State Complexity of Neighbourhoods and Approximate Pattern Matching

Timothy Ng, David Rappaport, and Kai Salomaa

School of Computing, Queen's University

DLT 2015, Liverpool, UK

July 27, 2015

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

1. A lower bound on the state complexity of neighbourhoods with respect to additive quasi-distances.

1. A lower bound on the state complexity of neighbourhoods with respect to additive quasi-distances.
2. State complexity of approximate pattern matching.

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. $d(x, y) \leq d(x, w) + d(w, y)$

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. $d(x, y) \leq d(x, w) + d(w, y)$

If condition (1) is relaxed to $d(x, y) = 0$ if $x = y$, then d is a quasi-distance.

Dovercourt $\rightarrow$ Davenpo_rt

Dovercourt → Davenpo_rt

Harbord → Harbourfront

Dovercourt → Davenpo_rt

Harbord → Harbourfront

Montréal → Montreal

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

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

A distance d on Σ^* is additive if for all factorizations $w = w_1 w_2$, we have for all $r \geq 0$

$$E(\{w\}, d, r) = \bigcup_{r_1 + r_2 = r} E(\{w_1\}, d, r_1) \cdot E(\{w_2\}, d, r_2)$$

A distance d on Σ^* is additive if for all factorizations $w = w_1 w_2$, we have for all $r \geq 0$

$$E(\{w\}, d, r) = \bigcup_{r_1 + r_2 = r} E(\{w_1\}, d, r_1) \cdot E(\{w_2\}, d, r_2)$$

Theorem (Calude, Salomaa, Yu 2002)

Let d be an additive quasi-distance on Σ^* and $L \subseteq \Sigma^*$ be a regular language. Then $E(L, d, r)$ is regular for all $r \geq 0$.

What is the state complexity of additive neighbourhoods?

What is the state complexity of additive neighbourhoods? We have upper bounds of

- $(r+2)^n$ for additive distances (Salomaa, Schofield 2007)

What is the state complexity of additive neighbourhoods? We have upper bounds of

- $(r+2)^n$ for additive distances (Salomaa, Schofield 2007)
- $(r+2)^n$ for additive quasi-distances (Ng, Rappaport, Salomaa 2015)

What is the state complexity of additive neighbourhoods? We have upper bounds of

- $(r+2)^n$ for additive distances (Salomaa, Schofield 2007)
- $(r+2)^n$ for additive quasi-distances (Ng, Rappaport, Salomaa 2015)

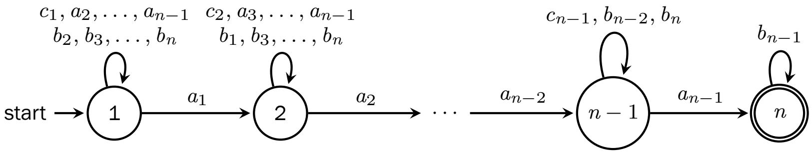
What is the lower bound for the state complexity of additive neighbourhoods?

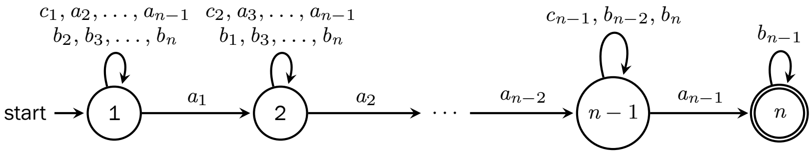
Lemma (Povarov 2007)

If a language $L \subseteq \Sigma^*$ is recognized by an n-state NFA, then the neighbourhood of L of radius r can be recognized by an NFA with $n(r+1)$ states.

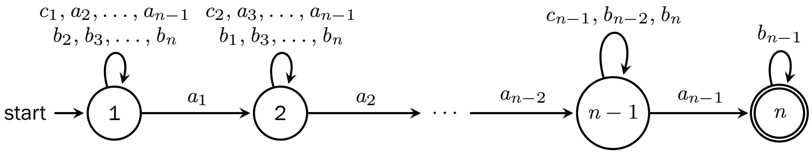
This construction gives an immediate upper bound of $2^{n(r+1)}$.

This construction gives an immediate upper bound of $2^{n(r+1)}$. We only need $(r+2)^n$ by only keeping track of the minimal distance computation.

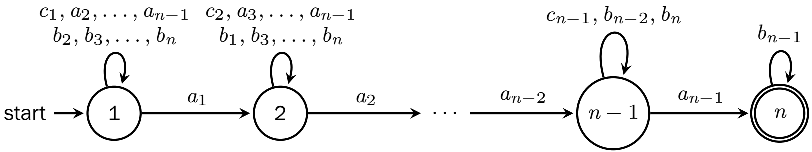

$c_1, a_2, \dots, a_{n-1}$
$b_2, b_3, \dots, b_n$
$c_2, a_3, \dots, a_{n-1}$
$b_1, b_3, \dots, b_n$
c_{n-1}, b_{n-2}, b_n
b_{n-1}
start
1
a_1
2
a_2
$\cdots$
a_{n-2}
$n-1$
a_{n-1}
n



- $d_r(a_i, c_i) = 0$ for $1 \leq i \leq n-1$

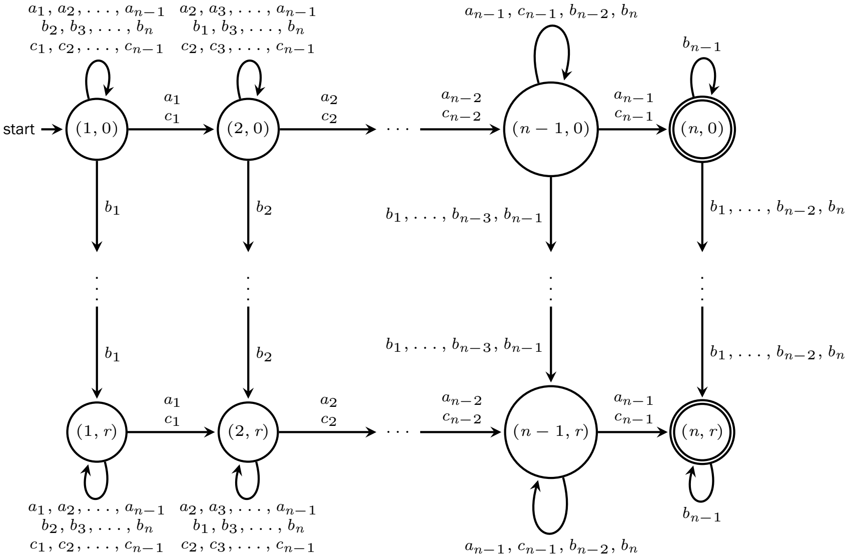


- $d_r(a_i, c_i) = 0$ for $1 \leq i \leq n-1$
- $d_r(b_i, b_j) = 1$ for $i \neq j$

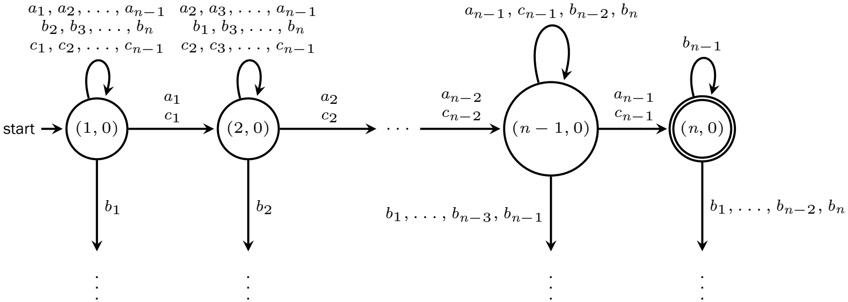


- $d_r(a_i, c_i) = 0$ for $1 \leq i \leq n-1$
- $d_r(b_i, b_j) = 1$ for $i \neq j$

- $d_r(a_i, a_j) = r+1$ for $i \neq j$
- $d_r(a_i, b_j) = d_r(c_i, b_j) = r+1$ for all $1 \leq i, j \leq n$
- $d_r(c_i, c_j) = r+1$ for all $1 \leq i, j \leq n$
- $d_r(a_i, c_j) = r+1$ for all $i \neq j$
- $d_r(\sigma, \varepsilon) = r+1$ for all $\sigma \in \Sigma$.



start
$(1, 0)$
$(2, 0)$
$(n-1, 0)$
$(n, 0)$
$(1, r)$
$(2, r)$
$(n-1, r)$
(n, r)
$a_1, a_2, \dots, a_{n-1}$
$b_2, b_3, \dots, b_n$
$c_1, c_2, \dots, c_{n-1}$
$a_2, a_3, \dots, a_{n-1}$
$b_1, b_3, \dots, b_n$
$c_2, c_3, \dots, c_{n-1}$
$a_{n-1}, c_{n-1}, b_{n-2}, b_n$
b_{n-1}
a_1
c_1
a_2
c_2
a_{n-2}
c_{n-2}
a_{n-1}
c_{n-1}
b_1
b_2
$b_1, \dots, b_{n-3}, b_{n-1}$
$b_1, \dots, b_{n-2}, b_n$
b_1
b_2
$b_1, \dots, b_{n-3}, b_{n-1}$
$b_1, \dots, b_{n-2}, b_n$
a_1
c_1
a_2
c_2
a_{n-2}
c_{n-2}
a_{n-1}
c_{n-1}
$a_1, a_2, \dots, a_{n-1}$
$b_2, b_3, \dots, b_n$
$c_1, c_2, \dots, c_{n-1}$
$a_2, a_3, \dots, a_{n-1}$
$b_1, b_3, \dots, b_n$
$c_2, c_3, \dots, c_{n-1}$
$a_{n-1}, c_{n-1}, b_{n-2}, b_n$
b_{n-1}



$$w(k_1, \ldots, k_n) = a_1 b_1^{k_1} a_2 b_2^{k_2} \cdots a_{n-1} b_{n-1}^{k_{n-1}} b_n^{k_n}$$

Theorem

If d is an additive quasi-distance, A is an NFA with n states and $r \in \mathbb{N}$,

$$\mathrm{sc}(E(L(A), d, r)) \leq (r+2)^n.$$

There exists an additive quasi-distance d_r and a DFA A with n states over an alphabet of size $3n-2$ such that $\mathrm{sc}(E(L(A), d_r, r)) = (r+2)^n$.

Given a pattern P of length m and a text T, does T contain substrings of length m having symbols differing from P in at most r positions? (El-Mabrouk 1997)

Given a pattern P of length m and a text T, does T contain substrings of length m having symbols differing from P in at most r positions? (El-Mabrouk 1997)

- $\binom{m+1}{r+1}$ for $P = a^m$.

Given a pattern P of length m and a text T, does T contain substrings of length m having symbols differing from P in at most r positions? (El-Mabrouk 1997)

- $\binom{m+1}{r+1}$ for $P = a^m$.
- $\sum_{i=1}^{r} \alpha_i b_i$, b_i the $(i+1)$th Catalan number; for m different symbols.

For a given FA A and quasi-distance d, what is the state complexity of the set of strings that contain a substring with distance r from $L(A)$?

For a given FA A and quasi-distance d, what is the state complexity of the language $\Sigma^* \cdot E(L(A), d, r) \cdot \Sigma^*$?

For a given FA A and quasi-distance d, what is the state complexity of the language $\Sigma^* \cdot E(L(A), d, r) \cdot \Sigma^*$?
For $r = 0$, we have $2^{n-2} + 1$ for $\Sigma^* \cdot L(A) \cdot \Sigma^*$ (Brzozowski, Jirásková, Li 2010).

Modify the neighbourhood DFA construction.

Modify the neighbourhood DFA construction.

- Start a new computation on each symbol.

Modify the neighbourhood DFA construction.

- Start a new computation on each symbol.
- We can accept once a final state is reached.

Modify the neighbourhood DFA construction.

- Start a new computation on each symbol.
- We can accept once a final state is reached.

For k final states, this gives us a machine with at most $(r+2)^{n-1-k}+1$ states.

Theorem

Let d be an additive quasi-distance on Σ^*. For any n-state NFA A and $r \in \mathbb{N}$ we have

$$\mathrm{sc}(\Sigma^* \cdot E(L(A), d, r) \cdot \Sigma^*) \leq (r+2)^{n-2} + 1.$$

For given $n, r \in \mathbb{N}$ there exists an additive distance d_r and an n-state NFA A defined over an alphabet of size $2n-1$ such that $\mathrm{sc}(\Sigma^* E(L(A), d_r, r)\Sigma^*) = (r+2)^{n-2} + 1$.

1. Lower bound of $(r+2)^n$ states for additive neighbourhoods

1. Lower bound of $(r+2)^n$ states for additive neighbourhoods
2. Approximate pattern matching with $(r+2)^{n-2}+1$ states

1. Lower bound of $(r+2)^n$ states for additive neighbourhoods
2. Approximate pattern matching with $(r+2)^{n-2}+1$ states

- Lower bound with fixed alphabet.

1. Lower bound of $(r+2)^n$ states for additive neighbourhoods
2. Approximate pattern matching with $(r+2)^{n-2}+1$ states

- Lower bound with fixed alphabet.
- State complexity for additive distances.

1. Lower bound of $(r+2)^n$ states for additive neighbourhoods
2. Approximate pattern matching with $(r+2)^{n-2}+1$ states

- Lower bound with fixed alphabet.
- State complexity for additive distances.
- State complexity for specific distances.