property?

Acknowledgement

We would like to thank Professor Jan H. van Schuppen for his useful comments.

References

1. Cassandras, C.G., Lafortune, S.: Introduction to discrete event systems, Second edition. Springer (2008)
2. Chrobak, M.: Finite automata and unary languages. Theoret. Comput. Sci. 47(2), 149–158 (1986), Errata: Theoret. Comput. Sci. 302, 497-498 (2003)
3. Feng, L., Wonham, W.M.: Computationally efficient supervisor design: Abstraction and modularity. In: Proc. of WODES 2006. pp. 3–8. Ann Arbor, USA (2006)
4. Feng, L., Wonham, W.M.: On the computation of natural observers in discrete-event systems. Discrete Event Dyn. Syst. 20, 63–102 (2010)
5. Flordal, H., Malik, R.: Compositional verification in supervisory control. SIAM J. Control Optim. 48(3), 1914–1938 (2009)
6. Geffert, V.: Magic numbers in the state hierarchy of finite automata. Inf. Comput. 205(11), 1652–1670 (2007)
7. Hill, R.C., Tilbury, D.M.: Modular supervisory control of discrete event systems with abstraction and incremental hierarchical construction. In: Proc. of WODES 2006. pp. 399–406. Ann Arbor, USA (2006)
8. Holzer, M., Jakobi, S., Kutrib, M.: The magic number problem for subregular language families. In: Proc. of DCFS 2010. EPTCS, vol. 31, pp. 110–119 (2010)
9. Holzer, M., Kutrib, M.: Descriptional complexity – an introductory survey. In: Scientific Applications of Language Methods, vol. 2. Imperial College Press (2010)
10. Iwama, K., Kambayashi, Y., Takaki, K.: Tight bounds on the number of states of DFAs that are equivalent to n-state NFAs. Theoret. Comput. Sci. 237, 485–494 (2000)
11. Iwama, K., Matsuura, A., Paterson, M.: A family of NFAs which need $2^n - \alpha$ deterministic states. Theoret. Comput. Sci. 301, 451–462 (2003)
12. Jirásek, J., Jirásková, G., Szabari, A.: Deterministic blow-ups of minimal nondeterministic finite automata over a fixed alphabet. IJFCS 19, 617–631 (2008)
13. Jirásková, G.: Note on minimal finite automata. In: Proc. of MFCS 2001. LNCS, vol. 2136, pp. 421–431. Springer (2001)

14. Jirásková, G.: State complexity of complements, stars, and reversals of regular languages. In: Proc. of DLT 2008. LNCS, vol. 5257, pp. 431–442. Springer (2008), Full version: http://im3.saske.sk/∼jiraskov/star/
15. Jirásková, G.: Magic numbers and ternary alphabet. In: Proc. of DLT 2009. LNCS, vol. 5583, pp. 300–311. Springer (2009)
16. Komenda, J., Masopust, T., van Schuppen, J.H.: Supervisory control synthesis of discrete-event systems using a coordination scheme. CoRR 1007.2707 (2010), http://arxiv.org/abs/1007.2707
17. Komenda, J., Masopust, T., van Schuppen, J.H.: Synthesis of safe sublanguages satisfying global specification using coordination scheme for discrete-event systems. In: Proc. of WODES 2010. pp. 436–441. Berlin, Germany (2010)
18. Komenda, J., van Schuppen, J.H.: Coordination control of discrete event systems. In: Proc. of WODES 2008. pp. 9–15. Göteborg, Sweden (2008)
19. Lupanov, O.B.: Über den vergleich zweier typen endlicher quellen. Probl. Kybernetik 6, 328–335 (1966), translation from Probl. Kibernetiki 9, 321-326 (1963)
20. Lyubich, Y.I.: Estimates for optimal determinization of nondeterministic autonomous automata. Sib. Matemat. Zhu. 5, 337–355 (1964), in Russian
21. Meyer, A.R., Fischer, M.J.: Economy of description by automata, grammars, and formal systems. In: Proc. of FOCS 1971. pp. 188–191. IEEE (1971)
22. Moore, F.R.: On the bounds for state-set size in the proofs of equivalence between deterministic, nondeterministic, and two-way finite automata. IEEE Trans. Comput. 20(10), 1211–1214 (1971)
23. Pena, P.N., Cury, J.E.R., Lafortune, S.: Polynomial-time verification of the observer property in abstractions. In: Proc. of ACC 2008. pp. 465–470. Seattle, USA (2008)
24. Pena, P.N., Cury, J.E.R., Malik, R., Lafortune, S.: Efficient computation of observer projections using OP-verifiers. In: Proc. WODES 2010, pp. 416-421 (2010)
25. Salomaa, K.: NFA to DFA conversion for finite languages over a k-letter alphabet. Personal communication (2011)
26. Salomaa, K., Yu, S.: NFA to DFA transformation for finite languages. In: Proc. of WIA 1996. LNCS, vol. 1260, pp. 149–158. Springer (1996)
27. Sipser, M.: Introduction to the theory of computation. PWS Publishing Company, Boston, USA (1997)
28. Szabari, A.: Descriptional Complexity of Regular Languages. Ph.D. thesis, Mathematical Institute, Slovak Academy of Sciences, Košice, Slovakia (2010)
29. Wong, K.: On the complexity of projections of discrete-event systems. In: Proc. of WODES 1998. pp. 201–206. Cagliari, Italy (1998)
30. Wonham, W.M.: Supervisory control of discrete-event systems, Lecture Notes, Dept. of Electrical and Computer Engineering, Univ. of Toronto, Canada (2009)
31. Yu, S.: Regular languages. In: Handbook of Formal Languages – Vol. I, pp. 41–110. Springer (1997)