On the commutative equivalence of bounded context-free and regular languages

Flavio D'Alessandro

Giornate di Informatica Teorica in memoria di Aldo de Luca

July 12, 2019

Regularity conditions for languages

Conditions that guarantee that a language is accepted by a finite automaton

[A. de Luca, S. Varricchio, 1999, EATCS]

"Finiteness and Regularity in Semigroups and Formal Languages"

A. de Luca, S. Varricchio, 1999, EATCS

BURNSIDE PROBLEM FOR SEMIGROUPS

Study of the conditions that guarantee that a finitely generated and periodic semigroup is finite

FINITENESS CONDITIONS VS LANGUAGES

Via the syntactic semigroup of L, a finiteness condition is a regularity condition for L

Brzozowski Conjecture, 1969

For $n, k \geq 1$, $B(k, n, n+1)$ is the free semigroup over k generators in the variety $x^n = x^{n+1}$

$$\varphi : A^* \longrightarrow B(k, n, n+1)$$

$$s \in B(k, n, n+1) \implies \varphi^{-1}(s) \in \mathsf{Rat}(A^*)$$

[A. de Luca, S. Varricchio, 1990]

Positive solution for $n \geq 5$

The Word problem is recursively decidable for $n \geq 5$

The Permutation property for semigroups

[A. Restivo, Ch. Reutenauer, 1984]

S finitely generated and periodic semigroup. Then S is finite if and only if S is permutable

[M. Curzio, P. Longobardi, M. Maj, D. Robinson, 1983, 1985]

Characterization of permutable groups

A proof of Restivo and Reutenauer result

[Ch. 3, Prop. 3.4.1]

Let S be a finitely generated and permutable semigroup

The set of the canonical representatives of S is a bounded language

Bounded languages

Definition

Let $L \subseteq A^*$. L is called n-bounded if there exist n words $u_1, u_2, \ldots, u_n$ such that

$$L \subseteq u_1^* u_2^* \cdots u_n^*$$

Bounded languages

Definition

Let $L \subseteq A^*$. L is called n-bounded if there exist n words $u_1, u_2, \ldots, u_n$ such that

$$L \subseteq u_1^* u_2^* \cdots u_n^*$$

$$L \subseteq \{u_1^{k_1} u_2^{k_2} \cdots u_n^{k_n} \ : \ k_1, \ldots, k_n \in \mathbb{N}\}$$

Bounded languages

Definition
Let $L \subseteq A^*$. L is called n-bounded if there exist n words $u_1, u_2, \ldots, u_n$ such that

$$L \subseteq u_1^* u_2^* \cdots u_n^*$$

$$L \subseteq \{u_1^{k_1} u_2^{k_2} \cdots u_n^{k_n} \; : \; k_1, \ldots, k_n \in \mathbb{N}\}$$

L is called bounded if it is n-bounded for some n

A characterization of bounded languages

[Ch. 2, Theorem 2.5.1, A. Restivo, Ch. Reutenauer, 1983]

L is bounded if and only if there exists some $n \geq 1$ such that

the set of factors of L does not contain n-divided words

Bounded languages

in

the theory of context-free languages

Main result

Every bounded context-free language L_1 is commutatively equivalent to a regular language L_2

Commutative Equivalence

Let $L_1, L_2 \subseteq A^*$

L_1 is commutatively equivalent to L_2 if there exists a bijection

$$f : L_1 \longrightarrow L_2$$

such that, for every $u \in L_1$,

$$\psi(u) = \psi(f(u))$$

The Parikh morphism

- $A = \{a_1, \ldots, a_t\}$
- $\psi : A^* \longrightarrow \mathbb{N}^t$
- $\forall u \in A^*, \quad \psi(u) = (|u|_{a_1}, |u|_{a_2}, \ldots, |u|_{a_t})$

Schützenberger Conjecture

(Schützenberger, 1956)

Every finite maximal (unique factorization, variable-length) code is commutatively equivalent to a prefix code

- Bounded and sparse context-free languages

- Our problem and its relations with the theory of formal languages

Bounded languages

Definition

Let $L \subseteq A^*$. L is called n-bounded if there exist n words $u_1, u_2, \ldots, u_n$ such that

$$L \subseteq u_1^* u_2^* \cdots u_n^*$$

Bounded languages

Definition

Let $L \subseteq A^*$. L is called n-bounded if there exist n words $u_1, u_2, \ldots, u_n$ such that

$$L \subseteq u_1^* u_2^* \cdots u_n^*$$

$$L \subseteq \{u_1^{k_1} u_2^{k_2} \cdots u_n^{k_n} \ : \ k_1, \ldots, k_n \in \mathbb{N}\}$$

L is called bounded if it is n-bounded for some n

Sparse languages

$L \subseteq A^*$

The counting function of L is the map

$$c_L : \mathbb{N} \longrightarrow \mathbb{N}$$

such that

$$c_L(n) = \mathrm{Card}(L \cap A^n)$$

Sparse and bounded languages

Definition

L is sparse or poly-slender if $c_L(n)$ is upper bounded by a polynomial in n

Sparse and bounded languages

Theorem (Latteux and Thierrin 1984; Ibarra and Ravikumar 1986; Raz 1997; Ilie, Rozenberg and Salomaa 2000)

A context-free language is sparse if and only if it is bounded

Sparse and bounded languages

Theorem (D., Intrigila, and Varricchio 2006)

Let L be a bounded context-free language over the alphabet A

Then there exists a regular language L' over an alphabet B such that, for all $n \geq 0$,

$$c_L(n) = c_{L'}(n)$$

The Problem

Given a bounded context-free language L, does it exist a regular language L' which is commutatively equivalent to L?

The Problem

Given a bounded context-free language L, does it exist a regular language L' which is commutatively equivalent to L?

Do there exist a regular language L' over the same alphabet of L and a bijection

$$f : L \longrightarrow L'$$

such that, for every $u \in L$, u and $f(u)$ have the same Parikh vector ?

- L commutatively equivalent to L' implies

$$\forall\, n \in \mathbb{N}, \quad c_L(n) = c_{L'}(n)$$

Main result

Theorem (D., Intrigila, 2014)

Every bounded context-free language is commutatively equivalent to a regular language. Moreover such construction is effective

Main result

Theorem (D., Intrigila, 2014)

Every bounded context-free language is commutatively equivalent to a regular language. Moreover such construction is effective

More generally, we prove that:

Theorem

Every bounded semi-linear language is commutatively equivalent to a regular language. Moreover such construction is effective

Example

$$L \subseteq a^*\mathbf{b}a^*$$

$$L = \{a^{1+x_1+2x_2}\mathbf{b}a^{x_2} \ : \ x_1, x_2 \geq 0\}$$

$$L \subseteq a^*\mathbf{b}a^*$$

$$L = \{a^{1+x_1+2x_2}\mathbf{b}a^{x_2} \ : \ x_1, x_2 \geq 0\}$$

$$L' = \{a^{1+x_1}\mathbf{b}a^{3x_2} \ : \ x_1, x_2 \geq 0\}$$

$$L \subseteq a^*\mathbf{b}a^*$$

$$L = \{a^{1+x_1+2x_2}\mathbf{b}a^{x_2} \ : \ x_1, x_2 \geq 0\}$$

$$L' = \{a^{1+x_1}\mathbf{b}a^{3x_2} \ : \ x_1, x_2 \geq 0\}$$

$$a^{1+x_1+2x_2}\mathbf{b}a^{x_2} \quad \xrightarrow{f} \quad a^{1+x_1}\mathbf{b}a^{3x_2}$$

A geometrical perspective

$$a^x \mathbf{b} a^y \longrightarrow (x, y) \in \mathbb{N}^2$$

Then the language

$$L = \{a^{1+x_1+2x_2} \mathbf{b} a^{x_2} \ : \ x_1, x_2 \geq 0\}$$

becomes

$$\{(1 + x_1 + 2x_2, \quad x_2) \ : \ x_1, x_2 \geq 0\}$$

A geometrical perspective

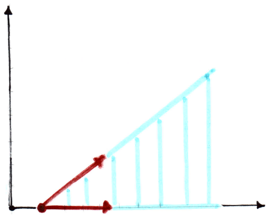
$$a^x \mathbf{b} a^y \longrightarrow (x, y) \in \mathbb{N}^2$$

and the language

$$L' = \{a^{1+x_1} \mathbf{b} a^{3x_2} \;:\; x_1, x_2 \geq 0\}$$

becomes

$$\{(1 + x_1, \quad 3x_2)\} \;:\; x_1, x_2 \geq 0\}$$



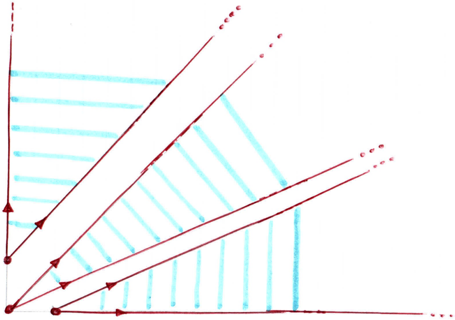
$(1 + x_1 + 2x_2,\ x_2)$

$(1 + x_1,\ 3x_2)$

$$L \subseteq a^*\mathbf{b}a^*$$

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

1. $$L_1 = a^{1+x_1+2x_2}\mathbf{b}a^{x_2}$$
2. $$L_2 = a^{2y_1+y_2}\mathbf{b}a^{y_1+2y_2}$$
3. $$L_3 = a^{z_1}\mathbf{b}a^{1+z_2+2z_1}$$



1. $L_1 = a^{1+x_1+2x_2}\mathbf{b}a^{x_2} \longrightarrow L'_1 = a^{1+x_1}\mathbf{b}a^{3x_2}$

2. $L_2 = a^{2y_1+y_2}\mathbf{b}a^{y_1+2y_2} \longrightarrow L'_2 = a^{3y_1}\mathbf{b}a^{3y_2}$

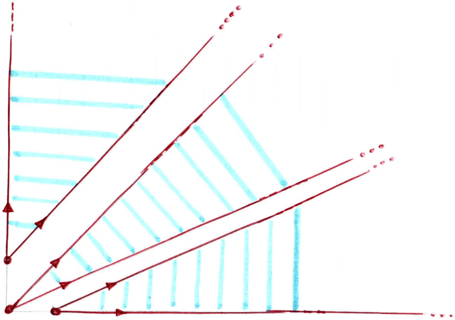
3. $L_3 = a^{z_1}\mathbf{b}a^{1+z_2+2z_1} \longrightarrow L'_3 = a^{3z_1}\mathbf{b}a^{z_2+1}$

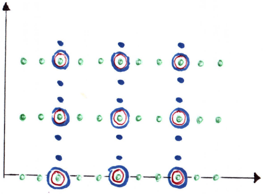
1. $L_1 = a^{1+x_1+2x_2}\mathbf{b}a^{x_2} \longrightarrow L_1' = a^{1+x_1}\mathbf{b}a^{3x_2}$

2. $L_2 = a^{2y_1+y_2}\mathbf{b}a^{y_1+2y_2} \longrightarrow L_2' = a^{3y_1}\mathbf{b}a^{3y_2}$

3. $L_3 = a^{z_1}\mathbf{b}a^{1+z_2+2z_1} \longrightarrow L_3' = a^{3z_1}\mathbf{b}a^{z_2+1}$

Obstruction : $a^3\mathbf{b}a^3 \in L_1' \cap L_2' \cap L_3'$





A regular language equivalent to $L = L_1 \cup L_2 \cup L_3$

$$
\begin{array}{ll}
(3x_1+1, \quad 3x_2) & \cup \\
(3x_1+1, \quad 3x_2+1) & \cup \\
(3x_1+1, \quad 3x_2+2) & \cup \\
(3y_1, \quad 3y_2) & \cup \\
(3z_1, \quad 3z_2+1) & \cup \\
(3z_1, \quad 3z_2+2) & \cup \\
(3z_1+2, \quad 3z_2+1) & .
\end{array}
$$

Some elements of the solution

Ambiguities of context-free languages

1. $L \subseteq u_1^* u_2^* \cdots u_k^*$ context-free bounded

2. Ambiguity of L as a context-free language

3. Ambiguity of L as a subset of the product

$$u_1^* \cdots u_k^*$$

Techniques

1. Faithful linear representation of bounded languages:
   Ginsburg and Spanier, 1966; Eilenberg Cross-section, 1974

2. Elementary number theory (on semi-linear sets):
   Eilenberg and Schützenberger result on semi-simple sets, 1969

3. Geometrical decomposition of semi-linear sets

4. Combinatorics of variable-length codes

Open problems

Gap Theorem

Theorem (Incitti 1999, Bridson and Gillman 1999)

A context-free language is either sparse or of exponential growth

Gap Theorem

Theorem (Incitti 1999, Bridson and Gillman 1999)

A context-free language is either sparse or of exponential growth

Theorem (Grigorchuk, Machí, 1998)

Existence of languages of intermediate growth

$$L = \{a^{n_1} b a^{n_2} b \cdots a^{n_{k-1}} b a^{n_k} \ : \ n_1 \leq n_2 \leq \cdots \leq n_k, \ k \geq 1\}$$

Languages of exponential growth

there exist languages of exponential growth that are not commutatively equivalent to regular languages

Languages of exponential growth

there exist languages of exponential growth that are not commutatively equivalent to regular languages

Search of conditions for and characterizations of languages of exponential growth commutatively equivalent to regular ones.

[D., Carpi, 2018, 2019]

Finite-index context-free languages

Languages of exponential growth

there exist languages of exponential growth that are not commutatively equivalent to regular languages

Search of conditions for and characterizations of languages of exponential growth commutatively equivalent to regular ones.

[D., Carpi, 2018, 2019]

Finite-index context-free languages

Minimal linear grammars

Minimal linear grammars and the commutative equivalence problem

Thank you