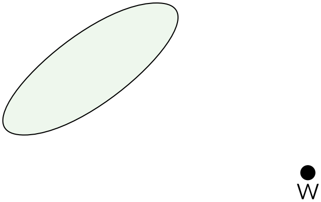
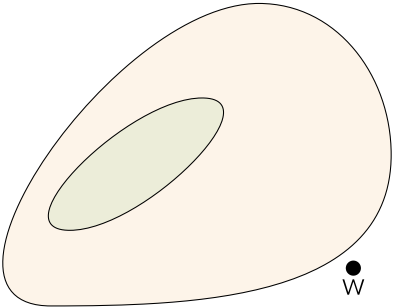
State Complexity of Prefix Distance of Subregular Languages

Timothy Ng David Rappaport Kai Salomaa

School of Computing, Queen's University, Kingston, Canada

DCFS 2016, Bucharest, Romania


W


W

We are interested in the state complexity of neighbourhoods of subregular language classes with respect to the prefix distance.

We are interested in the state complexity of neighbourhoods of subregular language classes with respect to the prefix distance. We show tight state complexity bounds for the following classes:

- finite languages
- prefix-closed regular languages
- prefix-free regular languages

We are interested in the state complexity of neighbourhoods of subregular language classes with respect to the prefix distance. We show tight state complexity bounds for the following classes:

- finite languages
- prefix-closed regular languages
- prefix-free regular languages

The state complexity for these classes is strictly less than for regular languages.

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 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 $w \in L$,

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

- 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)

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.

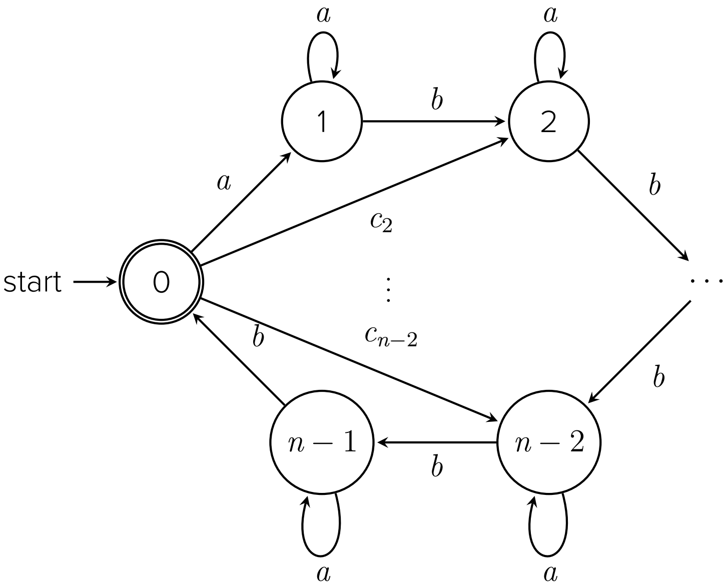
$$d_p(x, y) = |x| + |y| - 2 \cdot \max_{z \in \Sigma^*} \{|z| \mid x, y \in z\Sigma^*\}.$$

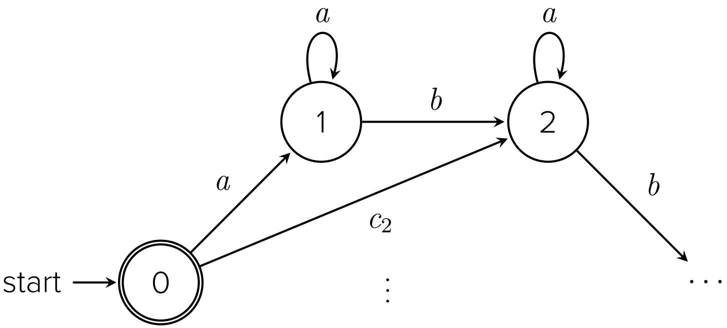
Harbord → Harbourfront

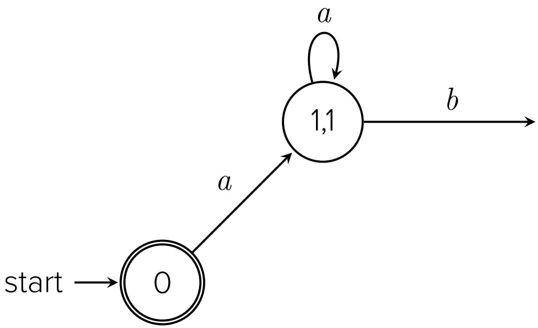
Theorem (Ng, Rappaport, Salomaa 2015)

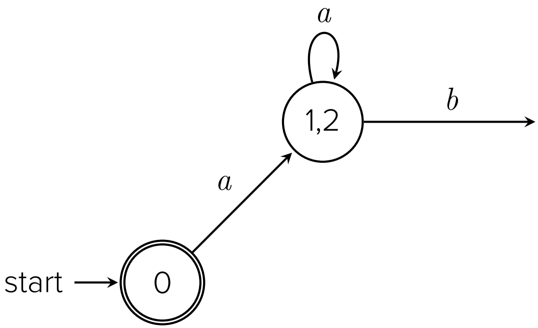
For $n > k \geq 0$, if $\mathrm{sc}(L) = n$ then

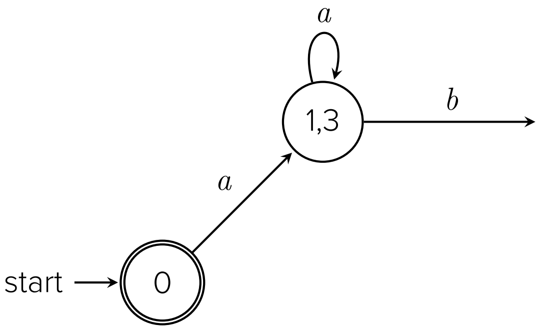
$$\mathrm{sc}(E(L, d_p, k)) \leq n \cdot (k+1) - \frac{k(k+1)}{2}.$$

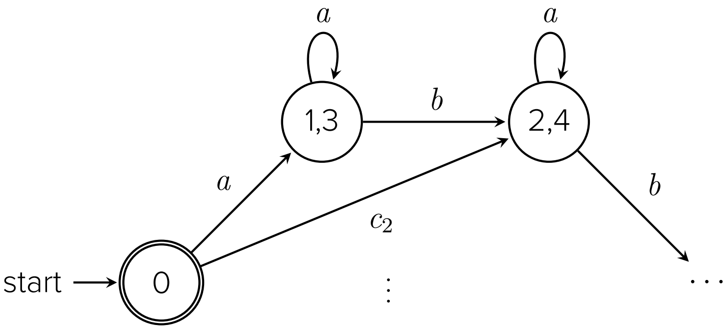

start
0
1
2
n − 1
n − 2
a
b
c2
cn−2
⋯
⋮

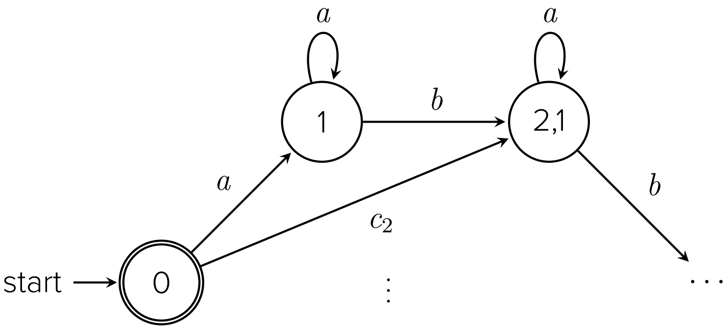

a
a
1
b
2
a
c2
b
start
0
⋮
…

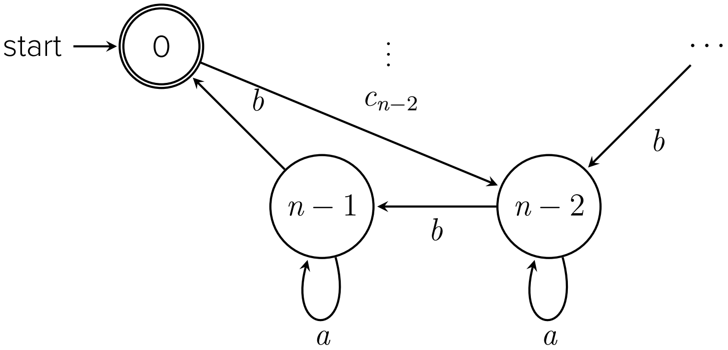

a
b
1,1
a
start
0

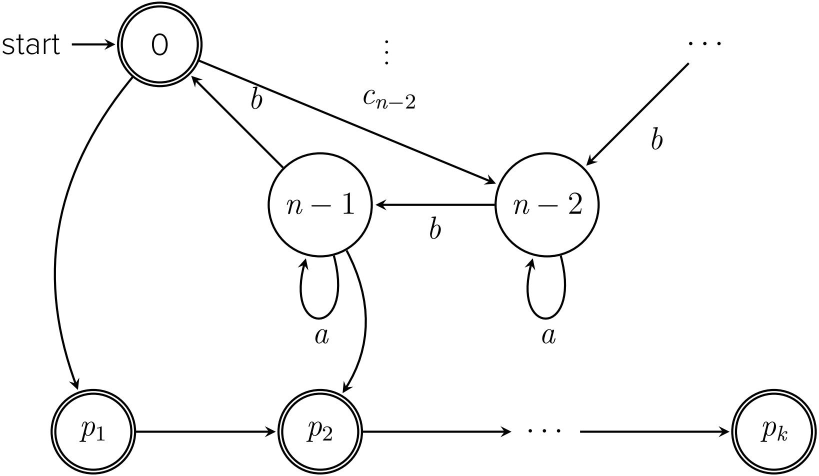

a
b
1,2
a
start
0


a
b
1,3
a
start
0


a
a
b
1,3
2,4
a
b
c_2
start
0
⋮
…

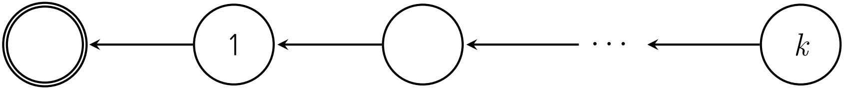

a
a
b
1
2,1
a
b
c2
start
0
⋮
⋯


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

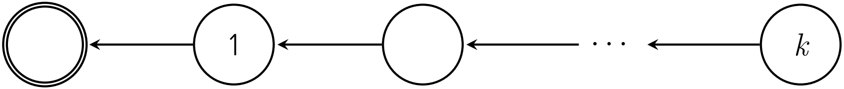

start
0
⋮
c_{n-2}
⋯
b
b
$n-1$
$n-2$
b
a
a
p_1
p_2
⋯
p_k

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.



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



There are at least $1 + 2 + \cdots + k$ unreachable states. The number of reachable states is at most

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

A language is finite if and only if it is recognized by an acyclic finite automaton.

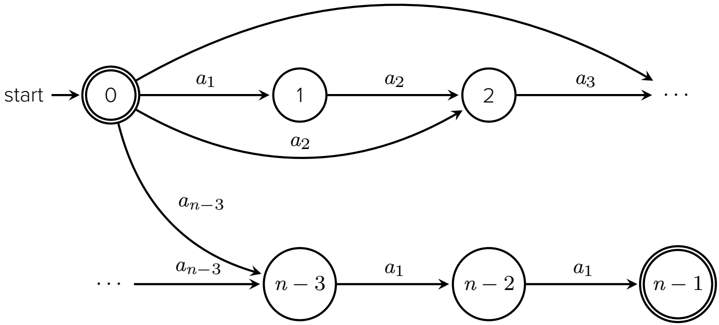
Theorem

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

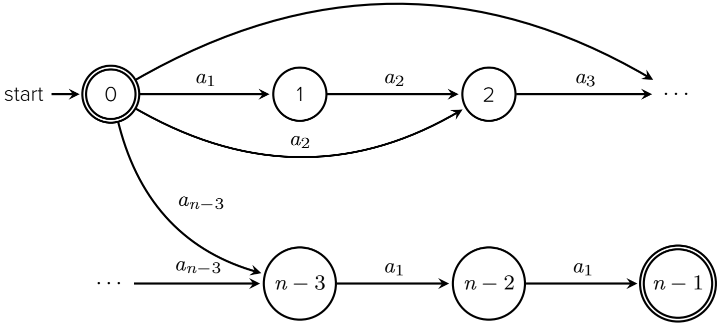
$$\mathrm{sc}(E(L, d_p, k)) \leq (n-2) \cdot (k+1) - k^2 + 2,$$

and this bound can be reached in the worst case.

Each state has a longest word that reaches it.



There must be at least 2 final states.



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

A regular language is prefix-closed if $x \in L$ implies $p \in L$ for every prefix p of x.

A regular language is prefix-closed if $x \in L$ implies $p \in L$ for every prefix p of x. A prefix-closed regular language is recognized by a finite automaton with all states final.

Theorem

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.

Since every state is a final state, $f = n$ and

$$\begin{aligned}(n-f)\cdot(k+1)+f+k &= (n-n)\cdot(k+1)+n+k\\ &= n+k.\end{aligned}$$

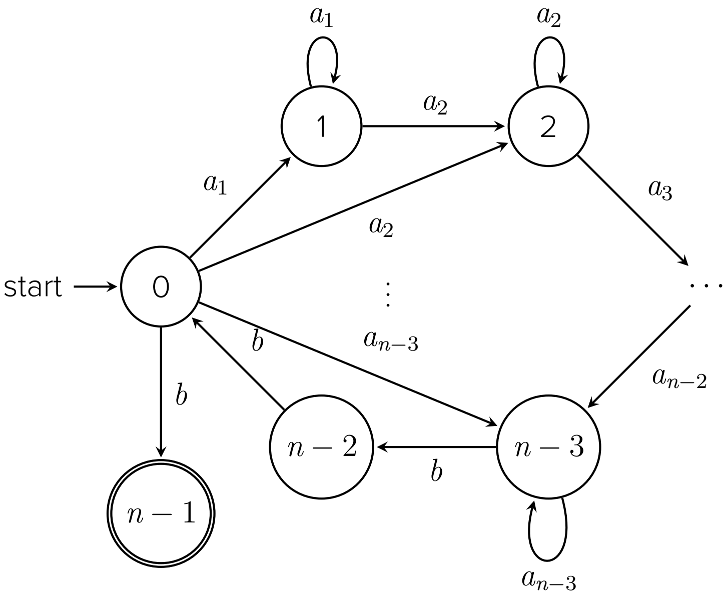
A language is prefix-free if for every $x \in L$, no prefix p of x is in L. A prefix-free regular language is recognized by a non-exiting finite automaton.

Theorem

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

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

and this bound can be reached in the worst case.



start
0
1
2
$\cdots$
$n-3$
$n-2$
$n-1$
a_1
a_2
a_3
$\vdots$
a_{n-3}
a_{n-2}
b

The state complexity neighbourhoods of some subregular language classes with respect to the prefix distance is strictly less than for regular languages.

Regular	$n \cdot (k+1) - \frac{k(k+1)}{2}$
Finite	$(n-2) \cdot (k+1) - k^2 + 2$
Prefix-closed	$n + k$
Prefix-free	$(n-1) \cdot k + 2 - \frac{k(k-1)}{2}$

Future work:

- Tight state complexity bounds for neighbourhoods of finite languages with respect to
  - additive distances
  - suffix and factor distances
- Other subregular language classes