State Complexity of Basic Operations on Finite Languages*

C. Câmpeanu[1], K. Culik II[2], K. Salomaa[3], and S. Yu[3]

[1] Fundamentals of Computer Science Department, Faculty of Mathematics
University of Bucharest, Romania cezar@funinf.math.unibuc.ro
[2] Department of Computer Science, University of South Carolina
Columbia, SC 29208, USA culik@cs.scarolina.edu
[3] Department of Computer Science, The University of Western Ontario
London, Ontario, Canada N6A 5B7 {ksalomaa, syu}@csd.uwo.ca

Abstract. The state complexity of basic operations on regular languages has been studied in [9–11]. Here we focus on finite languages. We show that the catenation of two finite languages accepted by an m-state and an n-state DFA, respectively, with $m > n$ is accepted by a DFA of $(m-n+3)2^{n-2}-1$ states in the two-letter alphabet case, and this bound is shown to be reachable. We also show that the tight upper-bounds for the number of states of a DFA that accepts the star of an n-state finite language is $2^{n-3}+2^{n-4}$ in the two-letter alphabet case. The same bound for reversal is $3 \cdot 2^{p-1}-1$ when n is even and 2^p-1 when n is odd. Results for alphabets of an arbitrary size are also obtained. These upper-bounds for finite languages are strictly lower than the corresponding ones for general regular languages.

1 Introduction

Many applications of regular languages use essentially finite languages. In [9–11], the state complexity of basic operations on regular languages has been studied. It is interesting and important to know whether those state-complexity results still hold for finite languages. For example, $(2m-1)2^{n-1}$ is the number of states of a minimal DFA, in the worst case, that accepts the catenation of an m-state and an n-state DFA language. Does the catenation of two DFA, each accepting a finite language, need the same number of states in the worst case? May it be significantly smaller?

It is known [4] that a minimal DFA that accepts the reversal of an n-state DFA language needs 2^n states in the worst case. This fact determines that Brzozowski's DFA minimization algorithm [1, 7], which uses two reversals, is exponential in time and space in the worst case. However, this algorithm is faster than other algorithms in many experiments. It is a natural question whether this algorithm has a polynomial time or space complexity in the case of finite

* The work reported here has been supported by the Natural Sciences and Engineering Research Council of Canada Grants OGP0041630 and OGP0147224.

languages. This question is very much related to the state complexity of the reversal of finite languages.

In this paper, we focus on the above mentioned problems and on the state complexity of basic operations on finite languages, in general. We show that for an n-state DFA A accepting a finite language L, a minimal DFA that accepts L^* has $2^{n-3}+2^{n-t-2}$ states in the worst case, where $t \geq 2$ is the number of final states in A (except the starting state). Note that for $t = 1$, this bound is simply $n-1$.

For the catenations of finite languages, we show that a minimal DFA that accepts the catenation of two finite languages, which are accepted by an m-state DFA and an n-state DFA, respectively, has at most

$$\sum_{i=0}^{m-2} \min\left\{k^i, \binom{n-2}{\leq i}, \binom{n-2}{\leq t-1}\right\} + \min\left\{k^{m-1}, \binom{n-2}{\leq t}\right\}$$

states, where k is the size of the alphabet and t is the number of final states in the first automaton. Notice that this bound depends very much on t. If t is a constant, then this bound is $O(mn^{t-1}+n^t)$, which is polynomial. In particular, when $t = 1$, it is $m+n-2$. In the case of a two-letter alphabet (with an arbitrary t), this bound is $(m-n+3)2^{n-2}-1$. We give examples to show that this bound is reachable.

We also show that $\sum_{i=0}^{t-1} k^i + 2^{n-1-t}$ is an upper bound on the number of states for a minimal DFA that accepts the reversal of a finite language accepted by an n-state DFA, where t is the smallest integer such that $2^{n-1-t} \leq k^t$. This bound is, in the case of a two-letter alphabet, $3 \cdot 2^{p-1}-1$ if $n = 2p$ or $2^p - 1$ if $n = 2p-1$. We also give examples to show that the latter bounds are reachable. Unfortunately, these results show that Brzozowski's DFA minimization algorithm is still exponential in the worst case even for finite languages.

We also consider the state complexity of operations on finite languages in the case of a one-letter alphabet.

2 Preliminaries

Let T be a finite set. Denote by $\#T$ the cardinality of T and by T^* the free monoid generated by T. The empty word, i.e., the neutral element of T^*, is denoted by λ and $T^+ = T^* - \{\lambda\}$. For $w \in T^*$, denote by $|w|$ the length of w. We define

$$T^l = \{w \in T^* \mid |w| = l\}, \quad T^{\leq l} = \bigcup_{i=0}^{l} T^i, \text{ and } T^{<l} = \bigcup_{i=0}^{l-1} T^i.$$

If $T = \{t_1, \ldots, t_k\}$ is an ordered set, $k > 0$, the lexicographical order on T^*, denoted $\preceq_l$, is defined by: $x \preceq_l y$ iff $x = y$ or $|x| < |y|$ or $|x| = |y|$ and $x = zt_iv$, $y = zt_ju$, $i < j$, for some $z, u, v \in T^*$ and $1 \leq i, j \leq k$. We say that x is a prefix of y, denoted $x \preceq_p y$ if $y = xz$ for some $z \in T^*$. The relation $\preceq_p$ is a partial order on T^*.

A deterministic finite automaton (DFA) is a quintuple $A = (Q, \Sigma, \delta, q_0, F)$, where Q is the finite nonempty set of states; Σ is the finite nonempty alphabet; $q_0 \in Q$ is the starting state; $F \subseteq Q$ is the set of final states; and $\delta : Q \times \Sigma \longrightarrow Q$ is the transition function. We extend δ from $Q \times \Sigma$ to $Q \times \Sigma^*$ by $\overline{\delta}(q, aw) = \overline{\delta}(\delta(q, a), w)$ and $\overline{\delta}(q, \lambda) = q$ for $q \in Q$, $a \in \Sigma$, and $w \in \Sigma^*$. We usually denote $\overline{\delta}$ by δ if there is no confusion.

The language recognized by the automaton A is $L(A) = \{w \in \Sigma^* \mid \delta(q_0, w) \in F\}$. Two automata are equivalent if they recognize the same language.

For simplicity, in what follows, we assume that $Q = \{0, 1, \ldots, \#Q - 1\}$ and $q_0 = 0$. We also assume that δ is a total function, i.e., that the automaton is complete.

Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA. Then

a) a state s is said to be accessible if there exists $w \in \Sigma^*$ such that $\delta(0, w) = s$;

b) a state s is said to be useful if there exists $w \in \Sigma^*$ such that $\delta(s, w) \in F$.

It is clear that for every DFA A there exists an automaton A' such that $L(A') = L(A)$ and every state of A' is accessible and at most one state is useless (the sink state). The DFA A' is called a *reduced* DFA. We will use only reduced DFA in the following.

A DFA $A = (\Sigma, Q, q_0, \delta, F)$ is said to be *minimal* if for every other automaton $A' = (\Sigma, Q', q_0', \delta', F')$ such that $L(A) = L(A')$, we have $\#Q \leq \#Q'$.

A minimal DFA has at most one useless state.

Let $L \subseteq \Sigma^*$ and $x, y \in \Sigma^*$. Then $x \equiv_L y$ if for all $z \in \Sigma^*$, $xz \in L$ iff $yz \in L$. Clearly, $\equiv_L$ is an equivalence relation on Σ^*. The number of states in a minimal DFA that accepts L is exactly the number of equivalence classes of $\equiv_L$ [3]. If $L = L(A)$ and p, q are states of the DFA $A = (\Sigma, Q, q_0, \delta, F)$ we denote also $p \equiv_L q$ (or simply $p \equiv q$) if for all $z \in \Sigma^*$, $\delta(p, z) \in F$ iff $\delta(q, z) \in F$.

For basic definitions and results in automata theory, the reader may refer to [5, 3, 11].

3 Star operation on finite languages

In [9] (also in [11]), it was shown that for any n-state (complete) DFA A, there exists a minimal DFA of at most $2^{n-1}+2^{n-2}$ states that accepts $L(A)^*$. Examples were also given to show that this bound is reachable. In this section, we show that in the case that A accepts a finite language rather than an infinite regular language, the corresponding bound is exactly $2^{n-3} + 2^{n-4}$. The latter is exactly one-fourth of the former.

Let A be an n-state DFA accepting a finite language. If A has only one final state, it is clear that a minimal DFA accepting $L(A)^*$ needs at most $n-1$ states. Note that this is not true in general for an n-state DFA accepting an infinite regular language. It has been shown that the upper bound $2^{n-1} + 2^{n-2}$ can be reached even for n-state DFA with only one final state.

In the following, we consider DFA with at least two final states.

Theorem 1. *Let $A = (Q, \Sigma, 0, \delta, F)$ be a DFA accepting a finite language L, where $0 \notin F$, $\#F = t \geq 2$, $\#Q = n \geq 4$. Then there exists a DFA of at most $2^{n-3} + 2^{n-t-2}$ states that accepts L^*.*

Proof. We first construct an NFA A' from A by adding a λ-transition from each final state $f \in F$ to 0. Formally, $A' = (Q, \Sigma, \delta', 0, F)$ where $\delta' : Q \times \Sigma \to 2^Q$ is defined for each $p \in Q$ and $a \in \Sigma$ as follows:

$$\delta'(p,a) = \begin{cases} \{q\} & \text{if } q = \delta(p,a) \text{ and } q \notin F, \\ \{q, 0\} & \text{if } q = \delta(p,a) \text{ and } q \in F. \end{cases}$$

Clearly, A' accepts $L(A)^+$.

Next we construct a DFA $B = (Q_B, \Sigma, \delta_B, 0_B, F_B)$ from A' using the standard subset-construction method [3, 8] and, furthermore, make the starting state of B a final state which guarantees that $L(B) = L(A)^*$. Then we have $Q_B \subseteq 2^Q$, $0_B = \{0\}$, $F_B = \{P \in Q_B \mid P \cap F \neq \emptyset\} \cup \{0_B\}$, and $\delta_B(P,a) = \cup_{p \in P} \delta'(p,a)$.

In the following, we assume that, in A, $(n-1)$ is the sink state and $(n-2)$ is the final state that has transitions only to $(n-1)$. Without loss of generality, we also assume that B is a reduced DFA.

Let $P \in Q_B$. Then the following three propositions can be easily proved:

(1) If $P \cap F \neq \emptyset$, then $0 \in P$.
(2) If $(n-1) \in P$, then $P \equiv_{L^*} P - \{n-1\}$.
(3) If $(n-2) \in P$ and $P \cap (F - \{n-2\}) \neq \emptyset$, then $P \equiv_{L^*} P - \{n-2\}$.

Using the above propositions, we can simplify the DFA B by merging all equivalent states. Let the resulting DFA be $B' = (Q'_B, \Sigma, \delta'_B, 0_B, F'_B)$. So, Q'_B has at most the following states:

(i) the starting state $0_B = \{0\}$ and the sink state $\{n-1\}$,
(ii) all P such that $P \subseteq (Q - F - \{0, n-1\})$ and $P \neq \emptyset$,
(iii) all $P = \{0\} \cup P' \cup P''$ such that $P' \subseteq (Q - F - \{0, n-1\})$ and $P'' \subseteq F - \{n-2\}$ and $P'' \neq \emptyset$,
(iv) all $P = P' \cup \{0, n-2\}$ where $P' \subseteq (Q - F - \{0, n-1\})$ and $P' \neq \emptyset$.

Note that in (iv) $P' \neq \emptyset$ because $\{0, n-2\}$ is equivalent to $\{0\}$ ($\{0\} \in F'_B$), which is included in (i).

Now we calculate the number of states in each of the items above: (i) 2, (ii) $2^{n-t-2} - 1$, (iii) $2^{n-t-2}(2^{t-1} - 1)$, and (iv) $2^{n-t-2} - 1$.

Hence we have $\#Q'_B \leq 2^{n-3} + 2^{n-t-2}$. □

As we have mentioned before, when $t = 1$, we can construct a DFA of at most $n-1$ states to accept L^*. So, when $t = 2$ we obtain the maximum number of states for the above formula, i.e., $2^{n-3} + 2^{n-4}$.

Note that if $0 \in F$, then for each $P \in Q_B$ such that $\{0, n-2\} \subseteq P$ we have $P \equiv_{L^*} (P - \{n-2\})$. Thus, all states of (iv) are included in (iii). Then we have (i) 2, (ii) 2^{n-t-1}, and (iii) $2^{n-t-1}2^{t-2}$. The total number is $2^{n-t-1} + 2^{n-3}$. However, if $t \leq 2$, then we can construct a DFA of at most $n-1$ states to accept L^*. So, this formula reaches its maximum when $t = 3$, i.e., $2^{n-3} + 2^{n-4}$, which is the same as the one in the case $0 \notin F$.

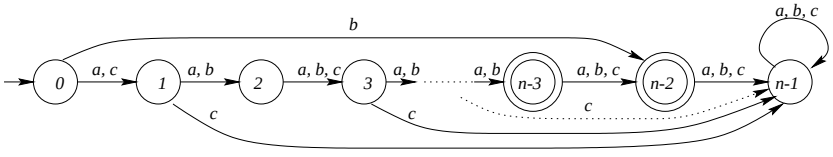
Corollary 1. *Let $A = (Q, \Sigma, \delta, 0, F)$ be a DFA accepting a finite language L, where $\#Q = n > 4$. Then there exists a DFA of at most $2^{n-3} + 2^{n-4}$ states that accepts L^*.*

Theorem 2. *There exists a DFA $A = (\Sigma, Q, \delta, 0, F)$ with $\#Q = n \geq 4$ such that any DFA recognizing $L(A)^*$ has at least $2^{n-3} + 2^{n-4}$ states.*

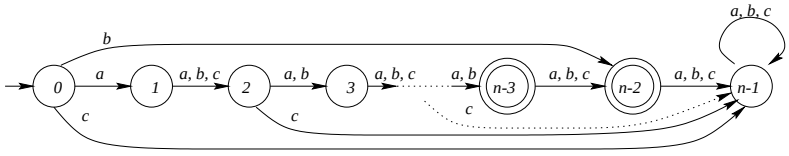
Proof. For an arbitrary integer $n \geq 4$, we define a DFA $A = (Q, \Sigma, \delta, 0, F)$, where $Q = \{0, 1, \ldots, n-1\}$, $\Sigma = \{a, b, c\}$, $F = \{n-3, n-2\}$, and δ:

$\delta(i, a) = i + 1$, for $0 \leq i \leq n-2$,
$\delta(i, b) = i + 1$, for $1 \leq i \leq n-2$, and $\delta(0, b) = n-2$,
$\delta(i, c) = i + 1$, for $0 \leq i \leq n-2$ and $n - i$ is odd,
$\delta(i, c) = n-1$, for $0 \leq i \leq n-2$ and $n - i$ is even,
$\delta(n-1, a) = n-1$, $\delta(n-1, b) = n-1$, $\delta(n-1, c) = n-1$.

The DFA A is shown in the figure below in two cases: (a) n is odd and (b) n is even.



(a) n is odd



(b) n is even

Fig. 1. DFA A of n states such that $L(A)^*$ needs $2^{n-3} + 2^{n-4}$ states

We construct a DFA $A' = (Q', \Sigma, \delta', 0', F')$ that accepts $L(A)^*$ following the two steps described in Theorem 1: (i) construct an NFA by adding a λ-transition from each final state to the starting state; (ii) construct a DFA from the resulting NFA of the previous step using the standard subset-construction algorithm.

In the following it suffices to show that every state specified in Theorem 1 is (1) reachable from the starting state $\{0\}$ and (2) in a distinct equivalence class with respect to $L(A)^*$.

We first prove that every state in the proof of Theorem 1 is reachable. For convenience, we denote the four disjoint subsets of Q' described in (i), (ii), (iii), and (iv) of Theorem 1 by $Q'_{(i)}$, $Q'_{(ii)}$, $Q'_{(iii)}$, $Q'_{(iv)}$, respectively. In particular, we have

$Q'_{(i)} = \{\{0\}, \{n-1\}\}$,
$Q'_{(ii)} = \{P \mid P \subseteq \{1, \ldots n-4\} \text{ and } P \neq \emptyset\}$,
$Q'_{(iii)} = \{P \cup \{0, n-3\} \mid P \subseteq \{1, \ldots n-4\}\}$,
$Q'_{(iv)} = \{P \cup \{0, n-2\} \mid P \subseteq \{1, \ldots n-4\} \text{ and } P \neq \emptyset\}$.

For $Q'_{(i)}$, obviously, the starting state $\{0\}$ and the sink state $\{n-1\}$ are both reachable. Now we prove the following claim:

Claim. Every state $q' \in Q'_{(iii)}$ is reachable (from the starting state $\{0\}$).

Let $q' \in Q'_{(iii)}$. Then $q' = P \cup \{0, n-3\}$ for some $P \subseteq \{1, \ldots, n-4\}$. We prove the claim by induction on the size of P. If $\#P = 0$, then $q' = \{0, n-3\}$. It is clear that $q' = \delta'(\{0\}, a^{n-3})$. Suppose that every state q' is reachable for $\#P = k$, $0 \leq k < n-4$. Consider the case when $\#P = k+1$. Let $q' = \{0, i_0, i_1, \ldots, i_k, n-3\}$. We know that $q'' = \{0, i_2 - i_1, \ldots, i_k - i_1, n-3-i_1, n-3\}$ is reachable by the induction hypothesis. Then it is clear that

$$\begin{aligned}
&\delta'(q'', ab^{i_1 - i_0 - 1}a^{i_0}) \\
&= \delta'(\{0, 1, i_2 - i_1 + 1, \ldots, i_k - i_1 + 1, n-3-i_1+1, n-2\}, b^{i_1-i_0-1}a^{i_0}) \\
&= \delta'(\{0, i_1 - i_0, i_2 - i_0, \ldots, i_k - i_0, n-3-i_0, n-2\}, a^{i_0}) \\
&= \{0, i_0, i_1, i_2, \ldots, i_k, n-3\} = q'.
\end{aligned}$$

Note that if $q' = \{0, i_0, i_1, n-3\}$, let $q'' = \{0, n-3-i_1, n-3\}$. Then again $q' = \delta'(q'', ab^{i_1-i_0-1}a^{i_0})$. If $q' = \{0, i_0, n-3\}$ ($k = 0$), let $q'' = \{0, n-3\}$ and $q' = \delta'(q'', ab^{n-3-i_0-1}a^{i_0})$. Therefore, we have proved the claim.
Note that the claim directly implies that any state $P_2 \in Q'_{(iv)}$ is reachable since for any $P_2 = \{0, i_1, \ldots, i_k, n-2\}$, where $0 < i_1 < \ldots < i_k < n-3$, we have $P'_2 = \{0, i_1 - 1, \ldots, i_k - 1, n-3\} \in Q'_{(iii)}$ such that $\delta'(P'_2, b) = P_2$. Note that it is possible that $i_1 - 1 = 0$.

It is also clear that every state $P \in Q'_{(ii)}$ is reachable since for any such state $P = \{i_1, \ldots, i_k\}$, where $0 < i_1 < \ldots < i_k < n-3$, we have $P' = \{0, i_2 - i_1, \ldots, i_k - i_1, n-2\}$ such that $\delta'(P', a^{i_1}) = P$. So, we have proved that every state specified in Theorem 1 is reachable from $\{0\}$.

Now, we prove that every state above is in a distinct equivalence class of $\equiv_{L^*}$.

It is clear that if two states p and q are from different sets of $Q'_{(i)}$, $Q'_{(ii)}$, $Q'_{(iii)}$, and $Q'_{(iv)}$, then $p \not\equiv q$ (with respect to L^*). It suffices in the remaining to prove that if there exists $i \in \{1, \ldots, n-4\}$ such that $i \in p - q$, then $p \not\equiv q$. If $n-i$ is odd, then both $\delta'(p, ca^{n-i-4})$ and $\delta'(p, ca^{n-i-3})$ are final, but $\delta'(q, ca^{n-i-4})$ and $\delta'(q, ca^{n-i-3})$ cannot be final at the same time. If $n-i$ is even and $i < n-4$, then both $\delta'(p, aca^{n-i-5})$ and $\delta'(p, aca^{n-i-4})$ are final, but $\delta'(q, aca^{n-i-5})$ and $\delta'(q, aca^{n-i-5})$ cannot be both final. If $i = n-4$, then $\delta'(p, a) \in F'$ but $\delta'(q, a) \notin F'$. Therefore, $p \not\equiv q$. □

We do not yet have an example for the two-letter alphabet case. It is still open whether there exists a lower upper bound for the two-letter alphabet case.

4 Catenation of finite languages

We now consider the state complexity of the catenation of two finite languages.

Without loss of generality, we assume that all the DFA we are considering are reduced and ordered. A DFA $A = (Q, \Sigma, \delta, 0, F)$ with $Q = \{0, 1, \ldots, n\}$ is called an ordered DFA if, for any $p, q \in Q$, the condition $\delta(p, a) = q$ implies that $p \leq q$.

For convenience, we introduce the following notation:

$$\binom{n}{\leq i} = \sum_{j=0}^{i} \binom{n}{j}.$$

Theorem 3. *Let $A_i = (Q_i, \Sigma, \delta_i, 0, F_i)$, $i = 1, 2$, be two DFA accepting finite languages L_i, $i = 1, 2$, respectively, and $\#Q_1 = m$, $\#Q_2 = n$, $\#\Sigma = k$, and $\#F_1 = t$. There exists a DFA $A = (Q, \Sigma, \delta, s, F)$ such that $L(A) = L(A_1)L(A_2)$ and*

$$\#Q \leq \sum_{i=0}^{m-2} \min\left\{k^i, \binom{n-2}{\leq i}, \binom{n-2}{\leq t-1}\right\} + \min\left\{k^{m-1}, \binom{n-2}{\leq t}\right\}. \quad (*)$$

Proof. The DFA A is constructed in two steps. First, an NFA A' is constructed from A_1 and A_2 by adding a λ-transition from each final state in F_1 to the starting state 0 of A_2. Then, we construct a DFA A from the NFA A' by the standard subset construction. Again, we assume that A is reduced and ordered.

It is clear that we can view each $q \in Q$ as a pair (q_1, P_2), where $q_1 \in Q_1$ and $P_2 \subseteq Q_2$. The starting state of A is $s = (0, \emptyset)$ if $0 \notin F_1$ and $s = (0, \{0\})$ if $0 \in F_1$. Let us consider all states $q \in Q$ such that $q = (i, P)$ for a particular state $i \in Q_1 - \{m-1\}$ and some set $P \subseteq Q_2$. Since A_1 is ordered and acyclic, the number of such states in Q is restricted by the following three bounds: (1) k^i, (2) $\binom{n-2}{\leq i}$, and (3) $\binom{n-2}{\leq t-1}$. We explain these bounds below informally.

We have (1) as a bound since all states of the form $q = (i, P)$ are at a level $\leq i$, which have at most k^{i-1} predecessors. By saying that a state p is at level i we mean that the length of the longest path from the starting state to q is i.

We now consider (2). Notice that if $q, q' \in Q$ such that $\delta(q, a) = q'$, $q = (q_1, P_2)$ and $q' = (q_1', P_2')$, then $\delta_1(q_1, a) = q_1'$ and $P_2' = \{\delta_2(p, a) \mid p \in P_2\}$ if $q_1' \notin F_1$ and $P_2' = \{0\} \cup \{\delta_2(p, a) \mid p \in P_2\}$ if $q_1' \in F_1$. So, $\#P_2' > \#P_2$ is possible only when $q_1' \in F_1$. Therefore, for $q = (i, P)$, $\#P \leq i$ if $i \notin F_1$ and $\#P \leq i + 1$ if $i \in F_1$. In both cases, the maximum number of distinct sets P is $\binom{n-2}{\leq i}$. The number $n-2$ comes from the exclusion of the sink state $n-1$ and starting state 0 of A_2. Note that, for a fixed i, either $0 \in P$ for all $(i, P) \in Q$ or 0 is not in any set P such that $(i, P) \in Q$.

(3) is a bound since for each state $i \in Q_1 - \{m-1\}$, there are at most $t-1$ final states on the path from the starting state to i (not including i).

For the second term of $(*)$, it suffices to explain that for each $(m-1, P)$, $P \subseteq Q_2$, $\#P$ is bounded by the total number of final states in F_1. □

Corollary 2. *Let $A_i = (Q_i, \Sigma, \delta_i, 0, F_i)$, $i = 1, 2$, be two DFA accepting finite languages L_i, $i = 1, 2$, respectively, and $\#Q_1 = m$, $\#Q_2 = n$, and $\#F_1 = t$, where $t > 0$ is a constant. Then there exists a DFA $A = (Q, \Sigma, \delta, s, F)$ of $O(mn^{t-1} + n^t)$ states such that $L(A) = L(A_1)L(A_2)$.*

We can simplify the formula in Theorem 3 for the case when $k = 2$, $m + 1 \geq n > 2$.

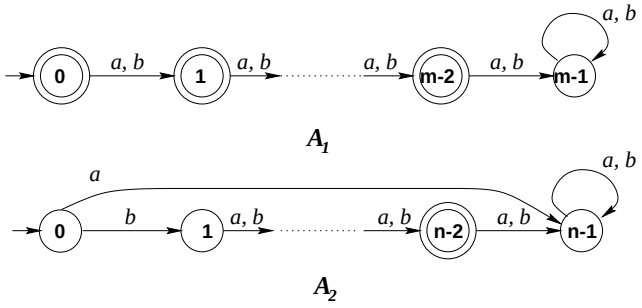
Corollary 3. *For $k = 2$ and $m+1 \geq n > 2$, the upper bound given in Theorem 3 is*

$$(m - n + 3)2^{n-2} - 1.$$

We omit the details of the mathematical calculation.

Theorem 4. *The upperbound given in Corollary 3 is reachable.*

Proof. Let $A_1 = (Q_1, \Sigma, \delta_1, 0, F_1)$ and $A_2 = (Q_2, \Sigma, \delta_2, 0, F_2)$, with $\Sigma = \{a, b\}$, $Q_1 = \{0, 1, \ldots, m-1\}$, $Q_2 = \{0, 1, \ldots, n-1\}$, and $m + 1 \geq n > 2$. A_1 and A_2 are shown below.



Let $L = L(A_1)L(A_2)$. We show that there are at least $(m - n + 3)2^{n-2} - 1$ equivalence classes of the relation $\equiv_L$ over Σ^*.

Consider all words $w \in \Sigma^*$ such that $|w| \leq m - 2$.

If $w_1, w_2 \in \Sigma^*$, $|w_1|, |w_2| \leq m - 2$, and $|w_1| < |w_2|$, then $w_1 \not\equiv_L w_2$ since $w_1 b^{n+m-4-|w_1|} \in L(A)$ but $w_2 b^{n+m-4-|w_1|} \notin L(A)$.

Let $|w_1| = |w_2|$ but $w_1 \neq w_2$ and w_1 and w_2 differ at the ith position from the right, $i \leq n - 2$. We assume that w_1 contains an a and w_2 contains a b at that position. Then $w_1 \not\equiv_L w_2$ since $w_1 a^{n-2-i} \notin L$ but $w_2 a^{n-2-i} \in L$.

So, for each k, $0 \leq k \leq n - 2$, words of length k belong to 2^k distinct equivalence classes of $\equiv_L$. For each k, $n - 2 < k \leq m - 2$, words of length k belong to at least 2^{n-2} distinct equivalence classes.

Therefore there are at least

$$1 + 2 + \ldots + 2^i + \ldots + 2^{n-2} + \underbrace{2^{n-2} + \ldots + 2^{n-2}}_{m-2-(n-2)+1 \text{ terms}}$$

$$= 2^{n-1} - 1 + (m-n+1)2^{n-2}$$
$$= (m-n+3)2^{n-2} - 1$$

equivalence classes of $\equiv_L$. □

5 Reversal of finite languages

Next we develop a tight upper bound for the state complexity of the reversal of a finite language.

Theorem 5. *Let $A = (Q, \Sigma, \delta, 0, F)$ be a DFA accepting a finite language L, where $\#Q = n \geq 3$ and $\#\Sigma = k \geq 2$. Let t be the smallest integer such that $2^{n-1-t} \leq k^t$. Then there exists a DFA $B = (Q_B, \Sigma, \delta_B, 0, F_B)$, with $\#Q_B \leq \sum_{i=0}^{t-1} k^i + 2^{n-1-t}$, that accepts L^R, i.e., the reversal of L.*

Proof. B is constructed by first reversing all the transitions of A and then determinizing the resulting NFA by the standard subset construction. Then each state in Q_B is a subset of Q. Recall that the level of a state in a finite automaton is the length of the shortest path from the starting state to this state. It is clear that the number of states at each level i of B is bounded by k^i. It is also not difficult to see that this number is bounded also by 2^{n-1-i} since they are subsets of at most $n-1-i$ states of A. Let l be the length of the longest word(s) in L (or L^R). The latter bound holds because for each i, $0 \leq i \leq l$, there exists at least one state of A that can be in a state of B of level i but not in any state of a higher level. Then the number of states at each level i is bounded by $min\{k^i, 2^{n-1-i}\}$. Since t is the smallest integer such that $2^{n-1-t} \leq k^t$, we have $\#Q_B \leq \sum_{i=0}^{t-1} k^i + 2^{n-1-t}$. Note that 2^{n-1-t} is the number of all remaining subsets of Q after the first $t-1$ levels. □

Corollary 4. *Let $|\Sigma| = 2$ and A be a DFA of $n \geq 3$ states, accepting a finite language $L \subseteq \Sigma^*$. Then there exists a DFA B that accepts L^R such that B has at most $3 \cdot 2^{p-1} - 1$ states if $n = 2p$ or $2^p - 1$ states if $n = 2p-1$.*

Proof. Since $k = 2$, we have $2^{n-1-t} \leq 2^t$, i.e. $n - 1 \leq 2t$. If $n = 2p$ then $t = p$ and $n - 1 - t = 2p - 1 - p = p - 1$. We have

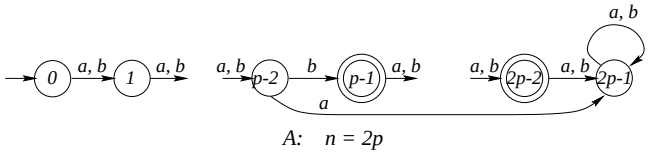
$$\sum_{i=0}^{t-1} 2^i + 2^{n-1-t} = 2^t - 1 + 2^{p-1} = 3 \cdot 2^{p-1} - 1.$$

If $n = 2p-1$ then $t = p-1$ and $n - 1 - t = 2p - 1 - 1 - p + 1 = p - 1$. We have

$$\sum_{i=0}^{t-1} 2^i + 2^{n-1-t} = 2^{p-1} - 1 + 2^{p-1} = 2^p - 1.$$

□

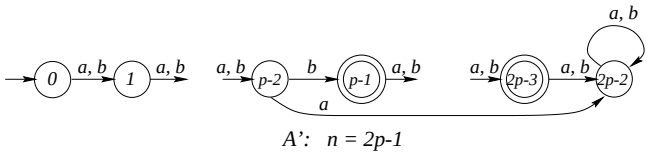
Theorem 6. *The bounds given by Corollary 4 are reachable.*



Proof. If $n = 2p$ for some integer $p > 1$, consider the DFA $A = (Q, \Sigma, \delta, 0, F)$ in the above figure.

Clearly, the reversal of A is equivalent to the catenation of A_1 and A_2 given in Theorem 4, with $m = n = p+1$. Then any DFA accepting $L(A)^R$ has at least $2^{p-1} + 2^p - 1 = 3 \cdot 2^{p-1} - 1$ states.

If $n = 2p-1$ for some integer $p > 1$, then look at the DFA $A' = (Q', \Sigma, \delta', 0, F')$ below:



The reversal of A' is equivalent to the catenation of A_1 and A_2 given in Theorem 4 with $m = p$ and $n = p + 1$. Thus, the number of states is at least $2^p - 1$. □

6 Operations on finite languages over a one-letter alphabet

We consider the case when $\#\Sigma = 1$. Without loss of generality, we assume that $\Sigma = \{a\}$.

Notice that if $A = (Q, \{a\}, 0, \delta, F)$ is a minimal DFA that accepts words of length at most l, then $\#Q = l + 1$.

Theorem 7. *Let $A_i = (Q_i, \{a\}, 0, \delta_i, F_i)$, $i = 1, 2$ be two minimal DFA, with $\#L(A_i) < \infty$, $\#Q_1 = m$, and $\#Q_2 = n$. Let $A = (Q, \{a\}, 0, \delta, F)$, $\#Q = k$, be a minimal DFA. Then we have the following:*
a) If $L(A) = L(A_1) \cup L(A_2)$, then $k = \max\{m, n\}$,
b) If $L(A) = L(A_1) \cap L(A_2)$, then $k \leq \min\{m, n\}$,
c) If $L(A) = L(A_1) - L(A_2)$, then $k \leq m$,
d) If $L(A) = L(A_1) \Delta L(A_2)$, then $k \leq \max\{m, n\}$,
e) If $L(A) = \{a\}^* - L(A_1)$, then $k = m$,
f) If $L(A) = L(A_1)L(A_2)$, then $k = m + n - 1$.
g) If $L(A) = L(A_1)^*$, then $k \leq m^2 - 7m + 13$ for $m > 4$ and $m = 3$, $k \leq 2$ otherwise.

h) If $L(A) = a \setminus L(A_1)$, *then* $k = m - 1$.
i) If $L(A) = (L(A_1)^R$, *then* $k = m$.

Proof. For a)–f) and h) the proof is obvious. For g), we give an informal proof in the following. It is clear that the length of the longest word accepted by A_1 is $m-2$. We consider the following three cases (1) A_1 has one final state; (2) A_1 has two final states; or (3) A_1 has three or more final states. If (1), then A has $m-1$ states. For (2), we need a lemma (Lemma 5.1 (iii)) from [9] which says that for two positive integers i and j, $(i,j) = 1$, the largest integer that cannot be presented as $ci + dj$ for any integers $c, d \geq 0$ is $i * j - (i + j)$. Let $i = m - 2$ and $j = m - 3$, i.e., $F_1 = \{m-2, m-3\}$. Then the length of the longest word that is not in $L(A)$ is

$$(m-2)(m-3) - (2m-5) = m^2 - 7m + 11.$$

Then A has exactly $m^2 - 7m + 13$ states. If (3), it is easy to see that A cannot have more than $m^2 - 7m + 13$ states. □

Remark 1. All the above bounds are the lowest upper bounds in the worst case. If the initial DFA A_1 and A_2 are not minimal, all the above equalities become inequalities.

References

1. J.A. Brzozowski, "Canonical regular expressions and minimal state graphs for definite events", *Mathematical Theory of Automata*, vol. 12 of MRI Symposia Series, Polytechnic Press, NY, 1962, 529-561.
2. C. Câmpeanu, "Regular languages and programming languages", *Revue Roumaine de Linguistique - CLTA*, 23 (1986), 7-10.
3. J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison Wesley (1979), Reading, Mass.
4. E. Leiss, "Succinct representation of regular languages by boolean automata", *Theoretical Computer Science* 13 (1981) 323-330.
5. A. Salomaa, *Theory of Automata*, Pergamon Press (1969), Oxford.
6. K. Salomaa and S. Yu, "NFA to DFA Transformation for Finite Languages over Arbitrary Alphabets", *Journal of Automata, Languages and Combinatorics*, 2 (1997) 3, 177-186.
7. B.W. Watson, *Taxonomies and Toolkits of Regular Language Algorithms*, PhD Dissertation, Department of Mathematics and Computing Science, Eindhoven University of Technology, 1995.
8. D. Wood, *Theory of Computation*, John Wiley and Sons, 1987.
9. S. Yu, Q. Zhuang, K. Salomaa, "The state complexities of some basic operations on regular languages", *Theoretical Computer Science* 125 (1994) 315-328.
10. S. Yu, Q. Zhuang, "On the State Complexity of Intersection of Regular Languages", *ACM SIGACT News*, vol. 22, no. 3, (1991) 52-54.
11. S. Yu, Regular Languages, in: *Handbook of Formal Languages, Vol. I,* G. Rozenberg and A. Salomaa eds., Springer Verlag, 1997, pp. 41–110.