ON PARTITIONS SEPARATING WORDS

ŠTĚPÁN HOLUB AND JUHA KORTELAINEN

Abstract. Partitions $\{L_k\}_{k=1}^m$ of A^+ into m pairwise disjoint languages L_1, $L_2, \ldots, L_m$ such that $L_k = L_k^+$ for $k = 1, 2, \ldots, m$ are considered. It is proved that such a closed partition of A^+ can separate the words $u_1, u_2, \ldots, u_m \in A^+$ (i.e. each L_k contains exactly one word of the sequence $u_1, u_2, \ldots, u_m$) if and only if for each pair i, j of distinct elements in $\{1, 2, \ldots, m\}$, the words u_i and u_j do not commute. Furthermore, it is proved that the separating languages can be chosen to be regular. In case that the Parikh images of the words are linearly independent, the choice of the separating languages may be based on geometrical intuition.

Formal languages; finite automata; separation by closed sets.

1. Introduction

The question whether a certain family of languages is closed under complement is one of standard questions of the classical language theory. It is well known, for instance, that the complement of a regular or context sensitive language is again regular or context sensitive, while the same is not in general true for context free or recursively enumerable languages.

Another case, interesting algebraically, is the class of languages that are also semigroups. In other words, these are languages closed under catenation, or, equivalently, under Kleene$^+$ (or positive closure, as it nowadays is sometimes called). Note that we are using the word "closed" in two different contexts. The family $\mathcal{S}$ of languages over an alphabet A is closed under complement if for each $L \in \mathcal{S}$ we have also $(A^+ \setminus L) \in \mathcal{S}$. (Note that in general A need not be equal to the set of letters used in the family S.) On the other hand, a language L itself is closed under a certain operation $\circ$ if $a \circ b \in L$ for each $a, b \in L$.

In algebra, a nice illustration of the above concepts are prime ideals. For example, the set $a \cdot \mathbb{Z}$, which is closed under multiplication, has the property that its complement (that is, the set $\mathbb{Z} \setminus a \cdot \mathbb{Z}$) is again closed under multiplication if and only if a is a prime number.

In topology, a set that is topologically closed and also its complement is topologically closed is called "clopen", since it is both closed and open. A celebrated theorem of Kuratowski [7] yields an interesting property of operations of (topological) closure and complement. The theorem says that at most fourteen different sets can be generated by a finite number of applications of those operations. This result holds also in closure systems that are not necessarily topological [5], [9].

As noted above, in formal languages Kleene$^+$ is a natural candidate for this kind of considerations, see [9]. The properties of semigroups and their complements, inspired by the topological point of view, are fairly extensively studied also in [2]. One of the central achievements of the paper tells that there exists a partition of A^+ into two semigroups separating two (nonempty) words if and only if the words

do not commute. We continue this research line and generalize the result to cover any number of words. Our main result says that any sequence $u_1, u_2, \ldots, u_m$ of noncommuting words in A^+ can be separated by a disjoint partition $\{L_i\}_{i=1}^m$ of A^+ where each of the languages $L_1, L_2 \ldots, L_m$ is a semigroup, and, moreover, regular. This separation result is thus connected to the classical Chomsky hierarchy of languages. The separation is then geometrically illustrated in a special case, i.e., when the Parikh-images of the words are pairwise linearly independent. The concepts and results are then generalized to separation of (certain types of) languages.

The paper is organized in the following way. The second section introduces basic definitions and preliminaries. The third section contains an encoding of words into integers and generalizes Theorem 14 in [2], which claims that two words can be separated by a closed partition if and only if they do not commute. Moreover, our separating languages are regular and we construct the corresponding automaton. The fourth section analyses techniques applied in [2] to separate words. We show by an example that although the resulting languages are always context-sensitive they are not necessarily context-free. We also show that if the Parikh images of the noncommuting words are linearly independent, then the separation can be carried out by using basic methods of geometry and combinatorial topology. The languages in the partition can be then chosen to be commutative and context-free. The last section contains some concluding remarks.

2. Preliminaries

Let $\mathbb{N} = \{0, 1, 2, \ldots\}$ be the set of all natural numbers and $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. By $\mathbb{Z}$ we denote the set of all integers and by $\mathbb{Q}$ the set of all rationals. Let $\gcd(m, n)$ be the greatest common divisor of the integers m and n. For each finite set S, let $|S|$ be the cardinality of S, i.e., the number of elements in S.

Let A be a finite alphabet and $w \in A^+$. The length of the word w is denoted by $|w|$; for each $a \in A$, let $|w|_a$ be the number of occurrences of the letter a in w, and let $\mathrm{alph}(w)$ denote the set of all letters occurring in w at least once. It will be mostly convenient to work with the semigroup A^+ which does not contain the empty word, rather than with the monoid A^*. We say that the words u and v *commute* if $uv = vu$.

Let $n \in \mathbb{N}_+$ and $A = \{a_1, a_2, \ldots, a_n\}$ be an alphabet of n symbols. The *Parikh map* Ψ_n (or Ψ when A is understood) from A^+ into $\mathbb{N}^n$ is the semigroup morphism defined by $\Psi_n(w) = (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_n})$. The vector $\Psi_n(w)$ is the *Parikh image* of the word w. The map is generalized to languages $L \subseteq A^*$ in the obvious way: $\Psi_n(L) = \{\Psi_n(w) \mid w \in L\}$. For each set $S \subseteq \mathbb{N}^n$, let $\Psi_n^{-1}(S) = \{w \in A^+ \mid \Psi_n(w) \in S\}$. The language L is *commutative* if for each $w_1 \in L$ and each $w_2 \in A^+$ such that $\Psi_n(w_1) = \Psi_n(w_2)$, the word w_2 is also in L. Clearly, a language L is commutative if and only if $L = \Psi_n^{-1}(\Psi_n(L))$.

A nonempty language $L \subseteq A^+$ is a semigroup if it is closed under Kleene$^+$, that is, if $L = L^+$; we shall also say that L is *positively closed* in such a case. Let $m \in \mathbb{N}_+$ and $L_1, L_2, \ldots, L_m \subseteq A^+$. Then $\{L_i\}_{i=1}^m$ is a *closed partition of* A^+ (into m languages) if

(i) the languages $L_1, L_2, \ldots, L_m$ are nonempty and pairwise disjoint;
(ii) $A^+ = \cup_{i=1}^m L_i$; and
(iii) $L_i = L_i^+$ for each $i \in \{1, 2, \ldots, m\}$.

The closed partition $\{L_i\}_{i=1}^m$ of A^+ *separates the words* $u_1, u_2, \ldots, u_m \in A^+$ if for each $i \in \{1, 2, \ldots, m\}$ there exists exactly one $j \in \{1, 2, \ldots, m\}$ such that $u_j \in L_i$.

We suppose that the reader is familiar with basic concepts of combinatorics on words as can be found in [8], where also a proof is given for the following fundamental result.

Lemma 1. *Two nonempty words commute if and only if they are powers of the same (primitive) word, i.e., they have the same primitive root.*

Recall that the (unique) *primitive root* of a nonempty word u is the shortest word r such that $u = r^i$ for some $i \in \mathbb{N}$.

3. Separation of words by regular languages

This section gives a proof of the following claim.

Theorem 2. *Let m be a positive integer and $u_1, u_2 \ldots, u_m$ nonempty words over an alphabet A. There exists a closed partition of A^+ separating the words $u_1, u_2 \ldots, u_m$ if and only if for all $i, j \in \{1, 2, \ldots, m\}$ such that $i \neq j$, the words u_i and u_j do not commute.*

Moreover, in such a case there always exist separating languages that are regular.

It is certainly enough to prove the theorem for the binary alphabet $B = \{0, 1\}$ only, since each alphabet can be encoded into a binary one.

The theorem holds trivially when $m = 1$, so in the following we implicitly assume that $m > 1$. Note also that the 'only if' part of the theorem is trivial, since commuting words must both be elements of the language containing their primitive root in a closed partition.

We shall exploit the one-to-one correspondence between the sets $\{0, 1\}^m$ and $\{0, 1, \ldots, 2^m - 1\}$, $m \in \mathbb{N}_+$, given by the binary enumeration system. The binary value of a word $w \in B^+$ is denoted by $\mathbf{b}(w)$. More rigorously, let $m \in \mathbb{N}_+$ and $w = x_1 x_2 \cdots x_m$ be a word such that $x_1, x_2, \ldots, x_m \in \{0, 1\}$. Then

$$\mathbf{b}(x_1 x_2 \cdots x_m) = \sum_{i=1}^{m} x_i \cdot 2^{m-i}.$$

Remark 1. The definition of $\mathbf{b}$ clearly implies that for all $u, v \in B^+$:

(i) $\mathbf{b}(uv) = \mathbf{b}(u)\, 2^{|v|} + \mathbf{b}(v)$;
(ii) $\mathbf{b}(u^n) = \mathbf{b}(u)\, \Sigma_{i=0}^{n-1} 2^{i|u|}$; and
(iii) $u = v$ if and only if $|u| = |v|$ and $\mathbf{b}(u) = \mathbf{b}(v)$.

The separation of noncommuting words is based on the following fact.

Lemma 3. *The words $u, v \in B^+$ commute if and only if $\frac{\mathbf{b}(u)}{2^{|u|}-1} = \frac{\mathbf{b}(v)}{2^{|v|}-1}$.*

Proof. The words u and v commute if and only if $uv = vu$. This, by item (iii) of Remark 1, is equivalent to $\mathbf{b}(uv) = \mathbf{b}(vu)$. By item (i) of Remark 1, $\mathbf{b}(uv) = \mathbf{b}(vu)$ is equivalent to $\mathbf{b}(u)\, 2^{|v|} + \mathbf{b}(v) = \mathbf{b}(v)\, 2^{|u|} + \mathbf{b}(u)$. The claim follows. □

Let us now prove the 'if' part of Theorem 2. Remember that $A = B$ $(= \{0, 1\})$.

Suppose thus that for all $i, j \in \{1, 2, \ldots, m\}$, $i \neq j$, the words u_i and u_j do not commute. By Lemma 3, the relation $\frac{\mathbf{b}(u_i)}{2^{|u_i|}-1} \neq \frac{\mathbf{b}(u_j)}{2^{|u_j|}-1}$ holds for all distinct

$i, j \in \{1, 2, \ldots, m\}$. Assume without loss of generality that

$$\frac{\mathbf{b}(u_1)}{2^{|u_1|}-1} > \frac{\mathbf{b}(u_2)}{2^{|u_2|}-1} > \cdots > \frac{\mathbf{b}(u_m)}{2^{|u_m|}-1} .$$

Let $p_i, q_i \in \mathbb{N}_+$, $i = 1, 2, \ldots, m-1$, be such that

$$\frac{\mathbf{b}(u_i)}{2^{|u_i|}-1} > \frac{q_i}{p_i} > \frac{\mathbf{b}(u_{i+1})}{2^{|u_{i+1}|}-1} .$$

Note that $0 < q_i < p_i$ since $\mathbf{b}(u_i) \leq 2^{|u_i|} - 1$. For each $i \in \{1, 2, \ldots, m-1\}$, let

$$T_i = \{\, w \in B^+ \mid p_i\, \mathbf{b}(w) - q_i\, (2^{|w|} - 1) \geq 0 \,\} .$$

We immediately note that the inclusion $T_i \subset T_{i+1}$ holds for $i = 1, 2, \ldots, m-2$. Moreover, from

$$p_i\, \mathbf{b}(uv) - q_i\, (2^{|uv|} - 1) = 2^{|v|}(p_i\, \mathbf{b}(u) - q_i\, (2^{|u|} - 1)) + (p_i\, \mathbf{b}(v) - q_i\, (2^{|v|} - 1))$$

we deduce that both T_i and $B^+ \setminus T_i$, $i = 1, 2, \ldots, m-1$, are positively closed. Let $L_1 = T_1$, $L_i = T_i \setminus T_{i-1}$ for $i = 2, \ldots, m-1$, and $L_m = B^+ \setminus T_{m-1}$. The construction implies that $u_i \in L_i$ for each $i = 1, 2, \ldots, m$, the languages are pairwise disjoint, and $B^+ = \cup_{i=1}^m L_i$. Since L_i, $i = 2, 3, \ldots, m-1$, is an intersection of two semigroups, namely T_i and $B^+ \setminus T_{i-1}$, we conclude that $L_i = L_i^+$ for each $i \in \{1, 2, \ldots, m\}$. Thus $\{L_i\}_{i=1}^m$ is a closed partition of B^+ separating the words $u_1, u_2, \ldots, u_m$. The proof of the first part of Theorem 2 is now complete.

In contrast to the fact that separating languages constructed in [2] are not always context-free (the proof is presented in the subsequent section), one can prove that the languages $T_1, T_2, \ldots, T_{m-1}$ are regular, whence also $L_1, L_2, \ldots, L_m$ are regular. We provide an explicit construction of automata accepting languages T_i in the rest of this section. For this, we need some additional concepts.

The binary numeration system (i.e., the binary value of any binary word) can be naturally extended to arbitrary sets of integer digits. Suppose that we have a word $z = z_1 z_2 \cdots z_n$, this time with $z_i \in \mathbb{Z}$ for $i = 1, 2, \ldots, n$, where $n \in \mathbb{N}_+$. Then we can define the (generalized) binary value of the word z by the same formula as for the alphabet $\{0, 1\}$:

$$\mathbf{b}(z) = \mathbf{b}(z_1 z_2 \cdots z_n) = \sum_{i=1}^{n} z_i \cdot 2^{n-i} .$$

The use of the same notation should cause no confusion, since the ordinary binary value is just a special case of the generalized one. In particular, note that

$$\mathbf{b}(uv) = 2^{|v|}\mathbf{b}(u) + \mathbf{b}(v)$$

holds for all $u, v \in \mathbb{Z}^+$.

For each $w \in B^+$ and $p, q \in \mathbb{Z}$, let $w(p, q)$ be the word obtained by replacing in w each occurrence of the symbol 1 with $(p - q)$ and each occurrence of the symbol 0 with $-q$. Then, we have

$$\mathbf{b}(w(p, q)) = p \cdot \mathbf{b}(w) - q \cdot (2^{|w|} - 1).$$

We shall construct a finite automaton accepting the language

$$\begin{aligned} T(p, q) &= \{w \in B^+ \mid p \cdot \mathbf{b}(w) - q \cdot (2^{|w|} - 1) \geq 0\} \\ &= \{w \in B^+ \mid \mathbf{b}(w(p, q)) \geq 0\} . \end{aligned}$$

Note that the set T_i is now equal to $T(p_i, q_i)$. We shall therefore suppose $p > q > 0$. The automaton can be viewed as a device deciding whether a word in the binary numeration basis written with digits $(p-q)$ and $-q$ represents a non-negative number or a negative one.

We need two easy observations:

Lemma 4. *Let $z = z_1 z_2 \cdots z_n$ be a word such that $n \in \mathbb{N}, n \geq 2$ and $z_i \in \mathbb{Z}$ for $i = 1, 2, \ldots, n$. Then the word $z' = (2z_1 + z_2)z_3 z_4 \cdots z_n$ of length $n-1$ in $\mathbb{Z}^+$ satisfies $\mathbf{b}(z') = \mathbf{b}(z)$.*

Proof. Easy:

$$\mathbf{b}(z') = (2z_1 + z_2)2^{n-2} + \sum_{i=3}^{n} z_i \cdot 2^{n-i} = \sum_{i=1}^{n} z_i \cdot 2^{n-i} = \mathbf{b}(z).$$

□

Lemma 5. *Let $x, y \in \mathbb{N}$ and let $z = z_1 z_2 \cdots z_n$, $n \in \mathbb{N}, n \geq 2$, be a word belonging to $\mathbb{Z}\{x, -y\}^+$, i.e., $z_1 \in \mathbb{Z}$ and $z_i \in \{x, -y\}$ for $i = 2, 3, \ldots, n$. If $z_1 \geq y$, then $\mathbf{b}(z) > 0$. If $z_1 \leq -x$, then $\mathbf{b}(z) < 0$.*

Proof. Let $z_1 \geq y$. Then

$$\begin{aligned}\mathbf{b}(z) &= \sum_{i=1}^{n} z_i \cdot 2^{n-i} = z_1 \cdot 2^{n-1} + \sum_{i=2}^{n} z_i \cdot 2^{n-i} \\ &\geq y \cdot 2^{n-1} - \sum_{i=2}^{n} y \cdot 2^{n-i} = y \cdot 2^{n-1} - y \cdot (2^{n-1} - 1) > 0.\end{aligned}$$

The other claim is proved similarly. □

We now describe the finite automaton $\mathcal{A} = (S, B, \delta, s_0, F)$ accepting $T(p, q)$, with $p > q > 0$. The set of states is

$$S = \{-p + q + 1, -p + q + 2, \ldots, 0, 1, \ldots, q - 1\} \cup \{-\infty, \infty\} .$$

The accepting states F consist of ∞ and non-negative integers $\{0, 1, \ldots, q-1\}$, the initial state is $s_0 = 0$.

As mentioned before, the automaton works on nonempty binary words w so that it accepts w if the word $w(p, q)$ represents a non-negative integer, that is, if $\mathbf{b}(w(p, q)) \geq 0$. The state transition function δ motivated by Lemma 4 and Lemma 5 is defined as follows. If $s \in \{-\infty, \infty\}$, then $\delta(s, x) = s$ for both $x \in B$. For $s \in S \setminus \{-\infty, \infty\}$ and $x \in B$ we define:

$$\delta(s, x) = \begin{cases} 2s + x(p-q, -q), & -p + q < 2s + x(p-q, -q) < q, \\ \infty, & 2s + x(p-q, -q) \geq q, \\ -\infty, & 2s + x(p-q, -q) \leq -p + q \ , \end{cases}$$

where $+$ naturally means normal (integer) addition. Note that $x(p-q, -q)$ denotes integer $p - q$ or $-q$ depending on whether $x = 1$ or $x = 0$ respectively.

Lemma 6. *The automaton $\mathcal{A} = (S, B, \delta, s_0, F)$ accepts the language $T(p, q) = \{w \in B^+ \mid \mathbf{b}(w(p, q)) \geq 0\}$.*

Proof. Note first, that if the automaton enters the state ∞, then it remains there and always accepts, and, similarly, if it enters the state $-\infty$, then it always rejects.

Suppose now that the automaton is in a state s that is a finite integer, and the suffix of the input word w not yet processed is xw', $x \in B$. If $\delta(s,x)$ is finite, then Lemma 4 implies that

$$\mathbf{b}(\delta(s,x) \cdot w'(p,q)) = \mathbf{b}(s \cdot (xw')(p,q)),$$

where $\cdot$ denotes the concatenation of words. Since

$$\mathbf{b}(0 \cdot w(p,q)) = \mathbf{b}(w(p,q)),$$

we obtain, by induction, that

$$\mathbf{b}(\delta(s,x) \cdot w'(p,q)) = \mathbf{b}(w(p,q)).$$

Therefore the decision of the automaton is correct if it ends in a finite integer state. In such case we have even calculated $\mathbf{b}(w(p,q))$.

On the other hand, Lemma 5 implies that the automaton enters the state ∞ (and eventually accepts) only if $\mathbf{b}(w(p,q)) > 0$. Similarly, the automaton enters the state $-\infty$ (and rejects) only if $\mathbf{b}(w(p,q)) < 0$. This completes the proof of the lemma and thus of Theorem 2. □

We wish to note that the situation becomes dramatically more complicated if we allow an infinite number of words. Consider, for example, the set `Prim` of all primitive words over A, which, obviously, pairwise do not commute. Then the set

$$\left\{ \frac{\mathbf{b}(u)}{2^{|u|}-1} \mid u \in \mathtt{Prim} \right\}$$

is a dense subset of the interval $[0,1]$. To see this, note that any number $r \in [0,1]$ with the infinite binary expansion $0.\alpha$, $\alpha \in \{0,1\}^\omega$, can be arbitrarily well approximated as

$$r \approx \frac{\mathbf{b}(\alpha')}{2^{|\alpha'|}} \approx \frac{\mathbf{b}(\alpha')}{2^{|\alpha'|}-1},$$

where α' is a sufficiently long prefix of α. If α' is not primitive, it can be replaced with its primitive root thanks to Lemma 3.

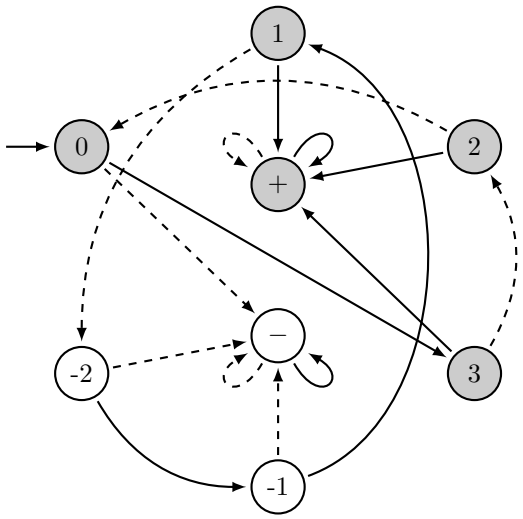
We conclude this section with two examples.

Example. Consider separation of words $u = 1000$ and $v = 1001$. We have

$$\frac{\mathbf{b}(v)}{2^{|v|}-1} = \frac{9}{15} > \frac{4}{7} > \frac{8}{15} = \frac{\mathbf{b}(u)}{2^{|u|}-1}.$$

Therefore we can choose $p = 7$ and $q = 4$. The automaton separating words u and v is given by the following figure, where the plus state stays for ∞ and the minus state for $-\infty$, normal arrows represent transitions corresponding to 1, and dashed

arrows transitions corresponding to 0. Accepting states are grey.



For the input word w the automaton tries to calculate $\mathbf{b}(w(7,4))$ and accepts the language $T(7,4)$.

Example. Consider words $u = 0^{n-1}1$ and $v = 0^{n-2}11$. In order to separate them, we need positive integers p and q satisfying

$$\frac{3}{2^n-1} \geq \frac{q}{p} > \frac{1}{2^n-1}.$$

For any such integers we have $p > 2^{n-2}$. This example shows that the size of our automaton separating two words can be exponential in the length of the words.

4. Geometrical considerations

In this section we provide an alternative view of the separation by closed partitions, which stems from a geometrical intuition. We discuss the way in which this intuition is applied in [2] in order to separate two noncommuting words. A natural limitation of the geometrical approach leads to an inductive construction, which obscures the original geometrical insight. We show by an example that the resulting separating languages can be strongly context-sensitive, that is, not context-free.

Assume that u and v are nonempty words over the n-symbol alphabet $A = \{a_1, a_2, \ldots, a_n\}$, $n \in \mathbb{N}_+$. Suppose furthermore that the vectors $\overline{u} = \Psi(u)$ and $\overline{v} = \Psi(v)$ are linearly independent (over $\mathbb{Q}$, the rationals). Recall, that for each pair $\overline{x} = (x_1, x_2, \ldots, x_n)$, $\overline{y} = (y_1, y_2, \ldots, y_n)$ of vectors in the vector space $\mathbb{Q}^n$, the *inner product* of $\overline{x}$ and $\overline{y}$ is defined by

$$\overline{x} \cdot \overline{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Choose an integer vector $\overline{p} = (p_1, p_2, \ldots, p_n)$ so that $\overline{u} \cdot \overline{p} = 0$ and $\overline{v} \cdot \overline{p} > 0$. By the basic results of linear algebra this can be done since $\overline{u}$ and $\overline{v}$ are linearly independent. This can be viewed as stating an $(n-1)$-dimensional hyperplane (with normal vector $\overline{p}$) that goes through the origin and splits the vector space $\mathbb{Q}^n$ into two parts; the vector (point) $\overline{u}$ lies in the hyperplane and the vector $\overline{v}$ on the

"positive" side of it. Let

$$S_1 = \{\overline{z} \in \mathbb{N}^n \mid \overline{z} \cdot \overline{p} \leq 0\} \qquad \text{and} \qquad S_2 = \{\overline{z} \in \mathbb{N}^n \mid \overline{z} \cdot \overline{p} > 0\}.$$

Obviously both S_1 and S_2 are subsemigroups of $\mathbb{N}^n$ and $\mathbb{N}^n = S_1 \cup S_2$. This means that the commutative languages $L_1 = \Psi^{-1}(S_1)$ and $L_2 = \Psi^{-1}(S_2)$ form a closed partition (L_1, L_2) of A^+ separating u and v. Clearly, L_1 and L_2 are commutative one-counter languages and thus context free (for more examples of this type of languages see [6]).

Theorem 7. *Let $u, v \in A^+$ be two words the Parikh images of which are linearly independent. Then there exists a closed partition of A^+ into two commutative context-free languages separating u and v.*

By the facts above, we can extend the separation results in the following way.

Theorem 8. *Let M_1, M_2 be languages over the alphabet $A = \{a_1, a_2, \ldots, a_n\}$ of n letters, $n \in \mathbb{N}_+$, and $\overline{p} \in \mathbb{Z}^n$ a vector such that $\Psi_n(u) \cdot \overline{p} \leq 0$ and $\Psi_n(v) \cdot \overline{p} > 0$ for all $u \in L_1$ and $v \in L_2$. Then there exists a closed partition of A^+ into two commutative context-free languages separating M_1 and M_2.*

A natural question arises whether or not Theorem 7 holds for any sequence $u_1, u_2, \ldots, u_m$ of words over a n-symbol alphabet A such that the Parikh images of u_j and u_k are linearly independent whenever $j \neq k$. We then would have $(n-1)$-dimensional hyperplanes $T_1, T_2, \ldots, T_{m-1}$ separating the words from each other. Moreover, the planes should bound sets whose inverse Parikh maps are context-free languages. To guarantee this, it should be sufficient (by the main result of [1]) that each pair of hyperplanes contains only one common point in $\mathbb{N}^n$, namely $\overline{0}$. It seems that the choice of planes can be carried out as described, but since we do not have an exact proof, we state the following

Conjecture 9. *Let $m \in \mathbb{N}_+$ and $u_1, u_2, \ldots, u_m \in A^+$ be words such that for all $j, k \in \{1, 2, \ldots, m\}$, $j \neq k$, the Parikh images of u_j and u_k are linearly independent. Then there exists a closed partition of A^+ into m commutative context-free languages separating the words $u_1, u_2, \ldots, u_m$.*

The paper [2] shows that two words can be separated by a closed partition if and only if they do not commute, which corresponds to the first part of our Theorem 2, restricted to the case $m = 2$. The construction of the closed partition in [2] is carried out inductively as follows.

Let u and v be the noncommuting words we want to separate. Assume that $|u| + |v| = 2$. Then, since u and v do not commute, we have $u = a$ and $v = b$ for some distinct letters $a, b \in A$. By choosing $L = a^+$ we get a closed partition $(L, A^+ \setminus L)$ separating u and v.

Let $|u| + |v| > 2$. We have two cases: 1° There exists $a \in A$ such that $\frac{|u|_a}{|u|} \neq \frac{|v|_a}{|v|}$; and 2° For all $a \in A$ the equality $\frac{|u|_a}{|u|} = \frac{|v|_a}{|v|}$ holds.

Case 1°. Let $a \in A$ be such that $\frac{|u|_a}{|u|} \neq \frac{|v|_a}{|v|}$. Assume without loss of generality that $\lambda = \frac{|u|_a}{|u|} < \frac{|v|_a}{|v|}$. Let $L = \{w \in A^+ \mid |w|_a \leq \lambda |w|\}$. Then $(L, A^+ \setminus L)$ is a closed partition separating u and v.

Case 2°. Let $a \in A$ be such that $\lambda = \frac{|u|_a}{|u|} = \frac{|v|_a}{|v|} > 0$. Certainly $\lambda < 1$; otherwise $u, v \in a^+$ and they commute. Let $\lambda = \frac{k}{m}$ where $k, m \in \mathbb{N}_+$, $m \geq 2$ and

$\gcd(k,m)=1$. Let Δ be a new alphabet of the size $|A|^m$. Let $g:\Delta\to A^m$ be a bijection and $\Phi:\Delta^+\to A^+$ the morphism induced by g. Let $r,s\in\Delta^+$ be such that $\Phi(r)=u$ and $\Phi(s)=v$. Obviously r and s do not commute; otherwise u and v would commute. Repeat the procedure for r and s to find a closed partition $(L,\Delta^+\setminus L)$ separating r and s. Let

$$A^{=}=\{w\in A^+\,|\,|w|_a=\lambda|w|\}\qquad\text{and}\qquad A^{<}=\{w\in A^+\,|\,|w|_a<\lambda|w|\},$$

and let

$$M=(\Phi(L)\cap A^{=})\cup A^{<}.$$

Then $(M,A^+\setminus M)$ is a closed partition separating u and v.

It is not difficult to see that both languages in the above constructed closed partition separating u and v are context-sensitive: The languages separating the rudimentary words r and s on which the construction of the closed partition separating u and v is inductively based, are either regular (if $|r|_a=0$ or $|s|_a=0$) or context-free (if $\mathrm{alph}(r)=\mathrm{alph}(s)$ but there exists $a\in\mathrm{alph}(r)$ such that $\frac{|r|_a}{|r|}\neq\frac{|s|_a}{|s|}$). The languages of the type $A^{=}$ and $A^{<}$ are context free as well. Since the family of context-sensitive languages is closed under nonerasing morphisms and intersection, the inductive process preserves context-sensitiveness. Context-sensitive languages are also closed under complement, so both resulting languages are context-sensitive.

One may pose a natural question whether or not the procedure possibly outputs always context-free languages. The following example shows that this is not the case. To understand the considerations completely, the reader should have some experience on semilinear sets and the structure of so called strictly bounded context-free languages as presented in [4].

Example. Let $u=abbaba$ and $v=baabab$. We shall show that the procedure previously described outputs a context-sensitive language that is not context-free. Choose $A=\{a,b\}$. Since $\Psi_2(u)=\Psi_2(v)=(3,3)$ and

$$\frac{|u|_a}{|u|}\;=\;\frac{|v|_a}{|v|}\;=\frac{1}{2}\quad\left(=\;\frac{|u|_b}{|u|}\;=\;\frac{|v|_b}{|v|}\right),$$

we may state $\lambda=\frac{1}{2}$ and $m=2$. Let $\Delta=\{a_1,a_2,a_3,a_4\}$, so that $|\Delta|=|A|^2$. This leads us to define the (injective) morphism $\Phi:\Delta^+\to A^+$ by

$$\Phi(a_1)=aa,\qquad\Phi(a_2)=bb,\qquad\Phi(a_3)=ab,\qquad\text{and}\qquad\Phi(a_4)=ba.$$

Let $r=a_3a_4a_4$ and $s=a_4a_3a_3$, so that

$$\Phi(r)=\Phi(a_3a_4a_4)=abbaba\qquad\text{and}\qquad\Phi(s)=\Phi(a_4a_3a_3)=baabab.$$

Now $\Psi_4(r)=(0,0,1,2)$ and $\Psi_4(s)=(0,0,2,1)$, and the language

$$L\;=\;\{\,w\in\Delta^+\,\big|\,|w|_{a_3}\leq\frac{1}{3}|w|\,\}$$

induces a closed partition $(L,\Delta^+\setminus L)$ separating r and s. Let

$$A^{=}=\{w\in A^+\,\big|\,|w|_a=\frac{1}{2}|w|\}\qquad\text{and}\qquad A^{<}=\{w\in A^+\,\big|\,|w|_a<\frac{1}{2}|w|\}.$$

Then the language

$$M\;=\;(\Phi(L)\cap A^{=})\,\cup\,A^{<}$$

gives us a closed partition $(M,A^+\setminus M)$ separating u and v.

We shall next show that M is not context-free. Consider the language $B = M \cap R$ where $R = (aa)^*(bb)^*(ab)^*$. Then $B = B_1 \cup B_2$ where

$$B_1 = \Phi(L) \cap A^{=} \cap R \qquad \text{and} \qquad B_2 = A^{<} \cap R.$$

Obviously B_1 consists of all words $(aa)^{n_1}(bb)^{n_2}(ab)^{n_3}$ such that $n_1,\ n_2,\ n_3 \in \mathbb{N}$, $3n_3 \leq n_1 + n_2 + n_3$, and $2n_1 + n_3 = 2n_2 + n_3$. Then

$$\begin{aligned} B_1 &= \{(aa)^{n_1}(bb)^{n_2}(ab)^{n_3} \,|\, n_1, n_2, n_3 \in \mathbb{N} \wedge\ 2n_3 \leq n_1 + n_2\ \wedge\ n_1 = n_2\,\} \\ &= \{(aa)^m(bb)^m(ab)^p \,|\, m, p \in \mathbb{N}\ \wedge\ m \geq p\,\}\ . \end{aligned}$$

On the other hand,

$$B_2\ =\ \{(aa)^m(bb)^n(ab)^p \,|\, m, n, p \in \mathbb{N} \wedge\ m < n\,\}\ .$$

Let $L' = \Phi^{-1}(B)$. Then

$$\begin{aligned} L' &= \Phi^{-1}(B_1) \cup \Phi^{-1}(B_2) \\ &= \{a_1^m a_2^m a_3^p \,|\, m, p \in \mathbb{N} \wedge m \geq p\,\} \ \cup\ \{a_1^m a_2^n a_3^p \,|\, m, n, p \in \mathbb{N} \wedge m < n\,\}\ . \end{aligned}$$

Consider the Parikh image $P = \Psi_3(\Phi^{-1}(B))$ of L'. Obviously $P = P_1 \cup P_2$ where $P_1 = \Psi_3(\Phi^{-1}(B_1))$ and $P_2 = \Psi_3(\Phi^{-1}(B_2))$. It is straightforward to see that

$$P_1\ = \{(m, m, p) \,|\, m, p \in \mathbb{N} \wedge m \geq p\,\}\ =\ \{m(1,1,1) + p\,(1,1,0) \,|\, m, p \in \mathbb{N}\,\}$$

and, as well, that

$$\begin{aligned} P_2 &= \{(m, n, p) \,|\, m, n, p \in \mathbb{N} \wedge m < n\,\} \\ &= \{m(1,1,0) + n(0,1,0) + p\,(0,0,1) \,|\, m, p \in \mathbb{N},\ n \in \mathbb{N}_+\,\}\ . \end{aligned}$$

Certainly $P_1 \cap P_2 = \emptyset$. The structure of P implies that in each presentation of P as a finite union of linear sets, (at least) one of the linear sets has to contain a period with three nonzero coordinates. Thus P is not a finite union of so called stratified linear sets (for the definition of the concept, see [4], page 159). Theorem 5.4.2 in [4] then implies that L' is not a context-free language. It is a well-known fact that the family of context-free languages is closed under inverse morphisms and intersection with regular sets, so we deduce that M is not context-free.

Note that the use of techniques exploiting geometrical properties of strictly bounded context-free languages cannot be easily avoided since the language L' satisfies the Pumping Lemma (even in its stronger form given by Ogden).

5. Concluding remarks

We have shown how to separate two noncommuting words by regular languages closed under catenation. We want to point out that automata we constructed resemble those used to accept solutions of linear Diophantine inequalities (see [3]).

Let us once more compare our approach to the approach used in [2]. Their construction has a nice geometrical interpretation, which, however, works only if the Parikh maps of the separated words are linearly independent. The price paid for the illustrative nature of the separation is rather complicated induction in case when the Parikh maps are linearly dependent, leading to languages which can be not context-free. The basic reason can be seen in the fact that the Parikh map counts the letters completely ignoring their positions. In contrast, the binary representation gives each position a specific weight.

Acknowledgements

The work on this paper has been supported by the research project MSM 0021620839.

The authors would like to thank an anonymous referee for valuable comments, which greatly helped to improve the presentation.

References

[1] J. Beauquier, M. Blattner and M. Latteux. On commutative context-free languages, *J. Comput. Syst. Sci.* **35** (1987), 311–320.
[2] J. Brzozowski, E. Grant and J. Shallit. Closures in formal languages and Kuratowski's Theorem, LNCS 5583 (2009), 125–144.
[3] A. Boudet and H. Comon. Diophantine equations, Presburger arithmetic and finite automata, LNCS 1059 (1996), 30–43.
[4] S. Ginsburg, *The Mathematical Theory of Context Free Languages*, McGraw-Hill, New York, 1966.
[5] P. C. Hammer. Kuratowski's closure theorem. *Nieuw Archief v. Wiskunde* **7** (1960), 74–80.
[6] J. Kortelainen. Remarks about commutative context-free languages. *J. Comput. Syst. Sci.* **56** (1998), 125–129.
[7] C. Kuratowski. Sur l'opération $\overline{A}$ de l'analysis situs. *Fund. Math.* **3** (1922), 182–199.
[8] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, Reading Massachusetts, 1983.
[9] D. Peleg. A generalized closure and complement phenomenon. *Discrete Math.* **50** (1984), 285–293.

Department of Algebra, Charles University in Prague,, Prague, Czech Republic
E-mail address: holub@karlin.mff.cuni.cz

Department of Information Processing Science, University of Oulu,, Oulu, Finland
E-mail address: jkortela@tols16.oulu.fi