

Note

On the separability of sparse context-free languages and of bounded rational relations[☆]Christian Choffrut^a, Flavio D'Alessandro^b, Stefano Varricchio^{c,*}^a *Laboratoire LIAFA, Université de Paris 7, 2, pl. Jussieu, 75251 Paris Cedex 05, France*^b *Dipartimento di Matematica, Università di Roma "La Sapienza", Piazzale Aldo Moro 2, 00185 Roma, Italy*^c *Dipartimento di Matematica, Università di Roma "Tor Vergata", via della Ricerca Scientifica, 00133 Roma, Italy*

Received 21 December 2006; received in revised form 3 April 2007; accepted 3 April 2007

Communicated by A.K. Salomaa

Abstract

This paper proves two results. (1) Given two bounded context-free languages, it is recursively decidable whether or not there exists a regular language which includes the first and is disjoint with the second and (2) given two rational k -ary bounded relations it is recursively decidable whether or not there exists a recognizable relation which includes the first and is disjoint with the second.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Bounded languages; Context-free languages; Rational relations; Decisional problems

1. Introduction

In the most general terms, the problem we tackle can be stated as follows. Given two families $\mathcal{F}_0, \mathcal{F}_1$ of subsets of a given set E , is it possible, given two subsets X, Y in \mathcal{F}_1 , to determine whether or not there exists a subset Z in \mathcal{F}_0 that *separates* them in the sense that $X \subseteq Z$ and $Y \cap Z = \emptyset$ holds? The problem is addressed in [2] where E is the direct product $A^* \times \mathbb{N}^k$ (where A^* is the free monoid generated by A and \mathbb{N} is the additive monoid of nonnegative integers), \mathcal{F}_1 is the family $\text{Rat}(A^* \times \mathbb{N}^k)$ of rational subsets of $A^* \times \mathbb{N}^k$ and \mathcal{F}_0 is the family $\text{Rec}(A^* \times \mathbb{N}^k)$ of recognizable subsets of $A^* \times \mathbb{N}^k$.

Here we consider two cases for which we give a positive answer based on the results of [2]. In the first case \mathcal{F}_1 is the family of bounded context-free languages and \mathcal{F}_0 is the family of regular languages. In the second case \mathcal{F}_1 is the family of bounded rational subsets of a direct product of finitely generated free monoids and \mathcal{F}_0 is their family of recognizable relations.

[☆] This work was partially supported by MIUR project ‘‘Linguaggi formali e automi: teoria e applicazioni’’. The second author acknowledges also the research funding ‘‘Facoltà di Scienze MM.FF.NN. 2004’’ of the University of Rome ‘‘La Sapienza’’.

* Corresponding author.

E-mail addresses: cc@liafa.jussieu.fr (C. Choffrut), dalessan@mat.uniroma1.it (F. D'Alessandro), varricch@mat.uniroma2.it (S. Varricchio).

URLs: <http://www.liafa.jussieu.fr/~cc> (C. Choffrut), <http://mat.uniroma1.it/people/dalessandro> (F. D'Alessandro), <http://mat.uniroma2.it/~varricch> (S. Varricchio).

To our knowledge the general problem where \mathcal{F}_1 is the unrestricted family of context-free languages is open and does not seem to be easy to solve. Indeed, if we were to consider \mathcal{F}_1 to be the family of deterministic context-free languages which is closed under complement, the decidability of the separability problem would entail the decidability of the question of whether or not given a subset in \mathcal{F}_1 belongs to \mathcal{F}_0 , which amounts to asking whether or not a deterministic context-free language is regular, a problem whose solution given by Stearns [13] and then improved by Valiant [14] is nontrivial.

2. Preliminaries

We assume that the reader is familiar with the basic notions of rational and recognizable subsets of an arbitrary monoid M , respectively denoted by $\text{Rat}(M)$ and $\text{Rec}(M)$ and with the notion of context-free languages. The reader is referred to the various textbooks on the topic [1,6,5,8,9]. In order to prevent any misunderstanding due to the not yet normalized use of these terms, we recall that a rational subset is expressed by a rational expression containing the operations of set union, set product and taking the submonoid generated, while a recognizable subset is a union of classes of a congruence of finite index on M . When M is the additive monoid \mathbb{N}^k , the family of rational subsets of \mathbb{N}^k coincides with the family of *semilinear sets*, i.e., finite unions of *linear sets* (cf. [12]).

2.1. Basic definitions

The basic notion underlying this work is the following.

Definition 1. Let M be a monoid. Two subsets X and Y of M are said to be *separable* if there exists a recognizable set Z of M such that:

$$X \subseteq Z, \quad Y \cap Z = \emptyset.$$

The subset Z *separates* X and Y .

Actually, the monoid that we are interested in is the free monoid. Given a finite *alphabet* Σ of *letters*, Σ^* denotes the free monoid that it generates. Its elements are called *words*.

The following theorem has been recently proven [2].

Theorem 1. Let $M = \Sigma^* \times \mathbb{N}^k$ be the direct product of the monoids Σ^* and \mathbb{N}^k , where Σ is a finite nonempty alphabet and \mathbb{N} is the additive monoid of nonnegative integers. Then it is decidable whether or not two rational sets of M are separable.

2.2. Bounded languages

In Section 3 we deal with context-free languages. The idea is to apply Theorem 1 by ignoring the component Σ^* and to convert rational subsets of \mathbb{N}^k into so-called k -bounded context-free languages of the free monoid. We are thus led to the following definition.

Definition 2. Let L be a language of a free monoid. For any positive integer k , L is called *k -bounded* if there exist nonempty words u_1, \dots, u_k such that

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

Moreover we say that L is *bounded* if there exists some integer $k \geq 1$ such that L is k -bounded.

We recall that bounded context-free languages are exactly the context-free languages for which the number of words belonging to the language and of a given length n is bounded by a polynomial in n [10,11]. These languages are thus also known as *sparse*. The counting function of sparse context-free languages and some related decision problems have been considered in [3,4].

Since the words $u_1, \dots, u_k \in \Sigma^*$ in the previous definition are fixed in the rest of the paper except if otherwise stated, the following proves to be useful.

Definition 3. Define $\phi(x_1, \dots, x_k) = u_1^{x_1} \cdots u_k^{x_k}$ for all $(x_1, \dots, x_k) \in \mathbb{N}^k$. Next let $A = \{a_1, \dots, a_k\}$ be a new alphabet of cardinality k . Consider the morphism defined by $h(a_i) = u_i$ for all $i = 1, \dots, k$ and the mapping $\theta : \mathbb{N}^k \rightarrow A^*$ defined as $\theta(x_1, \dots, x_k) = a_1^{x_1} \cdots a_k^{x_k}$. Then we have $\phi = h \circ \theta$.

The two main results on bounded languages used in this work are the following; see [8, Theorem 5.4.2] (actually a stronger result is proved) and [7, Theorem 1.2] respectively.

Theorem 2. Let $L \subseteq \Sigma^*$ be a bounded context-free language. Then $\phi^{-1}(L)$ is a rational subset of \mathbb{N}^k .

Theorem 3. Let $L \subseteq \Sigma^*$ be a bounded language. Let us have $k \in \mathbb{N}$ and let $u_1, \dots, u_k \in \Sigma^*$ such that $L \subseteq u_1^* \cdots u_k^*$. Then $L \in \text{Rec}(\Sigma^*)$ if and only if $\phi^{-1}(L) \in \text{Rec}(\mathbb{N}^k)$.

This theorem requires the subset of \mathbb{N}^k to be the inverse image of some subset in Σ^* . The next result, which is a consequence of the theorem, weakens the hypothesis.

Proposition 1. Let $R \in \text{Rec}(\mathbb{N}^k)$ and let $u_1, \dots, u_k \in \Sigma^*$. Then $\phi(R) \in \text{Rec}(\Sigma^*)$.

Proof. We use the notation of Definition 3. Because $R = \theta^{-1}(\theta(R))$ holds, we have $\theta(R) \in \text{Rec}(A^*)$ by the previous theorem. This yields $\phi(R) = h(\theta(R))$, which completes the proof. \square

2.3. Recognizable relations

Since the second result (Section 4) concerns relations of a direct product of free monoids, say $M = M_1 \times \cdots \times M_k$, we recall the characterization of recognizable relations of M in terms of the recognizable subsets of each component M_i (this result is attributed to Elgot and Mezei by Eilenberg in [6]).

Theorem 4. A subset of $M_1 \times \cdots \times M_k$ is recognizable if and only if it is a finite union of subsets of the form $X_1 \times \cdots \times X_k$ where each X_i is a recognizable subset of M_i , for $i = 1, \dots, k$.

3. Separating bounded context-free languages

We now have all the ingredients to prove our main result concerning separability of bounded context-free languages.

Theorem 5. It is decidable whether two context-free, bounded languages of the free monoid Σ^* are separable or not.

Proof. Let L_1 and L_2 be two bounded context-free languages of Σ^* . Since the family of bounded languages is closed with respect to the operations of product and union of sets, we can always suppose that there exist words $u_1, \dots, u_k \in \Sigma^+$ such that $L_1, L_2 \subseteq u_1^* \cdots u_k^*$. Let ϕ be the mapping defined by $\phi(x_1, \dots, x_k) = u_1^{x_1} \cdots u_k^{x_k}$. We claim that L_1 and L_2 are separable if and only if so are $\phi^{-1}(L_1)$ and $\phi^{-1}(L_2)$ which are rational subsets of \mathbb{N}^k by Theorem 2.

Indeed, if there exists a recognizable subset R of Σ^* satisfying $L_1 \subseteq R$ and $L_2 \cap R = \emptyset$, then by Theorem 3 the subset $\phi^{-1}(R)$ is recognizable in \mathbb{N}^k . Now, $L_1 \subseteq R$ implies $\phi^{-1}(L_1) \subseteq \phi^{-1}(R)$ and $L_2 \cap R = \emptyset$ implies $\phi^{-1}(L_2) \cap \phi^{-1}(R) = \phi^{-1}(L_2 \cap R) = \emptyset$.

Conversely, if $\phi^{-1}(L_1)$ and $\phi^{-1}(L_2)$ are separable by a recognizable subset $R \subseteq \mathbb{N}^k$, then by the previous proposition we have $\phi(R) \in \text{Rec}(\Sigma^*)$. Furthermore, $\phi^{-1}(L_1) \subseteq R$ implies $L_1 = \phi(\phi^{-1}(L_1)) \subseteq \phi(R)$. Finally, if $L_2 \cap \phi(R) = \phi(\phi^{-1}(L_2)) \cap \phi(R) \neq \emptyset$ then there exists an element $x \in R$ which maps into L_2 , implying $x \in \phi^{-1}(L_2)$, a contradiction.

The reduction to the result in [2] goes as follows. Let L_1 and L_2 be two bounded context-free languages. By a result of S. Ginsburg [8, Theorem 5.5.2], one can effectively compute nonempty words $v_1, \dots, v_p, w_1, \dots, w_r$, such that $L_1 \subseteq v_1^* \cdots v_p^*$ and $L_2 \subseteq w_1^* \cdots w_r^*$. Let $k = p + r$ and define

$$u_i = \begin{cases} v_i & \text{for } i = 1, \dots, p, \\ w_{i-p} & \text{for } i = p + 1, \dots, k. \end{cases}$$

The languages L_1 and L_2 may be viewed as bounded languages in $u_1^* \cdots u_k^*$. We now use the notation of Definition 3. Consider the Parikh map $\psi : A^* \rightarrow \mathbb{N}^k$ which assigns to each $u \in A^*$ the k -tuple $(|u|_{a_1}, \dots, |u|_{a_k})$ of number of

occurrences of each letter of A in u . Obviously $\phi^{-1}(L_1) = \psi(h^{-1}(L_1) \cap a_1^* \cdots a_k^*)$ and $\phi^{-1}(L_2) = \psi(h^{-1}(L_2) \cap a_1^* \cdots a_k^*)$. Since the languages $h^{-1}(L_1) \cap a_1^* \cdots a_k^*$ and $h^{-1}(L_2) \cap a_1^* \cdots a_k^*$ are context-free languages, we may resort to the well known Parikh theorem, which implies that the sets $\phi^{-1}(L_1)$ and $\phi^{-1}(L_2)$ are effective semilinear subsets of \mathbb{N}^k . Then apply the decision procedure to $\phi^{-1}(L_1)$ and $\phi^{-1}(L_2)$. \square

Lemma 1. *Let \mathcal{F} be a family of subsets of Σ^* closed under intersection with the recognizable subsets. Let $L_1, L_2 \in \mathcal{F}$ and assume $L_1 \subseteq R$ for some recognizable subset R . Then L_1 and L_2 are separable if and only if there exists a recognizable subset $S \subseteq R$ separating L_1 and $L_2 \cap R$.*

Proof. The condition is sufficient since if it holds then we have $L_1 \subseteq S$ and $L_2 \cap S = (L_2 \cap R) \cap S = \emptyset$. It is necessary since if $L_1 \subseteq S$ and $L_2 \cap S = \emptyset$ holds, then $L_1 \subseteq S \cap R$ and $(L_2 \cap R) \cap (S \cap R) = L_2 \cap (S \cap R) = \emptyset$ holds. \square

As a consequence, we have

Corollary 1. *Let L_1, L_2 be context-free languages of Σ^* and assume that L_1 is bounded. Then it is decidable whether L_1 and L_2 are separable or not.*

4. Separating bounded rational relations

In this last section we consider finite direct products of finitely generated free monoids, i.e., $A_1^* \times \cdots \times A_k^*$. It is well known that the family of recognizable subsets is strictly included in the family of rational subsets whenever at least two alphabets are non-empty. The problem posed in the introduction therefore makes sense in this setting. Here also, we show how the decidability is a consequence of the result in [2].

The following is a formal definition of bounded relations.

Definition 4. A relation $R \subseteq A_1^* \times \cdots \times A_k^*$ is *bounded* if there exist n_1 words $u_{1,1} \cdots u_{1,n_1} \in A_1^*$, etc ..., n_k words $u_{k,1} \cdots u_{k,n_k} \in A_k^*$ such that $R \subseteq u_{1,1}^* \cdots u_{1,n_1}^* \times \cdots \times u_{k,1}^* \cdots u_{k,n_k}^*$. Define the mapping $\phi : \mathbb{N}^{n_1 + \cdots + n_k} \rightarrow A_1^* \times \cdots \times A_k^*$ as

$$\phi(x_{1,1}, \dots, x_{1,n_1}, \dots, x_{k,1}, \dots, x_{k,n_k}) = (u_{1,1}^{x_{1,1}} \cdots u_{1,n_1}^{x_{1,n_1}}, \dots, u_{k,1}^{x_{k,1}} \cdots u_{k,n_k}^{x_{k,n_k}}).$$

The restriction of ϕ to \mathbb{N}^{n_i} is denoted by ϕ_i .

4.1. Closure properties of rational and recognizable subsets

Given two monoids M and N and a morphism $h : M \rightarrow N$, it is well known that the image under h of a rational subset of M is a rational subset of N and that the inverse image of a recognizable subset of N is a recognizable subset of M . Loosely speaking, this means that the family of rational subsets is closed under direct morphism and that the family of recognizable subsets is closed under inverse morphism: $h(\text{Rat}(M)) \subseteq \text{Rat}(N)$ and $h^{-1}(\text{Rec}(N)) \subseteq \text{Rec}(M)$. The inclusions $h(\text{Rec}(M)) \subseteq \text{Rec}(N)$ and $h^{-1}(\text{Rat}(N)) \subseteq \text{Rat}(M)$ do not hold in general. Here we show that they do hold under specific conditions on the monoids and the morphisms. Indeed, consider two direct products of free monoids $M = B_1^* \times \cdots \times B_k^*$ and $N = A_1^* \times \cdots \times A_k^*$ and morphisms $h : M \rightarrow N$ defined as follows. Let $h_i : B_i^* \rightarrow A_i^*$ be a morphism for $i = 1, \dots, k$ and define $h(w_1, \dots, w_k) = (h_1(w_1), \dots, h_k(w_k))$.

Proposition 2. *With the morphism defined as previously we have: If $R \in \text{Rec}(M)$ then $h(R) \in \text{Rec}(N)$. If $R \in \text{Rat}(N)$ then $h^{-1}(R) \in \text{Rat}(M)$.*

Proof. We show that if $R \in \text{Rat}(A_1^* \times \cdots \times A_k^*)$ then $h^{-1}(R) \in \text{Rat}(B_1^* \times \cdots \times B_k^*)$. By composition we may assume that the morphism leaves unchanged all components except one, e.g., that $h(u_1, u_2, \dots, u_k) = (h_1(u_1), u_2, \dots, u_k)$ holds.

Let \mathcal{A} be a k -tape automaton which accepts a (rational) relation $R \subseteq A_1^* \times \cdots \times A_k^*$. We may assume that the transitions of \mathcal{A} are of the kind $(q, (x_1, x_2, \dots, x_k), p)$, where for any i , $x_i \in A_i \cup \varepsilon$, and there exists at most one j such that $x_j \neq \varepsilon$. The k -tape automaton \mathcal{B} which accepts the inverse image of R under the morphism h is defined as follows. The set $Q_{\mathcal{B}}$ of the states of \mathcal{B} contains the set $Q_{\mathcal{A}}$ and new states of the kind $[q, u]$, where q is a state of $Q_{\mathcal{A}}$ and u is a nonempty suffix of some word of $h_1(B_1)$. Any transition of \mathcal{A} of the form $(q, (\varepsilon, x_2, \dots, x_k), p)$ is a

transition of \mathcal{B} as well as the transition $(q, (y, x_2, \dots, x_k), p)$ if $h_1(y) = \varepsilon$. Furthermore, it yields the new transitions $([q, u], (\varepsilon, x_2, \dots, x_k), [p, u])$.

For any $y \in B_1$ with $h_1(y) \neq \varepsilon$ and $q \in Q_{\mathcal{A}}$, we add to \mathcal{B} the transition

$$(q, (y, \varepsilon, \dots, \varepsilon), [q, h_1(y)]).$$

Finally, for any transition of \mathcal{A} of the form $(q, (a_1, \varepsilon, \dots, \varepsilon), p)$ we add the following ε -transitions to \mathcal{B} :

$$([q, a_1 \dots a_n], (\varepsilon, \varepsilon, \dots, \varepsilon), [p, a_2 \dots a_n]),$$

with $n \geq 2$, and

$$([q, a_1], (\varepsilon, \varepsilon, \dots, \varepsilon), p).$$

The initial state and the final states of \mathcal{B} are the same as those of \mathcal{A} . It is not difficult to see that the k -tape automaton \mathcal{B} accepts the set $h^{-1}(R)$.

We now show that if $R \in \text{Rec}(B_1^* \times \dots \times B_k^*)$ then $h(R) \in \text{Rec}(A_1^* \times \dots \times A_k^*)$. By the characterization of Elgot and Mezei, R is a finite union of direct products $X_1 \times \dots \times X_k$ where for $i = 1, \dots, k$, X_i is a recognizable set of B_i^* . It clearly suffices to consider the case where R is reduced to this product. But then we obtain $h(R) = h_1(X_1) \times \dots \times h_k(X_k)$ which is recognizable. \square

Proposition 3. Let $R \subseteq u_{1,1}^* \dots u_{1,n_1}^* \times \dots \times u_{k,1}^* \dots u_{k,n_k}^*$.

- (1) If R is rational then the set $\phi^{-1}(R)$ is rational.
- (2) If $S \subseteq \mathbb{N}^{n_1+\dots+n_k}$ is recognizable then $\phi(S)$ is recognizable.
- (3) R is recognizable if and only if $\phi^{-1}(R)$ is recognizable.

Proof. Claim 1. Consider for all $i = 1, \dots, k$ the alphabets $B_i = \{a_{i,1}, \dots, a_{i,n_i}\}$ of new symbols, the morphisms $h_i : B_i^* \rightarrow A_i^*$ defined by $h_i(a_{i,j}) = u_{i,j}$ and the Parikh mappings $g_i : B_i^* \rightarrow \mathbb{N}^{n_i}$. Set $g(w_1, \dots, w_k) = (g_1(w_1), \dots, g_k(w_k))$. Then we have

$$\phi^{-1}(R) = g \left(h^{-1}(R) \cap a_{1,1}^* \dots a_{1,n_1}^* \times \dots \times a_{k,1}^* \dots a_{k,n_k}^* \right).$$

The claim is a consequence of the previous proposition and the general closure properties of rational subsets.

Claim 2. If S is recognizable then, by the characterization of Elgot and Mezei, it is a finite union of direct products $X_1 \times \dots \times X_k$, where X_i is a recognizable set of \mathbb{N}^{n_i} , for $i = 1, \dots, k$. Then, $\phi(S)$ is a finite union of direct products $\phi_1(X_1) \times \dots \times \phi_k(X_k)$. By Proposition 1 each subset $\phi_i(X_i)$ is recognizable in A_i^* . This completes the proof.

Claim 3. If R is recognizable then, by the characterization of Elgot and Mezei, it is a finite union of direct products $Z = X_1 \times \dots \times X_k$, where for $i = 1, \dots, k$, X_i is a recognizable set of A_i^* included in $u_{i,1}^* \dots u_{i,n_i}^*$. Then, the subset $\phi_i^{-1}(X_i)$ is a recognizable subset of \mathbb{N}^{n_i} by Theorem 3. Therefore, since $\phi^{-1}(Z) = \phi_1^{-1}(X_1) \times \dots \times \phi_k^{-1}(X_k)$, then $\phi^{-1}(Z)$ is recognizable.

Conversely, if $\phi^{-1}(R)$ is recognizable in $\mathbb{N}^{n_1+\dots+n_k}$, then by claim 2 we have $R = \phi(\phi^{-1}(R))$ is recognizable in $A_1^* \times \dots \times A_k^*$. \square

We come to the main result of this section.

Theorem 6. Given two bounded rational subsets of a direct product of free monoids, it is recursively decidable whether or not they are separable.

Proof. The proof follows the same pattern as that for bounded context-free languages. The only point which requires some care concerns the effectiveness of the computation of the various words $u_{i,j}$. In the monoid which is a direct product of free monoids, it is recursively decidable whether or not a rational set is contained in a recognizable set [1] since this reduces to the emptiness problem for rational sets. Therefore, for fixed words $u_{1,1} \dots u_{1,n_1} \in A_1^*$, etc ..., $u_{k,1} \dots u_{k,n_k} \in A_k^*$, given a rational relation R , one can check the inclusion

$$R \subseteq u_{1,1}^* \dots u_{1,n_1}^* \times \dots \times u_{k,1}^* \dots u_{k,n_k}^*.$$

Since we know that these words exist, an exhaustive search can find them. \square

References

- [1] J. Berstel, *Transductions and Context-free Languages*. B. G. Teubner, 1979.
- [2] C. Hoffrut, S. Grigorieff, Separability of rational subsets by recognizable subsets is decidable in $\Sigma^* \times N^m$, *Inform. Process. Lett.* 99 (1) (2006) 27–32.
- [3] F. D’Alessandro, B. Intrigila, S. Varricchio, On the structure of the counting function of context-free languages, *Theoret. Comput. Sci.* 356 (2006) 104–117.
- [4] F. D’Alessandro, S. Varricchio, On the growth of context-free languages. Technical Report, Università di Roma La Sapienza, 2006.
- [5] A. de Luca, S. Varricchio, *Finiteness and Regularity in Semigroups and Formal Languages*, Springer-Verlag, 1999.
- [6] S. Eilenberg, *Automata, Languages and Machines*, vol. A, Academic Press, 1974.
- [7] S. Ginsburg, E.H. Spanier, Bounded regular sets, *Proc. Amer. Math. Soc.* 17 (1966) 1043–1049.
- [8] S. Ginsburg, *The mathematical theory of context-free languages*, McGraw-Hill Book Co., 1966.
- [9] J. Hopcroft, J. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley Pub. Co., 1979.
- [10] L. Ilie, G. Rozenberg, A. Salomaa, A characterization of poly-slender context-free languages, *Theor. Inform. Appl.* 34 (2000) 77–86.
- [11] M. Latteux, G. Thierrin, On bounded context-free languages, *Elektron. Informationsverarb. Kybernet.* 20 (1984) 3–8.
- [12] J. Sakarovitch, *Éléments de Théorie des Automates*, Vuibert, Paris, 2003.
- [13] R.E. Stearns, A regularity test for pushdown machines, *Information and Control* 11 (1967) 323–340.
- [14] L.G. Valiant, Regularity and related problems for deterministic pushdown automata, *Inform. Control* 22 (1975) 1–10.