Parikh Equivalence and Descriptional Complexity

Giovanni Pighizzini

Dipartimento di Informatica
Università degli Studi di Milano, Italy

Workshop on Descriptional and Computational Complexity of Languages
Project *Voices of CANTE* – CMUP, Porto, Portugal
January 24–25, 2014

Results from joint papers with Giovanna J. Lavado and Shinnosuke Seki
(SOFSEM 2012, DLT 2012, Inf. and Comput. 2013)

UNIVERSITÀ DEGLI STUDI
DI MILANO

NFAs vs DFAs

Subset construction: [Rabin&Scott '59]

NFA n states	$\Longrightarrow$	DFA 2^n states

The state bound cannot be reduced

[Lupanov '63, Meyer&Fischer '71, Moore '71]

What happens if we do not care of the order of symbols in the strings?

This problem is related to the concept of *Parikh Equivalence*

Parikh Equivalence

- $\Sigma = \{a_1, \ldots, a_m\}$ alphabet of m symbols
- *Parikh's map* $\psi : \Sigma^* \to \mathbb{N}^m$:

$$\psi(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_m})$$

  for each string $w \in \Sigma^*$
- *Parikh's image* of a language $L \subseteq \Sigma^*$:

$$\psi(L) = \{\psi(w) \mid w \in L\}$$

- $w' =_\pi w''$ iff $\psi(w') = \psi(w'')$
- $L' =_\pi L''$ iff $\psi(L') = \psi(L'')$

Parikh's Theorem

Theorem ([Parikh '66])

The Parikh image of a context-free language is a semilinear set, i.e, each context-free language is Parikh equivalent to a regular language

Example:

- $L = \{a^n b^n \mid n \geq 0\}$
- $R = (ab)^*$

$$\psi(L) = \psi(R) = \{(n, n) \mid n \geq 0\}$$

Different proofs after the original one of Parikh, e.g.

- [Goldstine '77]: a simplified proof
- [Aceto&Ésik&Ingólfsdóttir '02]: an equational proof
- ...
- [Esparza&Ganty&Kiefer&Luttenberger '11]: complexity aspects

Our Goal

We want to convert nondeterministic automata and context-free grammars into *small Parikh equivalent* deterministic automata

Problem (NFAs to DFAs)

NFA *n states*	$\Longrightarrow_\pi$	*DFA* *how many states?*

Problem (CFGs to DFAs)

CFG *size n*	$\Longrightarrow_\pi$	*DFA* *how many states?*

Why?

- Interesting theoretical properties:
  wrt Parikh equivalence regular and context-free languages are indistinguishable [Parikh '66]
- Connections of with:
  - Semilinear sets
  - Presburger Arithmetics [Ginsburg&Spanier '66]
  - Petri Nets [Esparza '97]
  - Logical formulas [Verma&Seidl&Schwentick '05]
  - Formal verification
    [Dang&Ibarra&Bultan&Kemmerer&Su'00, Göller&Mayr&To'09]
  - ...
- Unary case:
  size costs of the simulations of CFGs and PDAs by DFAs
  [Pighizzini&Shallit&Wang '02]

Converting NFAs

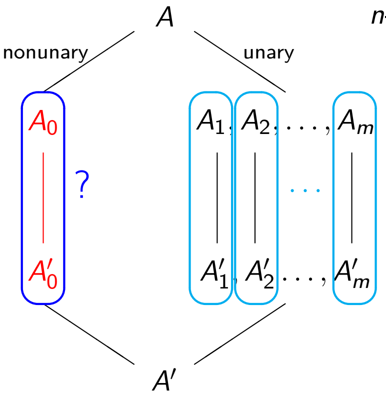
Problem (NFAs to DFAs)

NFA		*DFA*
n states	$\Longrightarrow_\pi$	*how many states?*

- *Upper bound*
  Subset construction: 2^n
- *Lower bound*
  Conversion *NFAs* $\rightarrow$ *DFAs* in the unary case: $e^{\Theta(\sqrt{n \ln n})}$

[Chrobak '86]

Converting NFAs: General Idea



n-state NFA over $\Sigma = \{a_1, \ldots, a_m\}$

$L(A_i) = L(A) \cap a_i^*$, $i \geq 1$

$L(A_0) = L - \bigcup_{i=0}^{m} L(A_i)$

Chrobak conversion: $e^{O(\sqrt{n \ln n})}$ states

Parikh equivalent DFAs

DFA Parikh equivalent to A

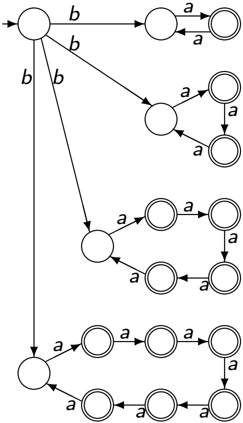
How much is the state cost of the conversion of NFAs accepting *only nonunary strings* into Parikh equivalent DFAs?

Only polynomial!

(less than in unary case)

An Example

$L = \{ba^n \mid n \bmod 210 \neq 0\}$



$L_1 = \{ba^n \mid n \bmod 2 \neq 0\}$

$L_1' = \{ba^n \mid n \bmod 2 \neq 0\}$

$L_2 = \{ba^n \mid n \bmod 3 \neq 0\}$

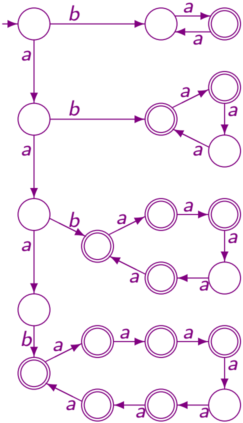
$L_2' = \{aba^{n-1} \mid n \bmod 3 \neq 0\}$

$L_3 = \{ba^n \mid n \bmod 5 \neq 0\}$

$L_3' = \{a^2ba^{n-2} \mid n \bmod 5 \neq 0\}$

$L_4 = \{ba^n \mid n \bmod 7 \neq 0\}$

$L_4' = \{a^3ba^{n-3} \mid n \bmod 7 \neq 0\}$



$DFA \geq 211$ states

$L' = L_1' \cup L_2' \cup L_3' \cup L_4'$

DFA with only 21 states!

Converting NFAs Accepting Only Nonunary Strings

The conversion uses a modification of the following result:

Theorem ([Kopczyński&To '10])

Given $\Sigma = \{a_1, \ldots, a_m\}$, there is a polynomial p s.t. for each n-state NFA A over Σ,

$$\psi(L(A)) = \bigcup_{i \in I} Z_i$$

where:

- *I is a set of at most $p(n)$ indices*
- *for $i \in I$, $Z_i \subseteq \mathbb{N}^m$ is a linear set of the form:*

  $$Z_i = \{\alpha_0 + n_1\alpha_1 + \cdots + n_k\alpha_k \mid n_1, \ldots, n_k \in \mathbb{N}\}$$

  with

  - $0 \leq k \leq m$
  - *the components of α_0 are bounded by $p(n)$*
  - *$\alpha_1, \ldots, \alpha_k$ are linearly independent vectors from $\{0, 1, \ldots, n\}^m$*

Converting NFAs Accepting Only Nonunary Strings

Outline: linear sets

Each above linear set

$$Z_i = \{\alpha_0 + n_1\alpha_1 + \cdots + n_k\alpha_k \mid n_1, \ldots, n_k \in \mathbb{N}\}$$

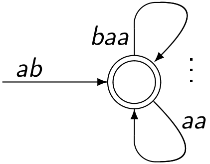
can be converted into a poly size DFA accepting a language

$$R_i = w_0(w_1 + \cdots + w_k)^*$$

s.t. $\psi(w_j) = \alpha_j$, $j = 0, \ldots, k$, and
$w_1, \ldots, w_k$ *begin with different letters*

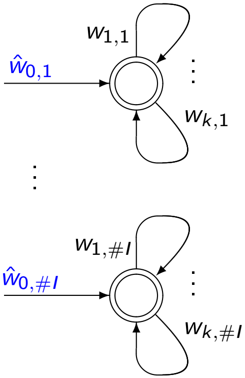
Example:

- $\{(1,1) + n_1(2,1) + n_2(2,0) \mid n_1, n_2 \geq 0\}$
- $ab(baa + aa)^*$



Converting NFAs Accepting Only Nonunary Strings

Outline: from linear to semilinear

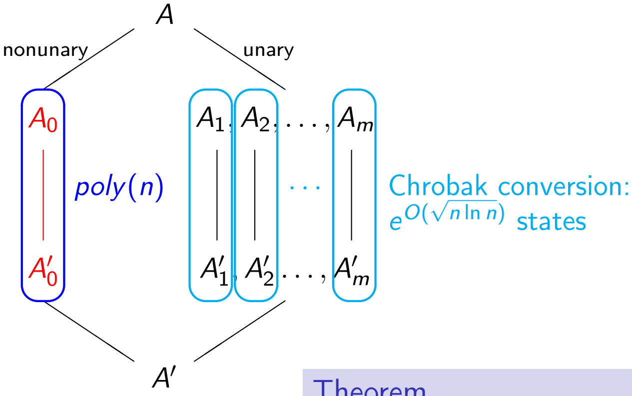


- Standard construction for union of DFAs: number of states = *product*
  $\#I \leq p(n)$ $\Rightarrow$ Too large!!!
- Strings $w_{0,i}$ can be replaced by Parikh equivalent strings $\hat{w}_{0,i}$ in such a way that $W_0 = \{\hat{w}_{0,i} \mid i \in I\}$ is a *prefix code*
- After this change: number of states $\leq$ *sum* Polynomial!!!

Theorem

For each n-state NFA accepting a language none of whose words are unary, there exists a Parikh equivalent DFA with a number of states polynomial in n

Converting NFAs: Back to the General Case



Theorem

For each n-state NFA there exists a Parikh equivalent DFA with $e^{O(\sqrt{n \ln n})}$ states. Furthermore this cost is tight

Converting CFGs

Problem (CFGs to NFAs and DFAs)

CFG *size h*	$\Longrightarrow_\pi$	*NFA/DFA* *how many states?*

- We consider CFGs in Chomsky Normal Form
- As a measure of size we consider the *number of variables*

[Gruska '73]

Problem (CFGs to NFAs)

CFG *Chomsky normal form* *h variables*	$\Longrightarrow_\pi$	*NFA* *how many states?*

Upper bound:

- $2^{2^{O(h^2)}}$ implicit construction from classical proof of Parikh's Th.
- $O(4^h)$ [Esparza&Ganty&Kiefer&Luttenberger '11]

Lower bound: $\Omega(2^h)$ Folklore

Problem (CFGs to DFAs)

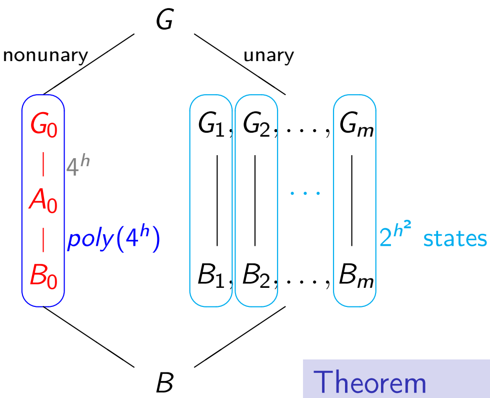
CFG
Chomsky normal form
h variables

$\Longrightarrow_\pi$

DFA
how many states?

- Upper bound: $2^{O(4^h)}$ — subset construction
- Lower bound: 2^{ch^2} — tight bound for the unary case $2^{\Theta(h^2)}$ [Pighizzini&Shallit&Wang '02]

Converting CFGs into Parikh Equivalent DFAs



CFG with h variables

Parikh equivalent NFA

Parikh equivalent DFAs

Theorem

For any CFG in Chomsky normal form with h variables, there exists a Parikh equivalent DFA with at most $2^{O(h^2)}$ states. Futhermore this bound is tight

Final Considerations

We obtained the following tight conversions:

	DFA	
NFA n states	$e^{O(\sqrt{n \ln n})}$ states	
CFG Cnf h variables	$2^{O(h^2)}$ states	

- In both cases the most expensive part is the unary one
- It could be interesting to investigate other conversions, e.g., automata minimization under Parkih equivalence, and computational complexity aspects

Final Considerations

Conversions into *two-way deterministic automata* (2DFAs)

	DFA	2DFA
NFA n states	$e^{O(\sqrt{n \ln n})}$ states	$poly(n)$ states
CFG Cnf h variables	$2^{O(h^2)}$ states	$2^{O(h)}$ states

Thank you for your attention!