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

F. D'Alessandro[1] B. Intrigila[2]

[1]Dipartimento di Matematica "G. Castelnuovo"
Università di Roma "La Sapienza"

[2]Dipartimento di Matematica
Università di Roma "Tor Vergata"

Main result

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

Main result

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

There exists a bijection $f : L_1 \longrightarrow L_2$ such that, for every $u \in L_1$, u and $f(u)$ have the same Parikh vector

Overview of the presentation

- Bounded and sparse context-free languages
- The problem
- Outline of the solution

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 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

Sparse and bounded languages

Definition

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

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'Alessandro, 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

- Commutative Equivalence of languages
- Our problem
- Some classical theorems on bounded context-free languages

The Parikh morphism

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

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 $u \in L_1$,

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

Main result

Theorem (D.I. 2011)

Let $L_1 \subseteq u_1^* \cdots u_k^*$ be bounded context-free language.

Then L_1 is commutatively equivalent to a regular language L_2

Main result

Theorem (D.I. 2011)

Let $L_1 \subseteq u_1^* \cdots u_k^*$ be bounded context-free language.

Then L_1 is commutatively equivalent to a regular language L_2

Obstruction:

- inherently ambiguity of bounded context-free languages
- ambiguity of the product $u_1^* \cdots u_k^*$ in the free monoid A^*

Some classical theorems on bounded context-free languages

Parikh Theorem

- **Definition**
  Given languages $L_1,\ L_2 \subseteq A^*$, L_1 is letter-equivalent (or Parikh equivalent) to L_2 if $\psi(L_1) = \psi(L_2)$.

- **Theorem (Parikh, 1966)**
  Given a context-free language L_1, there exists a regular language L_2 which is *letter-equivalent to* L_1

Parikh Theorem

- $L_1 = (ab)^* \cup (ba)^*$, $L_2 = (ab)^*$
- $\psi(L_1) = \psi(L_2) = \{(n, n) : n \in \mathbb{N}\}$
- L_1 cannot be commutatively equivalent to L_2

Ginsburg Theorems

Given words $u_1, \ldots, u_k \in A^+$, we define the function:

$$\phi : \mathbb{N}^k \longrightarrow u_1^* u_2^* \cdots u_k^*,$$

such that, for every $(n_1, \ldots, n_k) \in \mathbb{N}^k$,

$$\phi(n_1, \ldots, n_k) = u_1^{n_1} u_2^{n_2} \cdots u_k^{n_k}$$

Ginsburg Theorems

$$\phi : \mathbb{N}^k \longrightarrow u_1^* u_2^* \cdots u_k^*,$$

$$\phi(n_1, \ldots, n_k) = u_1^{n_1} u_2^{n_2} \cdots u_k^{n_k}$$

Theorem (Ginsburg 1966)

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

L is context-free iff $\phi^{-1}(L)$ is a finite union of linear sets, each having a stratified sets of periods

Ginsburg Theorems

Theorem (Ginsburg, 1966)

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

L is unambiguous iff $\phi^{-1}(L)$ is a finite union of disjoint linear sets, each with stratified and linearly independent periods

$L = \{a^i b^j c^k \mid i, j, k \in \mathbb{N},\ i = j \text{ or } j = k\}$ is ambiguous

Outline of the solution

Inherent ambiguity of L

1. Faithful representation of L by a semilinear set

2. "Geometrical decomposition of semi-linear sets"

   [D'Alessandro, Intrigila, and Varricchio, 2010, Quasi-polynomials, linear Diophantine equations and semi-linear sets, *to appear in Theoret. Comput. Sci.*]

Ambiguity of $u_1^* \cdots u_n^*$

3. Arguments of Combinatorics of variable-length codes

4. Arguments of elementary number theory

Theorem (Eilenberg Cross-section, 1974)

Let $\alpha : A^* \to B^*$ be a morphism and let L be a rational language of A^*. There exists a rational subset L' of L such that α maps bijectively L' of $\alpha(L)$

Theorem (Eilenberg and Schützenberger, 1969)

Every semi-linear set is represented as a finite and disjoint union of

unambiguous linear sets

Theorem (Eilenberg Cross-section, 1974)

Let $\alpha : A^* \to B^*$ be a morphism and let L be a rational language of A^*. There exists a rational subset L' of L such that α maps bijectively L' of $\alpha(L)$

Theorem (Eilenberg and Schützenberger, 1969)

Every semi-linear set is represented as a finite and disjoint union of

unambiguous linear sets

Every semi-linear set is semi-simple set

Faithful representation by semilinear set

$$\phi : \mathbb{N}^k \longrightarrow u_1^* u_2^* \cdots u_k^*,$$

$$\phi(n_1, \ldots, n_k) = u_1^{n_1} u_2^{n_2} \cdots u_k^{n_k}$$

Theorem

If L is bounded context-free, then there exists a semi-simple set B of $\mathbb{N}^k$ such that

$$\phi(B) = L$$

and ϕ is injective on B

The basic case

$$\phi : \mathbb{N}^k \longrightarrow u_1^* u_2^* \cdots u_k^*,$$

$$\phi(B) = L$$

$$B = \{b_0 \ + \ b_1 n_1 + \cdots \ + \ b_m n_m \ : \ n_i \in \mathbb{N}\}$$

$$b_0, b_1, \ldots, b_m \in \mathbb{N}^k$$

The basic case

$$\phi(B) = L$$

$$B = \{b_0 \ + \ b_1 n_1 + \cdots \ + \ b_m n_m \ : \ n_i \in \mathbb{N}\}$$

The basic case

$$\phi(B) = L$$

$$B = \{b_0 \ + \ b_1 n_1 + \cdots \ + \ b_m n_m \ : \ n_i \in \mathbb{N}\}$$

$$u = \phi(b_0 \ + \ n_1 b_1 + \cdots \ + \ n_m b_m)$$

The basic case

$$\phi(B) = L$$

$$B = \{b_0 \ + \ b_1 n_1 + \cdots \ + \ b_m n_m \ : \ n_i \in \mathbb{N}\}$$

$$u = \phi(b_0 \ + \ n_1 b_1 + \cdots \ + \ n_m b_m)$$

Because of some elementary properties of ϕ and ψ, one has:

The basic case

$$\phi(B) = L$$

$$B = \{b_0 \ + \ b_1 n_1 + \cdots \ + \ b_m n_m \ : \ n_i \in \mathbb{N}\}$$

$$u = \phi(b_0 \ + \ n_1 b_1 + \cdots \ + \ n_m b_m)$$

Because of some elementary properties of ϕ and ψ, one has:

$$\psi(u) = \psi(\phi(b_0)) \ + \ n_1 \psi(\phi(b_1)) + \cdots \ + \ n_m \psi(\phi(b_m))$$

The basic case

The latter formula suggests that a natural candidate for the commutative equivalence of L is:

$$L' = \phi(b_0)\phi(b_1)^* \cdots \phi(b_m)^*$$

The basic case

Indeed, taking

$$L' = \phi(b_0)\phi(b_1)^* \cdots \phi(b_m)^*$$

one defines the function

$$f : L \longrightarrow L'$$

as:

$$f(u) = f(\phi(b_0 + n_1 b_1 + \cdots + n_m b_m)) =$$

The basic case

Indeed, taking

$$L' = \phi(b_0)\phi(b_1)^* \cdots \phi(b_m)^*$$

one defines the function

$$f : L \longrightarrow L'$$

as:

$$f(u) = f(\phi(b_0 + n_1 b_1 + \cdots + n_m b_m)) = \phi(b_0)\phi(b_1)^{n_1} \cdots \phi(b_m)^{n_m}$$

The basic case

$$f(u) = \phi(b_0)\phi(b_1)^{n_1} \cdots \phi(b_m)^{n_m}$$

- f is a surjective map from L to L'
- $\forall\ u \in L,\ \psi(u) = \psi(f(u))$

The basic case

$$f(u) = \phi(b_0)\phi(b_1)^{n_1} \cdots \phi(b_m)^{n_m}$$

- f is a surjective map from L to L'
- $\forall\ u \in L,\ \psi(u) = \psi(f(u))$

Obstruction: f is not necessarily injective!
The product $L' = \phi(b_0)\phi(b_1)^* \cdots \phi(b_m)^*$ is not necessarily unambiguous

The basic case

Solution: Construction of a regular language which is "algebraically similar" to L'

$$w_0 w_1^* \cdots w_m^*$$

but unambiguous as a product of languages of A^*

The basic case

Solution: Construction of a regular language which is "algebraically similar" to L'

$$w_0 w_1^* \cdots w_m^*$$

but unambiguous as a product of languages of A^*

- Combinatorics of variable-length codes
- "Geometrical decomposition of semi-linear sets"

Lemma

Let

$$z_1, \ldots, z_k,$$

be a list of (possibly equal) non-empty words over the alphabet A and let $\psi : A^* \longrightarrow \mathbb{N}^t$ be the Parikh map. Let $\mathbf{v}_1, \ldots, \mathbf{v}_\ell$ be the Parikh vectors of words $z_i,\ i = 1, \ldots, k$, and let

$$\mathcal{S} = \{(\alpha_1, \mathbf{v}_1), \ldots, (\alpha_\ell, \mathbf{v}_\ell)\},$$

be the corresponding multiset of the vectors ($k = \alpha_1 + \cdots + \alpha_\ell$). Suppose that:

- for every $j = 1, \ldots, k$, z_j contains, at least, two different letters of A in its factorization;
- for every $j = 1, \ldots, \ell$, $|\mathbf{v}_j| = \beta$;
- for every $j = 1, \ldots, \ell$, every vector $\mathbf{v}_j$ has the form $\mathbf{v}_j = N_j \bar{\mathbf{v}}_j$ for some $N_j \in \mathbb{N}$ and some $\bar{\mathbf{v}}_j \in \mathbb{N}^t$.

If, for every $j = 1, \ldots, \ell$,

$$N_j \geq k(\gamma + 1)(n + 1),$$

then there exists a (variable length) uniform code $\mathcal{W}$ of k (distinct) words over the alphabet A such that

$$\forall\ i = 1, \ldots, \ell, \quad \mathrm{Card}(\{w \in \mathcal{W} \ \mid\ \psi(w) = \mathbf{v}_i\}) = \alpha_i, \tag{1}$$

that is, for every $i = 1, \ldots, \ell$, the number of words of $\mathcal{W}$ whose Parikh vector is $\mathbf{v}_i$ is α_i. Moreover every $w \in \mathcal{W}$ is not a factor of a word in $u_1^* \cdots u_n^*$.

given a distribution of Parikh vectors of k words, all of them of the same length and with at least two different symbols in their factorization, under the assumption that all the components of every vector are sufficiently large, then:

one can construct a uniform length code with the same distribution of Parikh vectors

Consider the following context-free language $L \subseteq a^*b^*a^*$ given as

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

with:

1. $L_1 = a^{n_1}ba^{n_1}a^{n_2}a$
2. $L_2 = a^{m_1}a^{2m_2}aba^{m_1}a^{m_2}$
3. $L_3 = a^{2p_1}a^{p_2}a^2ba^{p_1}$

with $n_1, n_2, m_1, m_2, p_1, p_2$ ranging over $\mathbb{N}$.