For submission to the Journal of Automata, Languages and Combinatorics
Created on April 19, 2017

STATE COMPLEXITY OF PREFIX DISTANCE OF SUBREGULAR LANGUAGES

Timothy Ng David Rappaport Kai Salomaa

School of Computing, Queen's University
Kingston, Ontario K7L 3N6, Canada
ng@cs.queensu.ca daver@cs.queensu.ca ksalomaa@cs.queensu.ca

ABSTRACT

The neighbourhood of a regular language of constant radius with respect to the prefix distance is always regular. We give upper bounds and matching lower bounds for the size of the minimal deterministic finite automaton (DFA) needed for the radius k prefix distance neighbourhood of an n state DFA that recognizes, respectively, a finite, a prefix-convex, a prefix-closed, a prefix-free, and a right ideal language. For prefix-closed languages the lower bound automata are defined over a binary alphabet. For finite and prefix-convex regular languages the lower bound constructions use an alphabet that depends on the size of the DFA and it is shown that the size of the alphabet is optimal.

1. Introduction

The neighbourhood of radius r of a language L consists of all words that are within distance at most r from some word of L. A distance measure d is said to be regularity preserving if the neighbourhood of any regular language with respect to d is regular. Calude et al. [3] have shown that *additive distances* are regularity preserving. Additivity requires, roughly speaking, that the distance is compatible with concatenation of words in a certain sense and best known examples of additive distances include the Levenshtein distance and the Hamming distance [3, 6].

The prefix distance of two words u and v is the sum of the lengths of the suffixes of u and v that begin after the longest common prefix of u and v. The suffix distance and the factor distance are defined analogously in terms of the longest common suffix (respectively, factor) of two words. It is known that the prefix, suffix and factor distance preserve regularity [5].

By the state complexity of a regularity preserving distance we mean the worst-case size of the minimal deterministic finite automaton (DFA) needed to recognize the radius r neighbourhood of an n state DFA language (as a function of n and r). Tight bounds for the state complexity of prefix distance were recently obtained by the authors [16].

Worst-case state complexity bounds for general regular languages typically cannot be matched by finite languages, as first observed by Câmpeanu et al. [4], and the same holds for other proper sub-families of the regular languages. Relations between different sub-regular language families have been investigated recently by Holzer and Truthe [13]. Bordihn, Holzer and Kutrib [1] have studied the state complexity of determinization of automata for the different sub-regular language families. The size blow-up of determinization and operational state complexity of bounded regular languages has been studied by Herrmann et al. [9]. Further recent work on the state complexity of sub-regular language families has been done by Holzer et al. [10, 12].

Here we study the state complexity of prefix distance for finite languages. Additionally, we concentrate on the class of prefix-convex languages [2, 18] and its subclasses because their corresponding restricting properties can be viewed to be related to the definition of the prefix distance measure. We give tight state complexity bounds for the prefix distance of finite, prefix-convex, prefix-closed, prefix-free, and right ideal languages. Except for the class of prefix-closed languages, the lower bound constructions use an alphabet that depends linearly on the size of the DFA. We establish that the general upper bound cannot be matched by languages defined over an alphabet of smaller size.

2. Preliminaries

We briefly recall some definitions and notation used in the paper. For all unexplained notions on finite automata and regular languages the reader may consult the textbook by Shallit [17] or the survey by Yu [19]. A survey of distances is given by Deza and Deza [6]. Recent surveys on descriptional complexity of regular languages include [7, 11, 15].

In the following Σ is always a finite alphabet, the set of words over Σ is Σ^* and ε is the empty word. The reversal of a word $x \in \Sigma^*$ is x^R. The set of nonnegative integers is $\mathbb{N}_0$. The cardinality of a finite set S is denoted $|S|$ and the powerset of S is 2^S. A word $w \in \Sigma^*$ is a *subword* or *factor* of x if there exist words $u, v \in \Sigma^*$ such that $x = uwv$. If $u = \varepsilon$, then w is a *prefix* of x. If $v = \varepsilon$, then w is a *suffix* of x.

A *nondeterministic finite automaton* (NFA) is a 5-tuple $A = (Q, \Sigma, \delta, Q_0, F)$ where Q is a finite set of states, Σ is an alphabet, δ is a transition function $\delta : Q \times \Sigma \to 2^Q$, $Q_0 \subseteq Q$ is a set of initial states, and $F \subseteq Q$ is a set of final states. We extend the transition function δ to a function $Q \times \Sigma^* \to 2^Q$ in the usual way. A word $w \in \Sigma^*$ is *accepted* by A if, for some $q_0 \in Q_0$, $\delta(q_0, w) \cap F \neq \emptyset$ and the language recognized by A consists of all words accepted by A. An ε-NFA is an extension of an NFA where transitions can be labeled by the empty word ε [17, 19], i.e., δ is a function $Q \times (\Sigma \cup \{\varepsilon\}) \to 2^Q$. It is known that every ε-NFA A has an equivalent NFA without ε-transitions and with the same number of states as A. An NFA $A = (Q, \Sigma, \delta, Q_0, F)$ is a *deterministic finite automaton* (DFA) if $|Q_0| = 1$ and, for all $q \in Q$ and $a \in \Sigma$, $\delta(q, a)$ either consists of exactly one state or is the empty set. We also say that the transition is undefined when the transition is empty. Two states p and q of a DFA A are equivalent if $\delta(p, w) \in F$ if and only if $\delta(q, w) \in F$ for every word $w \in \Sigma^*$. A DFA A is *minimal* if each state $q \in Q$ is reachable from the initial state, a final state is reachable from each state q, and no two states are equivalent.

Note that our definition of a DFA allows some transitions to be undefined, that is, by a DFA we mean an incomplete DFA. It is well known that, for a regular language L, the sizes of the minimal incomplete and complete DFAs differ by at most one. The constructions used in this paper are more convenient to formulate using incomplete DFAs but our results would not change in any significant way if we were to require that all DFAs are complete. The (incomplete deterministic) *state complexity* of a regular language L, $\mathrm{sc}(L)$, is the size of the minimal DFA recognizing L.

Convex languages were introduced in [18] and were studied more recently in [2]. A language L is *prefix-convex* if whenever $xyz \in L$ and $x \in L$, then $xy \in L$. A language L is *prefix-closed* if whenever $xy \in L$, then $x \in L$. A language L is *prefix-free* if no word $u \in L$ is a proper prefix of any other word in L. A language L is a *right ideal* if it is non-empty and satisfies $L = L\Sigma^*$. The class of prefix-convex languages contains the class of prefix-closed languages, the class of prefix-free languages, and the right ideal languages [2].

To conclude this section, we recall definitions of the distance measures used in the following. Generally, a function $d : \Sigma^* \times \Sigma^* \to [0, \infty)$ is a *distance* if it satisfies for all $x, y, z \in \Sigma^*$, the conditions $d(x, y) = 0$ if and only if $x = y$, $d(x, y) = d(y, x)$, and $d(x, z) \leq d(x, y) + d(y, z)$. The *neighbourhood* of a language L of radius k with respect to a distance d is the set

$$E(L, d, k) = \{w \in \Sigma^* \mid (\exists x \in L)\ d(w, x) \leq k\}.$$

Let $x, y \in \Sigma^*$. The *prefix distance* of x and y counts the number of symbols which do not belong to the longest common prefix of x and y [5]. Formally, it is defined by

$$d_p(x, y) = |x| + |y| - 2 \cdot \max_{z \in \Sigma^*}\{|z| \mid x, y \in z\Sigma^*\}.$$

The state complexity of prefix distance was established in [16].

Theorem 1 [16]. *For $n > k \geq 0$, if $\mathrm{sc}(L) = n$ then*

$$\mathrm{sc}(E(L, d_p, k)) \leq n \cdot (k + 1) - \frac{k(k + 1)}{2}$$

and this bound can be reached in the worst case.

To conclude this section we recall from [16] the construction of a DFA that recognizes the prefix-distance neighbourhood of a regular language. This construction will be used as the basis for our constructions in the rest of the paper.

Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA and $\varphi_A : Q \to \mathbb{N}_0$ be a function defined by

$$\varphi_A(q) = \min_{w \in \Sigma^*} \{|w| \mid \delta(q, w) \in F\}$$

The function $\varphi_A(q)$ gives the length of the shortest path from a state q to the closest reachable final state. Note that if $q \in F$, then $\varphi_A(q) = 0$.

We construct a DFA $A' = (Q', \Sigma, \delta', q_0', F')$ for the neighbourhood $E(L(A), d_p, k)$, $k \in \mathbb{N}$, as follows. We define the state set

$$Q' = ((Q - F) \times \{1, \ldots, k + 1\}) \cup F \cup \{p_1, \ldots, p_k\}. \qquad (1)$$

The machine A' has three types of states.

- States $q \in F$, which are final states of A. A word that reaches q is a word in $L(A)$.
- States $p_\ell, 1 \leq \ell \leq k$, are reached from the other types of states only on a transition that was undefined in A. A word that reaches p_ℓ is not a prefix of a word in $L(A)$ and has a distance of ℓ from $L(A)$.
- States $(i, j) \in (Q - F) \times \{1, \ldots, k+1\}$ are non-final states of A with a counter component. If a word reaches a state (i, j) in A', then it is a prefix of a word recognized by A and is j steps away from, or to, the closest final state of A. Note that the closest final state could have been reachable earlier in the computation of the input word and may not necessarily be reachable from i. If the closest final state is more than k steps away, then $j = k + 1$.

The initial state q_0' is defined by

$$q_0' = \begin{cases} q_0, & \text{if } q_0 \in F; \\ (q_0, \varphi_A(q_0)) & \text{if } q_0 \notin F \text{ and } \varphi_A(q_0) \leq k; \\ (q_0, k+1) & \text{if } q_0 \notin F \text{ and } \varphi_A(q_0) > k. \end{cases}$$

The set of final states is given by

$$F' = ((Q - F) \times \{1, \ldots, k\}) \cup F \cup \{p_1, \ldots, p_k\}.$$

Let $q_{i,a} = \delta(i, a)$ for $i \in Q$ and $a \in \Sigma$, if $\delta(i, a)$ is defined. Then for all $a \in \Sigma$, the transition function δ' is defined for states $i \in F$ by

$$\delta'(i, a) = \begin{cases} (q_{i,a}, 1), & \text{if } q_{i,a} \in Q - F; \\ q_{i,a}, & \text{if } q_{i,a} \in F; \\ p_1, & \text{if } \delta(i, a) \text{ is undefined.} \end{cases}$$

For states $(i, j) \in (Q - F) \times \{1, \ldots, k+1\}$, δ' is defined

$$\delta'((i, j), a) = \begin{cases} q_{i,a}, & \text{if } q_{i,a} \in F; \\ (q_{i,a}, \min\{j+1, \varphi_A(q_{i,a})\}), & \text{if } \varphi_A(q_{i,a}) \text{ or } j+1 \leq k; \\ (q_{i,a}, k+1), & \text{if } \varphi_A(q_{i,a}) \text{ and } j+1 > k; \\ p_{j+1}, & \text{if } \delta(i, a) \text{ is undefined and } j < k. \end{cases}$$

Note that if $\delta(i, a)$ is undefined and $j \geq k$, then the transition is undefined. Finally, we define δ' for states p_ℓ for $\ell = 1, \ldots, k-1$ by $\delta'(p_\ell, a) = p_{\ell+1}$.

The following Proposition 2 follows from the proof of Proposition 2 of [16]. Note that Proposition 2 of [16] establishes a stronger claim and the statement of the below proposition includes only the parts that we need in the later sections.

Proposition 2 [16]. *(a) The DFA A' recognizes the neighbourhood $E(L(A), d_p, k)$. (b) The elements of the set $S_{ur} = \{(q, j) \mid q \in Q - F, 1 \leq j \leq k+1, j > \varphi_A(q)\}$ are all unreachable as states of the DFA A'.*

3. Neighbourhoods of Finite Languages

We first consider the state complexity of neighbourhoods of finite languages with respect to the prefix distance.

Proposition 3. *Let L be a finite language recognized by a minimal DFA $A = (Q, \Sigma, \delta, q_0, F)$ with n states. Then for $n > 2k$,*

$$\mathrm{sc}(E(L, d_p, k)) \leq (n-2) \cdot (k+1) - k^2 + 2.$$

Proof. We know that the neighbourhood of L of radius k with respect to the prefix distance is recognized by a DFA $A' = (Q', \Sigma, \delta', q_0'.F')$ obtained from A as in Proposition 2 where, furthermore, all elements of the set S_{ur} are unreachable. We show that there are more unreachable states in the case of finite languages.

Since A is acyclic, the number and length of words that reach each state $q \in Q$ is bounded. For $q \in Q$, let ℓ_q denote the length of a longest word that reaches q from a final state without passing through another final state. Then for all states q with $\ell_q \leq k$, the states $(q, j) \in Q'$ with $j > \ell_q$ are unreachable as states of A' (where the set of states of A' is as in (1). That is, all states in the set

$$R_{ur} = \{(q, j) \mid q \in Q - F, 1 \leq j \leq k+1, j > \ell_q\}$$

are unreachable in A'. To see this, recall that by the definition of the transition function, the second component starts from zero at a final state and changes at every computation step, either by incrementing by one or taking on the value of the length of the path to the closest reachable final state, whichever is smaller. Since ℓ_q is the length of the longest word to reach the state q, j can take on a value of at most ℓ_q. Otherwise, j takes on the value of $\varphi(q)$, in which case, it would be less than ℓ_q.

By Proposition 2 (b) all elements of the set $S_{ur} = \{(q, j) \mid q \in Q - F, 1 \leq j \leq k+1, j > \varphi_A(q)\}$ are also unreachable in A'. We note that increasing the number of final states of A by one decreases the cardinality of Q' by k and decreases the cardinality of S_{ur} and R_{ur} by at most k. However, we observe that A must have at least two final states to reach the bound.

Since A is finite, consider the longest word of A and denote its length by ℓ. Then the last state of A is a state reachable on a word of length ℓ and no states are reachable on a word of length greater than ℓ. We will denote the last state of A by q_f. This state has no outgoing transitions and must be a final state since, otherwise, there are useless states. But this cannot be the only final state, since otherwise, for every state $q \in Q$ with $\varphi_A(q) > k$, only $(q, k+1)$ is reachable. Thus, the initial state q_0 must also be a final state.

As in [16], we note that the cardinality of S_{ur} is minimized when exactly one non-final state has a shortest path of length i that reaches q_f. From the above it then follows that reaching the upper bound requires exactly two final states, one of which must be the initial state and the other which must have no outgoing transitions. Since A is acyclic, the initial state cannot have any incoming transitions, so the states in S_{ur} consist of those that can reach the non-initial final state, giving $\frac{k(k+1)}{2}$ unreachable

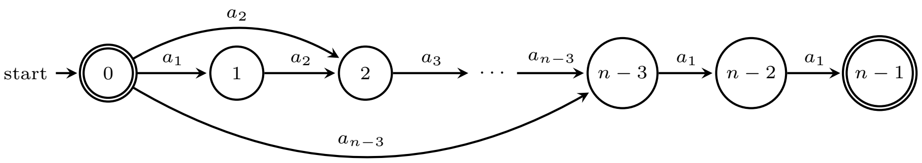

Figure 1: The DFA A_n.

states. Similarly, the cardinality of R_{ur} is minimized when exactly one non-final state has a longest word of length i which reaches it from q_0, giving $\frac{k(k+1)}{2}$ unreachable states.

Thus, the number of states of the minimal DFA for $E(L, d_p, k)$ is upper bounded by

$$(n-2)(k+1) + 2 + k - 2 \cdot \frac{k(k+1)}{2} = (n-2)(k+1) - k^2 + 2.$$

□

Next we give a lower bound construction that matches the upper bound of Proposition 3.

Lemma 4. *There exists a finite language recognized by a DFA with n states such that $E(L(A), d_p, k)$ requires at least $(n-2)(k+1) - k^2 + 2$ states.*

Proof. Let $A_n = (Q_n, \Sigma_n, \delta_n, q_0, F_n)$ where $Q_n = \{0, \ldots, n-1\}$, $\Sigma_n = \{a_1, \ldots, a_{n-3}\}$, $q_0 = 0$, $F_n = \{0, n-1\}$, and the transition function is defined by

- $\delta_n(0, a_i) = i$ for $1 < j \leq n-3$,
- $\delta_n(i, a_{i+1}) = i+1$ for $0 \leq i < n-3$,
- $\delta_n(i, a_1) = i+1$ for $i = n-3, n-2$.

The DFA A_n is depicted in Figure 1.

Let $A_n' = (Q_n', \Sigma_n, \delta_n', q_0', F_n')$ be the DFA constructed from A_n as in Proposition 2. First, we show that $(n-2)(k+1) - k^2 + 2$ states are reachable. State 0 is the initial state. State $n-1$ is reachable on the word $a_1 a_2 \cdots a_{n-3} a_1 a_1$. States of the form p_i with $1 \leq i \leq k$ are reachable from state $n-1$ on the word a_1^i. For states of the form $(i, j) \in (Q_n - F_n) \times \{1, \ldots, k+1\}$, with $\varphi_{A_n}(i) > k$ and $j \leq i$, each (i, j) is reachable on the word $a_{i-j} a_{i-j+1} \cdots a_i$. However, states (i, j) with $j > \varphi_{A_n}(i)$ are unreachable by definition of A_n' and states (i, j) with $i < j < k+1$ are unreachable. Thus the number of unreachable states in $(Q_n - F_n) \times \{1, \ldots, k+1\}$ is

$$\sum_{i=n-k}^{n-2} |\{i\} \times \{\varphi_{A_n}(i) + 1, \ldots, k+1\}| + \sum_{i=1}^{k} |\{i+1, \ldots, k+1\}|$$

$$= 2 \cdot \sum_{i=1}^{k} |\{i, \ldots, k+1\}| = 2 \cdot \sum_{i=1}^{k} i = 2 \cdot \frac{k(k+1)}{2}.$$

Thus the number of reachable states is

$$(n-2)(k+1)+2+k-2\cdot\frac{k(k+1)}{2}=(n-2)(k+1)-k^2+2.$$

Now, we show that all reachable states are pairwise inequivalent.

- For states of the form p_i and p_j, $i<j$, the word a_1^{k-i} takes the machine from state p_i to p_k and is accepted. However, from state p_j, the word a_1^{k-i} reaches state p_k on the prefix a_1^{k-j} with no further transitions to read a_1^{j-i} and thus, the word is not accepted.
- For states of the form (i,j) and p_ℓ with $\ell\le k$, we consider the word $z=w_ia_2^k$ with
$$w_i=a_{n-i+1}a_{n-i+2}\cdots a_{n-3}a_1a_1.$$
The prefix w_i takes the machine from state (i,j) to state $n-1$ and on the rest of the word a_2^k, the machine moves from $n-1$ to p_k and is accepted. However, from state p_ℓ, the computation on z reaches p_k before all of z is read, since $|z|=n-i+k>k-\ell$ and it is rejected.
- For states of the form (i,j) and (i',j') with $i<i'$ the states can be distinguished by $z=w_ia_2^k$ as above. For $i=i'$ and $j<j'$, let $z=a_ia_1^{k-j}$. From (i,j), the machine reads a_i and is taken to p_j, while from (i,j'), the machine is taken to $p_{j'}$. From above, p_j and $p_{j'}$ are distinguishable by a_1^{k-j}.
- State 0 is distinguished from every other state by the word $a_1a_2\cdots a_{n-3}a_1^{k+2}$. State $n-1$ is distinguished from every state of the form p_i for $1\le i\le k$ on the word a_1^k and from every state (i,j) on the word $w_ia_2^k$ as defined above.

Thus, we have shown that there are $(n-2)(k+1)-k^2+2$ reachable states and that all reachable states are pairwise inequivalent. □

Proposition 3 and Lemma 4 now yield a tight state complexity bound for the prefix distance neighbourhoods of regular languages.

Theorem 5. *Let L be a finite language. For $n>2k\ge 0$, if $\mathrm{sc}(L)=n$, then*

$$\mathrm{sc}(E(L,d_p,k))\le(n-2)\cdot(k+1)-k^2+2,$$

and this bound can be reached in the worst case.

Now, we consider the case when the radius k is larger than $\frac{n}{2}$, where n is the number of states in the DFA.

Theorem 6. *Let L be a finite language recognized by a minimal DFA $A=(Q,\Sigma,\delta,q_0,F)$ with n states. Then for $k\ge\frac{n}{2}$ and even n,*

$$\mathrm{sc}(E(L,d_p,k))\le\frac{n}{2}\cdot\left(\frac{n}{2}-1\right)+k+2.$$

and n is odd, then

$$\mathrm{sc}(E(L,d_p,k))\le\left(\frac{n-1}{2}\right)^2+k+2.$$

This bound can be reached in the worst case.

Proof. As in the proof of Proposition 3, the number of states is constrained by the maximal length of a word that can reach each state and the length of the shortest path to the next reachable final state. However, when we have $k \geq \frac{n}{2}$, the maximum number of reachable states decreases.

Let f be the number of final states of A. First, we consider the case when $n-f$ is even. We claim that there is no state (q,j) that is reachable with $j > \frac{n-f}{2}$. Suppose that there is some such reachable (q,j). Then it is reached by a word with length at least $\frac{n-f}{2}$. This implies that the length of the longest word that reaches q has length at least $\frac{n-f}{2}$. However, since $k > \frac{n-f}{2}$, we have $\varphi_A(q) \leq \frac{n-f}{2}$. But this means that $(q,j) \in S_{ur}$ and is unreachable.

Since these states are unreachable, we are left with at most

$$(n-f)(k+1)+k+f-\left(k-\frac{n-f}{2}\right)(n-f) = (n-f)\left(\frac{n-f}{2}+1\right)+k+f$$

possible reachable states. Now, we consider the sets R_{ur} and S_{ur} from the proof of Proposition 3, while excluding all states (q,j) with $j > \frac{n-f}{2}$, as they have already been counted. We now recall some properties of R_{ur} and S_{ur}.

(i) The cardinality of S_{ur} is minimized when exactly one non-final state has a shortest path of length i that reaches a final state. This implies that there must be a final state q_f that is the last state of A.

(ii) The cardinality of R_{ur} is minimized when exactly one non-final state has a longest word of length i which reaches it from from a final state. This implies that the initial state q_0 must be a final state.

This means that A must have at least two final states. Since increasing the number of final states reduces the cardinality of Q' by a factor of $\frac{n-f}{2}$, the size of Q' is maximized when $f=2$. It then follows that the sizes of S_{ur} and R_{ur} are both $\frac{1}{2}\cdot\frac{n-2}{2}\cdot\left(\frac{n-2}{2}+1\right)$. Then the number of reachable states is at most

$$(n-2)\left(\frac{n-2}{2}+1\right)+k+2-2\cdot\frac{1}{2}\cdot\frac{n-2}{2}\cdot\left(\frac{n-2}{2}+1\right) = \frac{n}{2}\left(\frac{n}{2}-1\right)+k+2.$$

Now, we consider when $n-f$ is odd. By a similar argument, we claim that there is no state (q,j) that is reachable with $j > \lceil\frac{n}{2}\rceil = \frac{n+1}{2}$. We are then left with at most

$$(n-f)(k+1)+k+f-\left(k-\frac{n-f-1}{2}\right)(n-f) = (n-f)\left(\frac{n-f-1}{2}+1\right)+k+f$$

possible reachable states. Again, we consider the size of the sets R_{ur} and S_{ur} and note that the sets are minimized when $f=2$ and have size $\frac{1}{2}\cdot\frac{n-3}{2}\cdot\left(\frac{n-3}{2}+1\right)$. Then the number of reachable states is at most

$$\begin{aligned}&(n-2)\left(\frac{n-3}{2}+1\right)+k+2-2\cdot\frac{1}{2}\cdot\frac{n-3}{2}\cdot\left(\frac{n-3}{2}+1\right)\\ =&(n-2)\left(\frac{n-1}{2}\right)-\left(\frac{n-1}{2}-1\right)\cdot\left(\frac{n-1}{2}\right)+k+2\\ =&\left(\frac{n-1}{2}\right)^2+k+2.\end{aligned}$$

We now show that this bound is reachable by considering the DFA $A_n = (Q_n, \Sigma_n, \delta_n, q_0, F_n)$ constructed in Lemma 4 and shown in Fig. 1. It was shown in Lemma 4 that there are $(n-2)(k+1)-k^2+2$ reachable and distinguishable states when $n > 2k$. We show that there are more unreachable states when $k \geq \frac{n}{2}$ and that this number coincides with the bound obtained above.

Recall from above that if n is even, then $\varphi_{A_n}(q) \leq \frac{n-2}{2}$ for all $q \in Q_n - F_n$ and thus states of the form $(q,j) \in (Q_n - F_n) \times \{\frac{n-2}{2}, \dots, k+1\}$ are unreachable. Then when counting the number of states of S_{ur} and R_{ur}, we consider only those states (q,j) with $1 \leq j \leq \frac{n-2}{2}$ and the number of unreachable states in S_{ur} and R_{ur} is

$$\begin{aligned}&\sum_{i=\frac{n}{2}}^{n-2}\left|\{i\}\times\left\{\varphi_{A_n}(i)+1,\dots,\frac{n-2}{2}\right\}\right|+\sum_{i=1}^{\frac{n-2}{2}}\left|\left\{i+1,\dots,\frac{n-2}{2}\right\}\right|\\ =&2\cdot\sum_{i=1}^{\frac{n-2}{2}}\left|\left\{i,\dots,\frac{n-2}{2}\right\}\right| = 2\cdot\sum_{i=1}^{\frac{n-2}{2}} i = 2\cdot\frac{1}{2}\cdot\frac{n-2}{2}\left(\frac{n-2}{2}+1\right).\end{aligned}$$

Then the number of reachable states is

$$\begin{aligned}&(n-2)(k+1)-2+k-(n-2)\left(k-\frac{n-2}{2}\right)-2\cdot\frac{1}{2}\cdot\frac{n-2}{2}\left(\frac{n-2}{2}+1\right)\\ =&\frac{n}{2}\cdot\left(\frac{n}{2}-1\right)+k+2.\end{aligned}$$

Now, if n is odd, then $\varphi_{A_n}(q) \leq \frac{n-3}{2}$ for all $q \in Q_n - F_n$. Then for the sets of unreachable states S_{ur} and R_{ur}, we consider only those states (q,j) with $1 \leq j \leq \frac{n-3}{2}$, so the number of states in S_{ur} and R_{ur} is

$$\begin{aligned}&\sum_{i=\frac{n-1}{2}+1}^{n-2}\left|\{i\}\times\left\{\varphi_{A_n}(i)+1,\dots,\frac{n-3}{2}\right\}\right|+\sum_{i=1}^{\frac{n-1}{2}-1}\left|\left\{i+1,\dots,\frac{n-3}{2}\right\}\right|\\ =&2\cdot\sum_{i=1}^{\frac{n-1}{2}-1}\left|\left\{i,\dots,\frac{n-3}{2}\right\}\right| = 2\cdot\sum_{i=1}^{\frac{n-1}{2}-1} i = 2\cdot\frac{1}{2}\cdot\frac{n-1}{2}\left(\frac{n-1}{2}-1\right).\end{aligned}$$

This gives a total of

$$(n-2)(k+1)+2+k-(n-2)\left(k-\frac{n-3}{2}\right)-2\cdot\frac{1}{2}\cdot\frac{n-1}{2}\left(\frac{n-1}{2}-1\right)$$
$$=\left(\frac{n-1}{2}\right)^2+k+2$$

reachable states. □

The lower bound construction of Lemma 4 uses, for a DFA with n states, an alphabet of cardinality $n-3$. To conclude this section we show that the construction is optimal in the sense that the upper bound of Theorem 5 cannot be reached with an alphabet of cardinality less than $n-3$.

Proposition 7. *Let A be a DFA recognizing a finite language with n states. If the state complexity of $E(L(A), d_p, k)$ equals $(n-2)(k+1)-k^2+2$, then the alphabet of A needs at least $n-3$ letters.*

Proof. Let $A = (Q, \Sigma, \delta, q_0, F)$ with $|Q| = n$. Let $A' = (Q', \Sigma, \delta', q_0' F')$ be the DFA recognizing $E(L(A), d_p, k)$ constructed in Proposition 2. Recall from the proof of Proposition 3 that in order for A' to have the maximal number of states $(n-2)(k+1)-k^2+2$, a necessary condition is that $F = \{q_0, q_f\}$ and that there can be only one state q_1 with $\varphi_A(q_1) = 1$.

Now for all $q \in Q - \{q_0, q_f, q_1\}$, $\varphi_A(q) \geq 2$. By definition of the transition function δ', if $\varphi_A(q) \geq 2$, the state $(q, 1)$ can only be reached by a direct transition from a final state. Since q_f does not have any outgoing transitions, q_0 must have $n-3$ outgoing transitions—one for each state q.

Furthermore, since A contains a final state q_f with no outgoing transitions, no additional symbols are required to reach p_1, as it can be reached from q_f via a direct transition on any symbol.

Since A is a DFA and q_0 has at least $n-3$ outgoing transitions, the cardinality of the alphabet must be at least $n-3$. □

4. Neighbourhoods of Prefix-Convex Languages

Next, we consider the state complexity of neighbourhoods of prefix-convex languages with respect to the prefix distance. First, we require the following characterization for minimal DFAs recognizing prefix-convex languages from Brzozowski and Sinnamon [2].

Proposition 8 [2]. *Let L be a regular language and $A = (Q, \Sigma, \delta, q_0, F)$ be a minimal DFA recognizing L. The following statements are equivalent:*

(i) *L is prefix-convex.*

(ii) *For all $p, q, r \in Q$, if p and r are final, q is reachable from p, and r is reachable from q, then q is final.*

(iii) Every state reachable in A from any final state is final.

Condition (iii) in the result stated in [2] assumes DFAs are complete and allows the possibility that a state reachable from the final state may be the sink state. We allow DFAs to be incomplete which means that a minimal DFA does not have a sink state.

We will show in the following that the structure of the minimal DFA recognizing a prefix-convex language gives rise to more unreachable states in the DFA obtained via the construction from Proposition 2.

4.1. When the number of non-final states is larger than the radius

First, we begin with stating the bound for the case when the number of non-final states of the given DFA is larger than the radius of the neighbourhood.

Proposition 9. *Let L be a prefix-convex language recognized by an n state DFA A with f final states. Then for $n - f > k > 0$, there is a DFA A' that recognizes the neighbourhood $E(L, d_p, k)$ with at most $(n - f) \cdot k + f + 1 - \frac{k(k-1)}{2}$ states.*

Proof. Let $A' = (Q', \Sigma, \delta', q_0', F')$ be the DFA constructed for the neighbourhood $E(L, d_p, k)$ as in Proposition 2. By Proposition 8, since L is prefix-convex, A has the property that every state reachable from a final state of A must also be a final state. This property creates additional unreachable states in A'.

For all non-final states $q \in Q - F$, the state $(q, 1)$ is reachable only if either $\varphi_A(q) = 1$ or there is a transition from a final state to q. However, since L is prefix-convex, no non-final states are reachable from any final state, thus no final states may have any transitions to a non-final state. Then the only states q where $(q, 1)$ is reachable are those with $\varphi_A(q) = 1$. However, for all such states q, the states (q, i) with $2 \leq i \leq k + 1$ are unreachable. Thus, to reach the upper bound on the number of states, the number of states q with $\varphi_A(q) = 1$ must be minimized if $k \geq 2$. If $k = 1$, then for each state $q \in Q - F$, either $(q, 1)$ is reachable or $(q, k + 1)$ is reachable, so the number of states with $\varphi_A(q) = 1$ need not be minimized.

Then the set of states Q' has $(n - f - 1) \times k + k + f + 1$ elements but they cannot all be reachable. From Proposition 2 (b), elements of the set

$$S_{ur} = \{(q, j) \mid q \in Q - F, 2 \leq j \leq k + 1, j > \varphi_A(q)\}$$

are unreachable as states of A'. Now, recall that the set S_{ur} is minimized when exactly one non-final state q_i in the DFA A for each $1 \leq i \leq k$ has a shortest path of length i that reaches a final state $q_f \in F$. In this case, we have $|S_{ur}| = \frac{k(k-1)}{2}$.

Thus, in order to maximize the number of reachable states of A', the DFA A has a single state q_1 with $\varphi_A(q_1) = 1$. Note that we assume $k \geq 2$ since the case $k = 1$ was already handled above. This gives us at most

$$(n - f - 1) \cdot k + k + f + 1 - \frac{k(k-1)}{2} = (n - f) \cdot k + f + 1 - \frac{k(k-1)}{2}$$

states of A' which are reachable. □

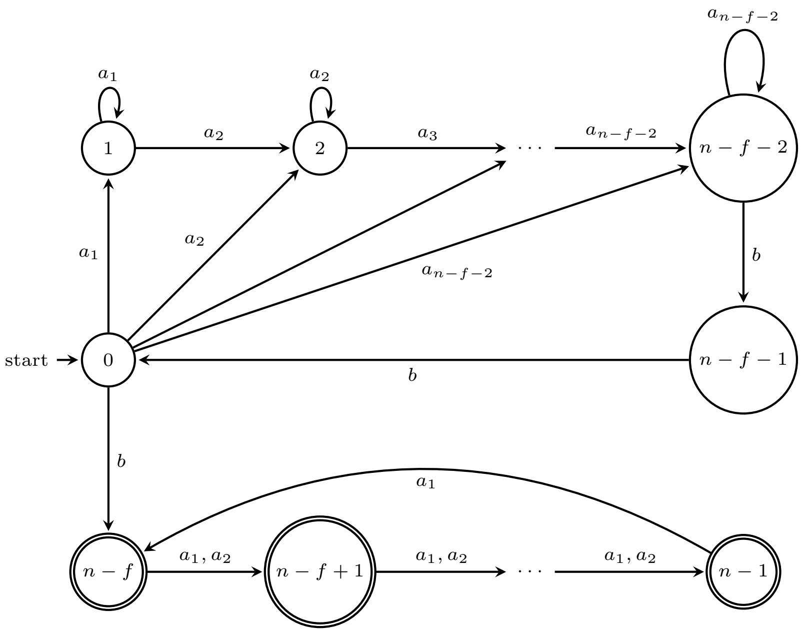

Figure 2: The DFA $A_{n,f}$.

Next we present a lower bound construction that matches the bound of Proposition 9.

Lemma 10. *There exists a DFA $A_{n,f}$ with n states and f final states recognizing a prefix-convex language such that a DFA recognizing the neighbourhood $E(L(A_{n,f}), d_p, k)$ requires at least $(n-f) \cdot k + f + 1 - \frac{k(k-1)}{2}$ states.*

Proof. We define the DFA $A_{n,f} = (Q_n, \Sigma_{n,f}, \delta_{n,f}, q_0, F_f)$ by setting

- $Q_n = \{0, \ldots, n-1\}$,
- $\Sigma_{n,f} = \{a_1, \ldots, a_{n-f-3}, b\}$,
- $F_f = \{n-f, \ldots, n-1\}$,
- $q_0 = 0$,

and the transition function $\delta_{n,f}$ is given by

- $\delta_{n,f}(0, a_i) = i$ for $i = 1, \ldots, n-f-2$,
- $\delta_{n,f}(i, a_i) = i$ for $i = 1, \ldots, n-f-2$,
- $\delta_{n,f}(i, a_{i+1}) = i+1$ for $i = 1, \ldots, n-f-3$,
- $\delta_{n,f}(n-f-2, b) = n-f-1$, $\delta_{n,f}(n-f-1, b) = 0$, $\delta_{n,f}(0, b) = n-f$,
- $\delta_{n,f}(n-j, a_1) = \delta_{n,f}(n-j, a_2) = n-j+1$ for $2 \leq j \leq f$,
- $\delta_{n,f}(n-1, a_1) = n-1$.

The DFA $A_{n,f}$ is shown in Figure 2.

We transform $A_{n,f}$ into the DFA $A'_{n,f} = (Q'_n, \Sigma_{n,f}, \delta_{n,f}, q_0, F_f)$ via the construction from Proposition 2. To determine the reachable states of Q'_n, we first note that the state $(0,1)$ is reachable as it is the initial state. Note that the initial state is $(0,1)$ since $\varphi_{A_{n,f}}(0) = 1$. Each final state $n-j$ for $1 \leq j \leq f$ is reachable on the word ba_2^{j-1} from the initial state $(0,1)$. Now consider states $p_1, \ldots, p_k$. The state p_ℓ is reachable on the word $b^{\ell+1}$.

Now consider states of the form $(i,j) \in (Q_n - \{0, n-1\}) \times \{2, \ldots, k+1\}$. Recall that states $(i,1)$ are unreachable for any state $i \in Q_n$ with $\varphi_{A_{n,f}} > 1$. Then for states $i \in Q_n$ with $\varphi_{A_{n,f}}(i) > k$ and each $2 \leq j \leq k+1$, we can reach state (i,j) from $(0,1)$ via the word a_i^{j-1}. For states $i \in Q_n$ with $\varphi_{A_{n,f}}(i) \leq k$, we can reach state (i,j) via the word a_i^{j-1} for $j = 2, \ldots, \varphi_{A_{n,f}}(i)$ and states (i,j) with $j > \varphi_{A_{n,f}}(i)$ are unreachable by definition of $A'_{n,f}$.

Finally, we can reach state $(n-f-1, 2)$ via the word $a_{n-f-2}b$ and states $(n-f-1, j)$ are unreachable for $j > 2$ since $\varphi_{A_{n,f}}(n-f-1) = 2$. Thus the number of unreachable states in $(Q_n - \{0, n-1\}) \times \{2, \ldots, k+1\}$ is

$$\sum_{i=n-f-k}^{n-f-1} |\{i\} \times \{\varphi_{A_{n,f}}(i)+1, \ldots, k+1\}| = \sum_{i=1}^{k} |\{i+1, \ldots, k+1\}| = \sum_{i=1}^{k} i = \frac{k(k-1)}{2}.$$

Now, we show that all reachable states are pairwise inequivalent. First, we show that no states $q \in F$ are equivalent to any state of the form (i,j) in $A'_{n,f}$. From q, reading the word b^k takes the machine to the state p_k. However, for $1 \leq i \leq n-f-3$, reading the word b^k from (i,j) takes the machine to p_k on the prefix b^{k-j} and the computation fails since p_k has no outgoing transitions. For $i = 0, n-f-1, n-f-2$, we distinguish (i,j) from q by the word a_3^k using the same argument as above.

Next, we distinguish states of the form (i,j) from states of the form p_ℓ. For each $1 \leq i \leq n-f-2$, reading the word a_i^k from state (i,j) takes the machine to state $(i, \min\{\varphi_{A_{n,f}}(i), k+1\})$. Then subsequently reading $a_{i+1}a_{i+2}\cdots a_{n-f-2}bbb$ takes the machine to the final state $n-f$. However, for every state p_ℓ, reading a_i^k forces the machine beyond state p_k, after which there are no transitions defined. The state $(n-f-1, 2)$ is distinguished from all p_ℓ by the word b^{2+k}, $(0,1)$ by b^{1+k}.

Next, without loss of generality, let $\ell < \ell'$ and consider states p_ℓ and $p_{\ell'}$. Choose $z = b^{k-\ell}$. The word z takes state p_ℓ to the state p_k, where it is accepted. However, the computation on word z from state $p_{\ell'}$ is undefined since $\ell' + k - \ell > k$.

Now consider states of the form $(i,j) \in (Q-F) \times \{1, \ldots, k+1\}$. Let $i < i'$ and consider the states (i,j) and (i',j'). Let $z = a_{i+1}a_{i+2}\cdots a_{n-f-2}bbbb^k$. From state (i,j), the word z goes to state $n-1$ on $a_{i+1}\cdots a_{n-f-2}bbb$. Then by reading b^k from state $n-1$, we reach state p_k, an accepting state. However, when reading z from state (i',j'), we immediately reach state $p_{j'+1}$ on a_{i+1}, since the transition on a_{i+1} is defined only for states $(0,1)$ and (i,j). Since the rest of the word z is of length greater than k, reading it takes us to state p_k with no further defined transitions for the rest of the word.

Next, consider the two states (i,j) and (i,j'), where $j < j'$. First, consider the case when $\varphi_{A_n}(i) > k$. Then let $z = a_i^{k-j}$. Reading z from (i,j) takes us to state (i,k), which is a final state. However, from (i, j'), reading z brings us to state $(i, k+1)$ and so the computation is rejected.

Now, consider the case when $\varphi_{A_n}(i) \leq k$. Let $z = bb^{k-j-1}$. From state (i, j), reading b takes the machine to state p_{j+1} and reading b^{k-j-1} puts the machine in the accepting state p_k. However, reading z from (i, j') takes us to state p_k with $b^{j'-j}$ still unread since $j' + k - j - 1 > k$ and thus, with no further transitions available, the computation is rejected.

Finally, we consider two states $i, i' \in F$. Without loss of generality, let $i < i'$ and consider the word $a_2^{n-i}b^k$. Then $\delta'_{n,f}(i, a_2^{n-i}) = n - f$. But reading a_2^{n-i} from i' brings the machine to state $n-1$ on the prefix $a_2^{n-i'}$. Since a_2 is undefined from $n-1$, the machine goes to p_1 and the computation fails before the machine can finish reading the rest of the input.

Thus, we have shown that there are $(n-f) \cdot k + f + 1 - \frac{k(k-1)}{2}$ reachable states and that all reachable states are pairwise inequivalent. □

Combining Proposition 9 and Lemma 10 we have:

Theorem 11. *Let L be a prefix-convex language. For $n > k \geq 0$, if $\mathrm{sc}(L) = n$, then*

$$\mathrm{sc}(E(L, d_p, k)) \leq (n-f) \cdot k + f + 1 - \frac{k(k-1)}{2},$$

and this bound can be reached in the worst case.

The construction of Lemma 10 that establishes the lower bound for Theorem 11 uses an alphabet of size $n-f-2$, where n is the number of states of the DFA and f is the number of final states. The below result establishes that the size of the alphabet cannot be reduced.

Proposition 12. *Let A be a DFA recognizing a prefix-convex language with n states and f final states. If for $n > k > 0$ the state complexity of $E(L(A), d_p, k)$ is $(n-f) \cdot k + f + 1 - \frac{k(k-1)}{2}$, then the alphabet of A requires at least $n - f - 2$ letters.*

Proof. Let $A = (Q, \Sigma, \delta, q_0, F)$ with $|Q| = n$ and $|F| = f$. Let $A' = (Q', \Sigma, \delta', q_0', F')$ be the DFA recognizing $E(L(A), d_p, k)$ constructed in Proposition 2. Recall that since A recognizes a prefix-convex language, no non-final states are reachable from a final state. Recall also from the proof of Proposition 9 that in order for A' to have the maximal number of states $(n-f) \cdot k + 1 + f - \frac{k(k-1)}{2}$, a necessary condition is that there can be only one state q_1 with $\varphi_A(q_1) = 1$ and one state q_2 with $\varphi_A(q_2) = 2$.

Now for all $q \in Q - (F \cup \{q_1, q_2\})$, $\varphi_A(q) \geq 3$. Recall that since no final states have outgoing transitions to any non-final states, states $(q, 1)$ are reachable only if $\varphi_A(q) = 1$. Then by definition of the transition function δ', if $\varphi_A(q) \geq 3$, the state $(q, 2)$ can only be reached by a direct transition from a state q with $\varphi_A(q) = 1$. Thus, q_1 must have $n-f-2$ outgoing transitions—one for each state q with $\varphi_A(q) \geq 3$. and one additional transition to a final state. Note that q_2 requires no direct transition from q_1 since $\varphi_A(q_2) = 2$ and thus $(q_2, 2)$ is the only reachable state of the form (q_2, j).

Since A is a DFA and q_1 has at least $n-f-2$ outgoing transitions, the cardinality of the alphabet must be at least $n-f-2$. □

From the above result, we can derive state complexity bounds for subclasses of prefix-convex languages. First, we need the following characterization from Brzozowski and Sinnamon [2].

Proposition 13 [2]. *Let L be a non-empty prefix-convex language and let $A = (Q, \Sigma, \delta, q_0, F)$ be a minimal DFA recognizing L.*

(i) *L is prefix-closed if and only if $q_0 \in F$.*

(ii) *L is prefix-free if and only if A has a unique final state p with no outgoing transitions.*

(iii) *L is a right ideal if and only if A has a unique final state p and $\delta(p,a) = p$ for all $a \in \Sigma$.*

In fact, (ii) gives a characterization for the prefix-free languages if L were a regular language, not just a prefix-convex language. However, it is not difficult to see that any language L that satisfies (ii) is a prefix-convex language by definition.

Theorem 14. *Let L be a prefix-free regular language recognized by a minimal n-state DFA A. Then for $n > k > 0$, there is a DFA A' with at most $(n-1)k + 2 - \frac{k(k-1)}{2}$ states that recognizes the neighbourhood $E(L, d_p, k)$ and this bound is reachable.*

Proof. Let $A' = (Q', \Sigma, \delta, q_0', F')$ be the DFA constructed for the neighbourhood $E(L, d_p, k)$ as in Proposition 2. Since L is prefix-free, it is a prefix-convex language and by Proposition 13, A must have a unique final state with no outgoing transitions. Then by Proposition 9, we have $f = 1$ and thus A' has at most $(n-1)\cdot k + 2 - \frac{k(k-1)}{2}$ states.

To show that this bound is reachable, we define a DFA $B_n = (Q_n, \Sigma_n, \gamma_n, 0, F)$ by modifying the DFA $A_{n,f}$ from Lemma 10 by setting $F = \{n-1\}$ and $\gamma_n(n-1, a) = \emptyset$ for all $a \in \Sigma$. The resulting DFA is shown in Figure 3.

We obtain the DFA B_n' by following the construction from Proposition 2. Then B_n' will have $(n-1)\cdot k + 2 - \frac{k(k+1)}{2}$ states. Using a similar, but simpler, argument as in the proof of Lemma 10, it is seen that all states are reachable and pairwise inequivalent. □

Theorem 15. *Let L be a right ideal recognized by a minimal n state DFA A. Then for $n > k > 0$, there is a DFA A' with at most $(n-1)\cdot k + 1 - \frac{k(k-1)}{2}$ states that recognizes the neighbourhood $E(L, d_p, k)$ and this bound is reachable.*

Proof. Since L is a right ideal, by Proposition 13, it is a prefix-convex language and A must have a unique final state $q_f \in F$ such that $\delta(q_f, a) = q_f$ for all $a \in \Sigma$. We obtain a DFA A' by following the construction in Proposition 2. However, since the sole final state q_f has no undefined transitions, the state p_1 is unreachable in A'. Since $f = 1$, Proposition 9 implies that B has at most $(n-1)\cdot k + 1 - \frac{k(k-1)}{2}$ states.

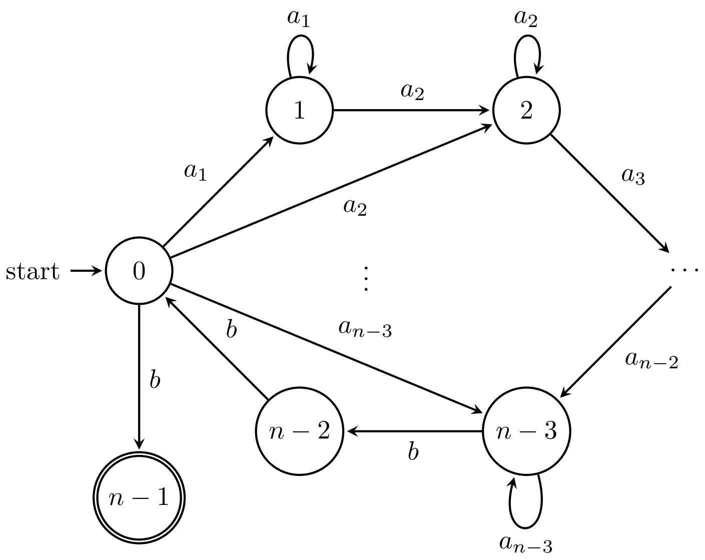

Figure 3: The DFA B_n.

To show that this bound is reachable, we define the DFA $C_n = (Q_n, \Sigma \cup \{c\}, \rho_n, 0, F)$, where $Q_n = \{0, \ldots, n-1\}$, $F = \{n-1\}$, and ρ_n is the same as the transition function γ_n of the DFA B_n in the proof of Theorem 14 except that we set $\rho_n(n-1, a) = n-1$ for all $a \in \Sigma$. We note that we add an extra symbol $c \notin \Sigma$ to the alphabet to ensure there is at least one undefined transition from state 0. We obtain the DFA C'_n by following the construction from Proposition 2. It then follows from the proof of Theorem 14 and the above argument that there are $(n-1) \cdot k + 1 - \frac{k(k-1)}{2}$ reachable states in C'_n and that all reachable states are pairwise inequivalent. □

Finally, we give a tight bound for the state complexity of prefix distance neighbourhoods of prefix-closed languages.

Theorem 16. *Let L be a prefix-closed regular language recognized by an n-state DFA A. Then there is a DFA A' that recognizes the neighbourhood $E(L, d_p, k)$ with at most $n + k$ states and this bound is reachable.*

Proof. We recall that by Proposition 13, a prefix-closed language is also prefix-convex and the initial state of A must be a final state. Since L is prefix-convex, by Proposition 8 every state reachable from the initial state must also be a final state. These two properties imply that every state of A must be a final state. If A has n states, this means that the DFA A' constructed in Proposition 2 for the radius k neighbourhood has $n + k$ states.

We now define a prefix-closed regular language L_n such that a DFA recognizing $E(L_n, d_p, k)$ requires at least $n + k$ states. Let $L_n = \{a^i \mid 0 \leq i \leq n\}$. Then we define $A_n = (Q_n, \{a, b\}, \delta_n, q_0, F_n)$ where $Q_n = F_n = \{0, \dots, n-1\}$, $q_0 = 0$, and the transition function δ_n is defined by $\delta_n(i, a) = i + 1$ for $0 \leq i \leq n - 1$.

Then we define the DFA recognizing $E(L_n, d_p, k)$ by $A' = (Q'_n, \{a, b\}, \delta'_n, q_0, F'_n)$ where $Q'_n = F'_n = Q_n \cup \{p_1, \dots, p_k\}$ and the transition function defined by

- $\delta'_n(i, a) = i + 1$ for $0 \leq i < n - 1$,
- $\delta'_n(n - 1, a) = p_1$,
- $\delta'_n(i, b) = p_1$ for $0 \leq i < n - 1$,
- $\delta'_n(p_i, a) = \delta'_n(p_i, b) = p_{i+1}$ for $1 \leq i < k$.

Every state i, $0 \leq i \leq n-1$, is reachable on the word a^i and every state p_i, $1 \leq i \leq k$ is reachable on the word b^i. The states $0 \leq i, i' \leq n-1$ are distinguished by the word b^{k-i} and the states p_i, p'_i, $1 \leq i, i' \leq k$ are also distinguished by the word b^{k-i}. The states i, $0 \leq i \leq n-1$ and p_j, $1 \leq j \leq k$ are distinguished by the word $a^{n-j}b^k$. Thus, there are $n + k$ reachable states and they are all pairwise distinguishable. □

4.2. When the radius is at least the number of non-final states

Now, we consider the state complexity of prefix distance neighbourhoods when $k \geq n - f$. We show that there are fewer reachable states and that the bound is reachable with the same via the same witness as in Lemma 10.

Theorem 17. *Let L be a prefix-convex language recognized by an n state DFA A with f final states. Then for $k \geq n - f > 0$, there is a DFA A' that recognizes the neighbourhood $E(L, d_p, k)$ with at most $\frac{(n-f-1)(n-f)}{2} + k + f + 1$ states. Furthermore, this bound is reachable.*

Proof. We begin by giving an estimation for the number of states of A' without assuming that L is prefix-convex. Let L be recognized by a DFA $A = (Q, \Sigma, \delta, q_0, F)$ and follow the construction from Proposition 2. This gives us a DFA A' with $(n - f)(k + 1) + k + f$ states. Recall that all elements of the set

$$S_{ur} = \{(q, j) \mid q \in Q - F, 1 \leq j \leq k + 1, j > \varphi_A(q)\}$$

are unreachable as states of A'. By the definition of φ_A, if for some state p, the value $\varphi_A(p) = \ell \geq 2$, then there must exist another non-final state p' with $\varphi_A(p') = \ell - 1$. Now, we observe that there is a path from every state of A to a final state and that the length of the shortest such path for each state is at most $n - f < k$.

Then the cardinality of S_{ur} is minimized when for each $1 \leq i \leq n - f$, exactly one non-final state q_i has a shortest path to a final state of length i. In this case, $S_{ur} = \{(q_i, j) \mid i < j \leq k + 1, i = 1, \dots, n - f\}$ and

$$|S_{ur}| = \sum_{i=1}^{n-f} (k + 1 - i) = (n - f)(k + 1) - \frac{(n - f)(n - f + 1)}{2}.$$

Then the total number of reachable states is

$$(n - f)(k + 1) + k + f - |S_{ur}| = \frac{(n - f)(n - f + 1)}{2} + k + f.$$

Now, suppose that L is a prefix-convex language. Recall that for a prefix-convex language, states $(q,1)$ with $q \in Q - F$ are unreachable except when $\varphi_A(q) = 1$. Thus, the number of unreachable states is

$$|S_{ur}| = \sum_{i=1}^{n-f-1} (k-i) = (n-f-1) \cdot k - \frac{(n-f-1)(n-f)}{2}.$$

This gives us a total of at most $\frac{(n-f-1)(n-f)}{2} + k + f + 1$ reachable states.

Now, we show that this bound is reachable. Consider the DFA $A_{n,f} = (Q_n, \Sigma_{n,f}, \delta_{n,f}, q_0, F_f)$ from Lemma 10, shown in Figure 2. Since $k \geq n - f$, we have that $\varphi_{A_{n,f}}(i) \leq k$ for all $i \in Q_n - F_f$ and we can reach (i,j) for $j = 2, \ldots, \varphi_{A_{n,f}}(i)$. This gives us at least $(n-f)k + f + 1$ reachable states. However, states (i,j) with $j > \varphi_{A_{n,f}}(i)$ are unreachable by definition and the number of unreachable states in S_{ur} is

$$\begin{aligned}\sum_{i=1}^{n-f-1} |\{i\} \times \{\varphi_{A_{n,f}}(i)+1, \ldots, k\}| &= \sum_{i=1}^{n-f-1} |\{i+1, \ldots, k\}| \\ = \sum_{i=1}^{n-f-1} (k-i) = (n-f-1) \cdot k - \frac{(n-f-1)(n-f)}{2}&.\end{aligned}$$

This gives us a total of $\frac{(n-f-1)(n-f)}{2} + k + f + 1$ reachable states and these states are all pairwise distinguishable as in Lemma 10. □

From the preceding proof, we also get the following result for regular languages in general.

Corollary 18. *Let L be a regular language recognized by a DFA with n states and f final states. For $k \geq n - f > 0$, there is a DFA recognizing $E(L, d_p, k)$ with at most $\frac{(n-f)(n-f+1)}{2} + k + f$ states.*

Theorem 17 gives us the following corollary concerning prefix-free languages and right ideals.

Corollary 19. *Let L be a prefix-convex language recognized by an n state DFA A with f final states. For $k \geq n - f > 0$, let A' be a DFA that recognizes the neighbourhood $E(L, d_p, k)$.*

(i) If L is a prefix-free language, then A' has at most $\frac{(n-1)(n-2)}{2} + k + 2$ states.

(ii) If L is a right ideal, then A' has at most $\frac{(n-1)(n-2)}{2} + k + 1$ states.

Proof.

(i) By Proposition 13, a prefix-free language has a single final state with no outgoing transitions. Following Theorem 17, we have $f = 1$ and the bound follows.

(ii) By Proposition 13, a right ideal has a single final state $q_f \in F$ with transitions $\delta(q_f, a) = q_f$ for all $a \in \Sigma$. Then the state p_1 is unreachable and by Theorem 9, we have $f = 1$ and the bound follows.

□

5. Conclusion

We have given tight state complexity bounds for the prefix-distance neighbourhood of, respectively, finite, prefix-convex, prefix-closed, prefix-free, and right ideal languages. As can, perhaps, be expected the bound for prefix-closed languages is relatively easier to obtain and the matching lower bound construction uses a binary alphabet. The upper bound constructions for the finite and the prefix-convex languages are more involved and the lower bound constructions use a variable size alphabet. Furthermore, we have shown that, in both cases, the alphabet size is optimal.

Since the reversal of a DFA is not, in general, deterministic, the state complexity bounds for suffix-distance (or factor-distance) neighbourhoods differ significantly from the corresponding bounds for prefix-distance neighbourhoods. Tight lower bounds are not known for suffix-distance neighbourhoods of general regular languages [16] or for various sub-regular language families. Such questions can be a topic for further research.

Acknowledgements

We thank Janusz Brzozowski for pointing out the related work on prefix-convex languages.

References

[1] Bordihn, H., Holzer, M., Kutrib, M.: Determination of finite automata accepting subregular languages. Theor. Comput. Sci. 410(35) (2009) 3209-3222

[2] Brzozowski, J., Sinnamon, C.: Complexity of Prefix-Convex Regular Languages. (2016). arXiv:1605.06697

[3] Calude, C.S., Salomaa, K., Yu, S.: Additive distances and quasi-distances between words. J. Univ. Comput. Sci. 8(2) (2002) 141–152

[4] Câmpeanu, C., Culik II, K., Salomaa, K., Yu, S.: State complexity of basic operations on finite languages. Proc. WIA'99 (1999) 60–70

[5] Choffrut, C., Pighizzini, G.: Distances between languages and reflexivity of relations. Theoretical Computer Science **286**(1) (2002) 117–138

[6] Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer Berlin Heidelberg (2009).

[7] Gao, Y., Moreira, N., Reis, R., Yu, S.: A review on state complexity of individual operations. Faculdade de Ciencias, Universidade do Porto, Technical Report DCC-2011-8 `www.dcc.fc.up.pt/dcc/Pubs/TReports/TR11/dcc-2011-08.pdf` To appear in *Computer Science Review.*

[8] Han, Y.S., Salomaa, K., Wood, D.: State Complexity of Prefix-Free Regular Languages. In: Proceedings of the 8th International Workshop on Descriptive Complexity of Formal Systems. (2006) 165–176.

[9] Herrmann, A., Kutrib, M., Malcher, A., Wendlandt, M.: Descriptional complexity of bounded regular languages. In: Proc. of DCFS'16. Lect. Notes Comput. Sci. 9777, Springer-Verlag (2016) 138–152.

[10] Holzer, M., Jakobi, S., Kutrib, M.: The Magic Number Problem for Subregular Language Families. Int. J. Found. Comput. Sci. 23(1) (2012) 115-131.

[11] Holzer, M. Kutrib, M.: Descriptional and computational complexity of finite automata — A survey. *Inform. Comput.* **209** (2011) 456–470.

[12] Holzer, M., Kutrib, M., Meckel, K.: Nondeterministic state complexity of star-free fanguages. Proc. CIAA 2011 (2011) 178-189.

[13] Holzer, M., Truthe, B.: On relations between some subregular language families. Proc. NCMA 2015 (2015) 109-124.

[14] Kao, J.Y., Rampersad, N., Shallit, J.: On NFAs where all states are final, initial, or both. Theoretical Computer Science **410**(47-49) (nov 2009) 5010–5021.

[15] Kutrib, M., Pighizzini, G.: Recent trends in descriptional complexity of formal languages. *Bulletin of the EATCS* 111 (2013) 70–86.

[16] Ng, T., Rappaport, D., Salomaa, K.: State Complexity of Prefix Distance. In: Implementation and Application of Automata (CIAA 2015). (2015) 238–249.

[17] Shallit, J.: A second course in formal languages and automata theory. Cambridge University Press, Cambridge, MA (2009).

[18] Thierrin, G.: Convex Languages. In: Nivat, M. (ed.) International Colloquium on Automata, Languages and Programming. pp. 481–492 North-Holland Publishing Company, Paris (1972).

[19] Yu, S.: Regular languages. In Rozenberg, G., Salomaa, A., eds.: Handbook of Formal Languages. Springer-Verlag, Berlin, Heidelberg (1997) 41–110.