State Complexity of Prefix Distance

Timothy Ng, David Rappaport, and Kai Salomaa

School of Computing, Queen's University
CIAA 2015, Umeå, Sweden

August 21, 2015

Are neighbourhoods of a regular language also regular? What is the state complexity of the neighbourhood of a regular language?

- Additive distances are regularity preserving (Calude, Salomaa, Yu 2002)

- Additive distances are regularity preserving (Calude, Salomaa, Yu 2002)
- The state complexity of these neighbourhoods is $(k+2)^n$
    - Upper bound (Salomaa, Schofield 2007)
    - Lower bound (Ng, Rappaport, Salomaa 2015)

- Additive distances are regularity preserving (Calude, Salomaa, Yu 2002)
- The state complexity of these neighbourhoods is $(k+2)^n$
  - Upper bound (Salomaa, Schofield 2007)
  - Lower bound (Ng, Rappaport, Salomaa 2015)
- Asymptotic lower bounds for neighbourhoods with respect to Hamming distance
  - $r = 1$ (Povarov 2007)
  - $r > 1$ (Shamkin 2011)

1. Tight state complexity bounds for neighbourhoods with respect to the prefix distance.

1. Tight state complexity bounds for neighbourhoods with respect to the prefix distance.
2. Tight nondeterministic state complexity bounds for neighbourhoods with respect to the prefix, suffix, and substring distances.

A distance is a function $d : \Sigma^* \times \Sigma^* \to [0, \infty)$ such that

1. $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, w) + d(w, y)$

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.

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

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.

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

Analogously, we can define the suffix distance d_s and substring distance d_f.

Harbord $\rightarrow$ Harbourfront

Harbord $\rightarrow$ Harbourfront

Kingston$\rightarrow$ Eglinton

Harbord $\rightarrow$ Harbourfront

Kingston $\rightarrow$ Eglinton

Church $\rightarrow$ Bathurst

The neighbourhood of a language $L \subseteq \Sigma^*$ of radius $k \geq 0$ with respect to a distance measure d is the set of all words u with $d(w, u) \leq k$ for some $w \in L$,

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

Theorem

For a regular language $L \subseteq \Sigma^*$ recognized by an NFA with n states and an integer $k \geq 0$,

$$\text{nsc}(E(L, d_p, k)) \leq n + k.$$

- Let $A = (Q, \Sigma, \delta, q_0, F)$.
- Let $\varphi(q)$ be the length of the shortest word w such that $\delta(q, w) \cap F \neq \emptyset$.

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

- $Q' = Q \cup \{p_1, \ldots, p_k\}$

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

- $Q' = Q \cup \{p_1, \ldots, p_k\}$
- $F' = F \cup \{p_q, \ldots, p_k\} \cup \{q \in Q \mid \varphi(q) \leq k\}$

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

- $Q' = Q \cup \{p_1, \ldots, p_k\}$
- $F' = F \cup \{p_q, \ldots, p_k\} \cup \{q \in Q \mid \varphi(q) \leq k\}$
- $\delta'(q, a) = \delta(q, a) \cup \{p_1\}$ for all $q \in F$,

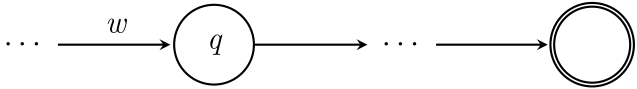
Let $A' = (Q', \Sigma, \delta', q_0', F')$.

- $Q' = Q \cup \{p_1, \dots, p_k\}$
- $F' = F \cup \{p_q, \dots, p_k\} \cup \{q \in Q \mid \varphi(q) \leq k\}$
- $\delta'(q, a) = \delta(q, a) \cup \{p_1\}$ for all $q \in F$,
- $\delta'(q, a) = \delta(q, a) \cup \{p_{\varphi(q)+1}\}$ for all $q \in Q$ with $\varphi(q) < k$,

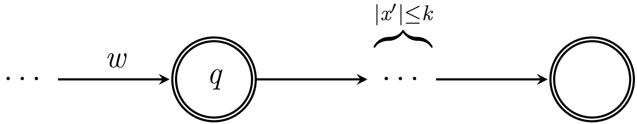
Let $A' = (Q', \Sigma, \delta', q_0', F')$.

- $Q' = Q \cup \{p_1, \dots, p_k\}$
- $F' = F \cup \{p_q, \dots, p_k\} \cup \{q \in Q \mid \varphi(q) \leq k\}$
- $\delta'(q, a) = \delta(q, a) \cup \{p_1\}$ for all $q \in F$,
- $\delta'(q, a) = \delta(q, a) \cup \{p_{\varphi(q)+1}\}$ for all $q \in Q$ with $\varphi(q) < k$,
- $\delta'(p_i, a) = p_{i+1}$ for $i = 1, \dots, k - 1$.

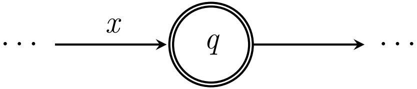
If $x = wx'$ with $x' \in \Sigma^*$,



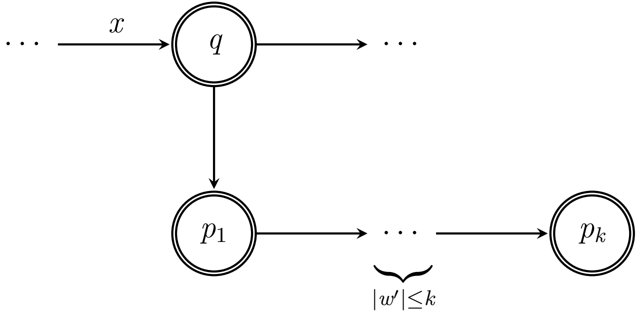
If $x = wx'$ with $x' \in \Sigma^*$,



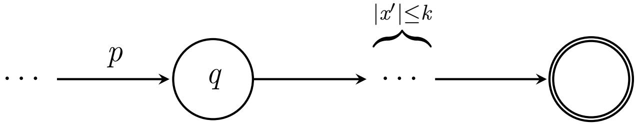
If $w = xw'$ with $w' \in \Sigma^*$,



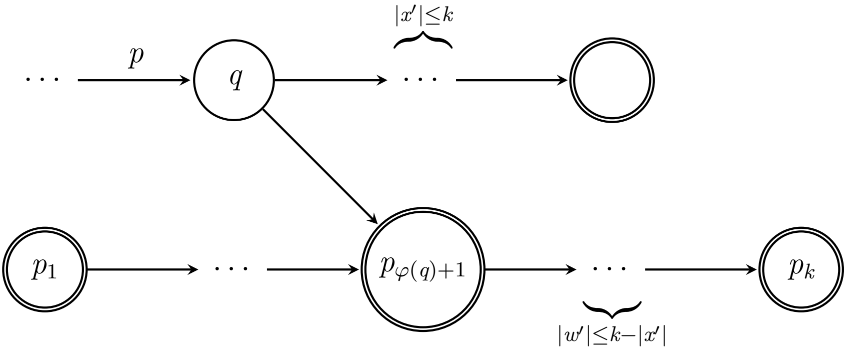
If $w = xw'$ with $w' \in \Sigma^*$,



If $w = pw'$ and $x = px'$ with $p, w', x' \in \Sigma^*$,



If $w = pw'$ and $x = px'$ with $p, w', x' \in \Sigma^*$,



Theorem

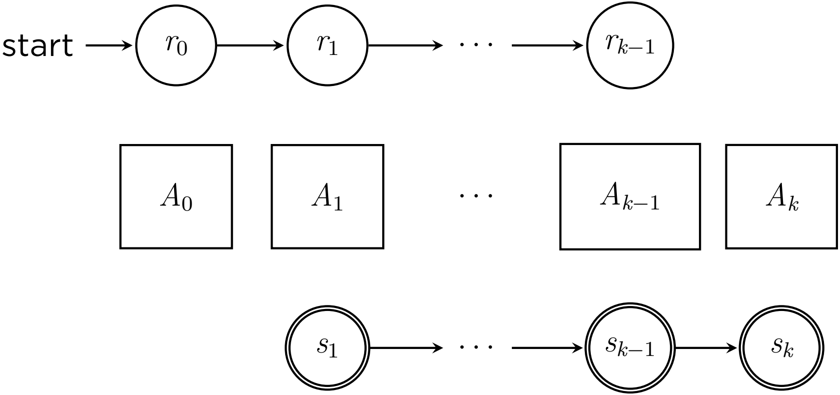
For a regular language $L \subseteq \Sigma^*$ recognized by an NFA with n states and an integer $k \geq 0$,

$$\operatorname{nsc}(E(L, d_s, k)) \leq n + k.$$

Theorem

If L has an NFA with n states and $k \in \mathbb{N}_0$,

$$\mathrm{nsc}(E(L, d_f, k)) \leq (k+1) \cdot n + 2k.$$


start
r_0
r_1
$\cdots$
r_{k-1}
A_0
A_1
$\cdots$
A_{k-1}
A_k
s_1
$\cdots$
s_{k-1}
s_k

- Let $A = (Q, \Sigma, \delta, q_0, F)$.
- Recall that $\varphi(q)$ is the length of the shortest word w such that $\delta(q, w) \notin F$.

Let $A' = (Q', \Sigma, \delta', q'_0, F')$.

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

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

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

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

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

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

Let $q_{i,a} = \delta(i, a)$ for $i \in Q$ and $a \in \Sigma$, if $\delta(i, a)$ is defined.

Let $A' = (Q', \Sigma, \delta', q_0', F')$.

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

Let $q_{i,a} = \delta(i, a)$ for $i \in Q$ and $a \in \Sigma$, if $\delta(i, a)$ is defined.
First, for states $p_\ell, \ell = 1, \ldots, k-1$, we have
$\delta'(p_\ell, a) = p_{\ell+1}$.

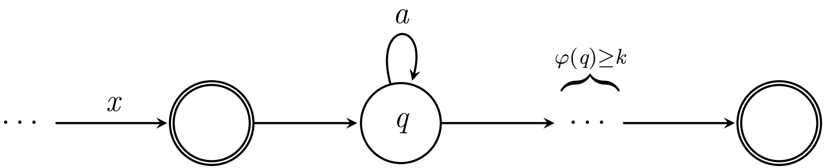
For final states, we have

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

For states $(i, j) \in Q - F \times \{1, \ldots, k+1\}$, we have

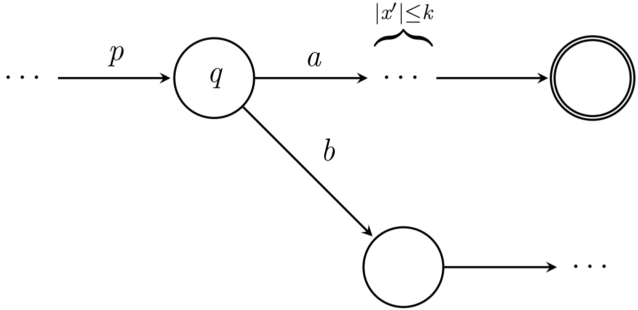
$$\delta'((i,j),a) =$$

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



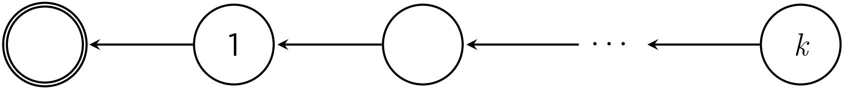
a
$\varphi(q) \geq k$
x
q

If $w = pw'$ and $x = px'$ with $p, w', x' \in \Sigma^*$,



This gives us $(n-f)\cdot(k+1)+k+f$ states in total, however not all of these states are reachable.

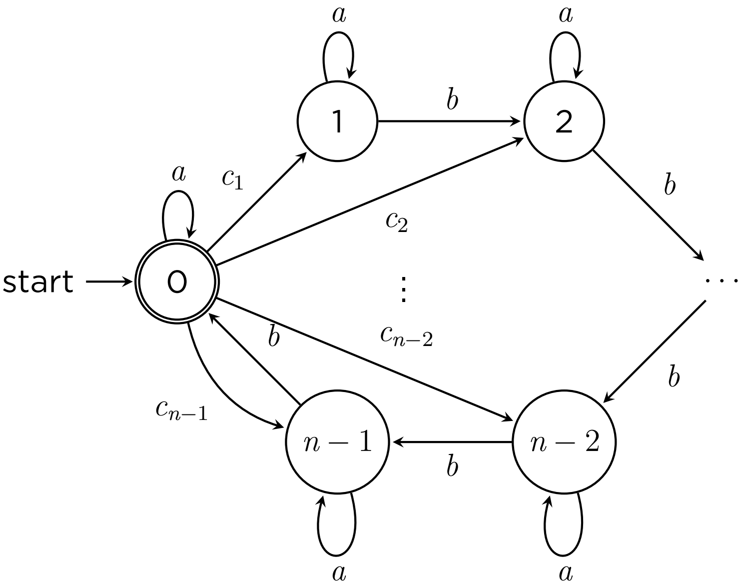
This gives us $(n-f)\cdot(k+1)+k+f$ states in total, however not all of these states are reachable.



Theorem

For $n > k \geq 0$*, if* $\mathrm{sc}(L) = n$ *then*

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


start
0
1
2
$n-1$
$n-2$
a
b
c_1
c_2
c_{n-2}
c_{n-1}
$\cdots$
$\vdots$

In summary,

1. Nondeterministic state complexity of $n + k$ for prefix and suffix neighbourhoods
2. Nondeterministic state complexity of $(k+1) \cdot n + 2k$ for substring neighbourhoods
3. Deterministic state complexity of $(k+1) \cdot n - \frac{k(k+1)}{2}$ for prefix neighbourhoods

Future work:

- DFA constructions for suffix and substring neighbourhoods
- Lower bound examples for suffix and substring neighbourhoods
- Properties of regularity-preserving distances