数理解析研究所講究録
第 695 巻 1989 年 45-54

On Learning Equal Matrix Languages *

Yuji Takada

*International Institute for Advanced Study of
Social Information Science (IIAS-SIS)
FUJITSU LIMITED
140, Miyamoto, Numazu, Shizuoka 410-03, Japan*

1 Introduction

In this paper, we consider the learning problem for a restricted family of matrix languages called *strongly bounded equal matrix languages*. The languages consist of strings of the form $a_1^{n_1} \cdots a_m^{n_m}$, where each a_i is a symbol and n_i is a nonnegative integer, and are defined in terms of certain parallel rewriting grammars called *equal matrix grammars*. Also, the languages closely related to semilinear subsets of the Cartesian product of nonnegative integers. The family contains a language which is not context-free and does not contain any context-free languages.

We show that (1) the family of strongly bounded equal matrix languages is not learnable from positive examples, while there exists a meaningful subfamily which is learnable from positive examples, (2) given any teacher called an *ideal* teacher, who presents elements of any language L for the question whether $L \subseteq L(G)$ for any grammar G and eventually gives sufficient examples for learning, the subfamily is learnable in polynomial time of the size of inputs.

2 Preliminaries

Let Σ be an *alphabet*, i.e., a finite set of symbols and Σ^* be the set of all strings over Σ containing the null string λ. For each string w, $w^0 = \lambda$ and $w^i = w^{i-1}w$ for each integer $i \geq 1$, and $w^* = \{w^i \mid i \geq 0\}$. A *language* over Σ is a subset of Σ^*.

Definition A language over an alphabet Σ is said to be *strongly bounded* if and only if $L \subseteq a_1{}^* \cdots a_k{}^*$ where $\Sigma = \{a_1, \ldots, a_k\}$.

*This is a part of the work in the major R&D of the Fifth Generation Computer Project, conducted under program set up by MITI.

Definition An *equal matrix grammar* (abbreviated *EMG*) *of order* k is a 4-tuple $G = (N, \Sigma, \Pi, S)$, where

1. S is the *initial symbol.*
2. N is a finite nonempty set consisting of k-tuples $(A_1, A_2, \dots, A_k)$, called a *nonterminal,* such that for any pair $(A_1, A_2, \dots, A_k)$ and $(B_1, B_2, \dots, B_k)$ of N, $\{A_1, A_2, \dots, A_k\} \cap \{B_1, B_2, \dots, B_k\} = \emptyset$.
3. Π is a finite nonempty set consisting of the following types of *matrix rules*;
   (a) $[S \rightarrow w_1 A_1 w_2 A_2 \cdots w_k A_k]$,
   (b) $[A_1 \rightarrow w_1 B_1,\ A_2 \rightarrow w_2 B_2,\ \dots A_k, \rightarrow w_k B_k]$,
   (c) $[A_1 \rightarrow w_1,\ A_2 \rightarrow w_2,\ \dots,\ A_k \rightarrow w_k]$,
   where S is the initial symbol, and $(A_1, A_2, \dots, A_k)$, $(B_1, B_2, \dots, B_k)$ are nonterminals, $w_1, w_2, \dots, w_k \in \Sigma^*$.

An *equal matrix grammar* is an *EMG* of any *finite* order k.

We denote $\Sigma \cup N \cup \{S\}$ by V.

Let $G = (N, \Sigma, \Pi, S)$ be an *EMG* of order k. We define the relation $\Longrightarrow$ between strings in V^*. For any $x, y \in V^*$, $x \Longrightarrow y$ if and only if either (1) x is the initial symbol S and the initial matrix rule $[S \rightarrow y]$ is in Π or (2) there exist strings $u_1, \dots, u_k$, $v_1, \dots, v_k$ over Σ such that $x = u_1 A_1 v_1 \cdots u_k A_k v_k$, $y = u_1 z_1 v_1 \cdots u_k z_k v_k$, and the matrix rule $[A_1 \rightarrow z_1, \cdots, A_k \rightarrow z_k]$ is in Π. $\overset{*}{\Longrightarrow}$ denotes the reflexive and transitive closure of $\Longrightarrow$.

The *language generated by* G, denoted $L(G)$, is the set $L(G) = \{w \in \Sigma^* \mid S \overset{*}{\Longrightarrow} w\}$.

Definition A language L is said to be an *equal matrix language* (abbreviated *EML*) if and only if there exists an *EMG* G such that $L = L(G)$ holds.

In this paper, we consider the learning problem for a *strongly bounded equal matrix language* (abbreviated *SBEML*). The family of *SBEML*s contains context-sensitive languages. For example, the context-sensitive language $\{a^n b^n c^n \mid n \geq 1\}$ is an *SBEML*. Also, there exists a context-free language which is not an *SBEML*. For example, the context-free language $\{a^n b^n \mid n \geq 1\}^*$ is not an *SBEML* (Ibarra [3]).

3 Algebraic Characterization

Let $\mathcal{N}$ denote the nonnegative integers. For each integer $k \geq 1$, let $\mathcal{N}^k = \mathcal{N} \times \cdots \times \mathcal{N}$ (k times) and for each $n \in \mathcal{N}$, $n^k = (n, \dots, n)$ (k times). We regard $\mathcal{N}^k$ as a subset of the vector space of all k-tuples of rational numbers over the rational numbers.

Given an element $\mathbf{c}$ and a subset P of $\mathcal{N}^k$, let $Q(\mathbf{c}, P)$ denote the set

$$Q(\mathbf{c}, P) = \{\mathbf{q} \mid \mathbf{q} = \mathbf{c} + n_1\mathbf{p}_1 + \cdots + n_r\mathbf{p}_r,\ n_i \in \mathcal{N},\ \mathbf{p}_i \in P\}.$$

$\mathbf{c}$ is called the *constant* and each $\mathbf{p}_i$ is called a *period* of $Q(\mathbf{c}, P)$.

A subset Q of $\mathcal{N}^k$ is said to be *linear* if and only if there exist an element $\mathbf{c}$ and a finite subset P of $\mathcal{N}^k$ such that $Q = Q(\mathbf{c}, P)$. Q is said to be *semilinear* if and only if Q is the union of a finite number of linear sets. Furthermore, a subset $Q = Q(\mathbf{c}, P)$ of $\mathcal{N}^k$ is said to be *simple* if and only if the elements of P are linearly independent. A subset Q is said to be *semi-simple* if and only if Q is a finite disjoint union of simple sets.

We note that any linear set has more than one description in terms of constants and periods, and so does any semilinear set. Therefore, we distinguish between a semilinear set Q and a description $Q(\mathbf{c}_1, P_1) \cup \cdots \cup Q(\mathbf{c}_n, P_n)$ of Q.

Definition A description $Q(\mathbf{c}, P)$ of a linear set is said to be *canonical* if and only if each period is not linear sum of the other periods. Also, description $Q(\mathbf{c}_1, P_1) \cup \cdots \cup Q(\mathbf{c}_n, P_n)$ of a semilinear set is said to be *canonical* if and only if each description $Q(\mathbf{c}_i, P_i)$ of a linear set is canonical.

Note that for any linear subset Q of $\mathcal{N}^k$, a canonical description $Q(\mathbf{c}, P)$ is unique because $\mathbf{c} \in \mathcal{N}^k$ and P is a finite subset of $\mathcal{N}^k$. We also note that for any linear set Q, a canonical description is effectively found from a description of Q. However, there exists a semilinear subset such that a canonical description is not unique.

The Parikh mapping defined as follows connects *EML*s with semilinear subsets of $\mathcal{N}^k$.

Definition Let $\Sigma = \{a_1, \ldots, a_k\}$ be an alphabet. *The Parikh mapping* $\psi_{(a_1,\ldots,a_k)}$ or ψ when $(a_1, \ldots, a_k)$ is understood, is the function from Σ^* into $\mathcal{N}^k$ defined by $\psi(w) = (\#_{a_1}(w), \ldots, \#_{a_k}(w))$, where $\#_{a_i}(w)$ is the number of occurrences of a_i in w.

We call $\psi(L) = \{\psi(w) \mid w \in L\}$ *the Parikh set of an EML L.*

The following theorem is due to Siromoney [4]:

Theorem 3.1 (Siromoney) *Let $\Sigma = \{a_1, \ldots, a_k\}$ be an alphabet. For any strongly bounded language L over Σ, L is generated by an EMG G of order k if and only if the Parikh set of L is a semilinear subset Q of $\mathcal{N}^k$. Moreover, an EMG G is effectively found from a description of Q and vice versa.*

For any semilinear set Q, an *EMG G* which generates an *SBEML* is effectively constructed from a description of Q in the following manner: It is enough to show the case that Q is a linear set. Let $Q(\mathbf{c}, \{\mathbf{p}_1, \ldots, \mathbf{p}_r\})$ be a description of the linear set Q. Also, let $\mathbf{c} = (c_1, \ldots, c_k)$ and $\mathbf{p}_i = (p_i^1, \ldots, p_i^k)$. Then $G = (N, \Sigma, \Pi, S)$ where $\Sigma = \{a_1, \ldots, a_k\}$, $N = \{(A_1, \ldots, A_k)\}$, and Π consists of the following matrix rules:

$$[S \to a_1^{c_1} A_1 \cdots a_k^{c_k} A_k],\ [A_1 \to \lambda, \ldots, A_k \to \lambda]$$

$$[A_1 \to a_1^{p_i^1} A_1, \ldots, A_k \to a_k^{p_i^k} A_k] \qquad \text{for each } i$$

From Theorem 3.1, we may regard the learning problem for *SBEMLs* as the learning problem for semilinear sets.

From these, we can consider meaningful subfamilies of *SBEMLs*:

Definition For each positive integer n, an *SBEML* L is said to be *n-linears SBEML* if and only if $\psi(L)$ is a union of exactly n linear sets and there is no $i < n$ such that $\psi(L)$ is a union of i linear sets.

Thus, a 1-linear *SBEML* is an *SBEML* whose Parikh set is a linear set.

4 Learnabilities from Positive Examples

On learning of formal languages, Angluin [1] presented a necessary and sufficient condition for languages to be learnable from positive examples.

Condition 1 An indexed family of nonempty languages *satisfies Condition 1* if and only if there exists an effective procedure which on any input $i \geq 1$ enumerates a set of strings T_i such that (1) T_i is finite, (2) $T_i \subseteq L_i$, and (3) for all $j \geq 1$, if $T_i \subseteq L_j$ then L_j is not a proper subset of L_i.

The next theorem shows that Condition 1 is a necessary and sufficient condition for a family of languages to be learnable from positive examples.

Theorem 4.1 (Angluin) *An indexed family of nonempty recursive languages is learnable from positive examples if and only if it satisfies Condition 1.*

The following condition is simply Condition 1 with the requirement of effective enumerability of T_i dropped.

Condition 2 We say an indexed family of nonempty recursive languages $L_1, L_2, L_3, \ldots,$ *satisfies Condition 2* provided that, for every $i \geq 1$, there exists a finite set $T_i \subseteq L_i$ such that for every $j \geq 1$, if $T_i \subseteq L_j$ then L_j is not a proper subset of L_i.

Theorem 4.2 (Angluin) *If $L_1, L_2, L_3, \ldots,$ is an indexed family of recursive languages that is learnable from positive examples, then it satisfies Condition 2.*

This theorem may be used to show that a family of languages is not learnable from positive examples.

We note that the Angluin's results described above are concerned with only the recursiveness of languages. Hence, all of them are applicable to the learning problem for recursive sets, straightforwardly. In the sequel, we apply them to the problem for semilinear subsets of $\mathcal{N}^k$.

Let $\preceq$ be the relation on $\mathcal{N}^k$ defined by $\mathbf{u} \preceq \mathbf{v}$ for elements $\mathbf{u} = (u_1, \ldots, u_k)$ and $\mathbf{v} = (v_1, \ldots, v_k)$ if and only if $u_i \leq v_i$ for each i. The relation $\preceq$ is a partial order on $\mathcal{N}^k$.

Definition Let Q be a linear subset of $\mathcal{N}^k$ and $Q(\mathbf{c}, \{\mathbf{p}_1, \ldots, \mathbf{p}_r\})$ be a canonical description of Q. Then, a *characteristic set of* Q is the finite set

$$C(Q) = \{\mathbf{c}\} \cup \{\mathbf{c} + \mathbf{p}_i \mid 1 \leq i \leq r\}.$$

We note that, given the characteristic set $C(Q)$ of a linear set Q, a canonical description of Q is effectively found. That is, the constant $\mathbf{c}$ is the unique minimum element of $C(Q)$ with respect to $\preceq$ and then the set of periods is $\{\mathbf{p}_i \mid \mathbf{q}_i - \mathbf{c},\ \mathbf{q}_i \in C(Q) - \{\mathbf{c}\}\}$.

Let $Q((c_1, \ldots, c_k), P)$ be a description of a linear subset of $\mathcal{N}^k$. Then, for each element $\mathbf{q} = (q_1, \ldots, q_k)$ of Q, we denote $(q_1 - c_1)^2 + \cdots + (q_k - c_k)^2$ by $|\mathbf{q}|_c$. The next lemma immediately follows from definitions Q and $C(Q)$:

Lemma 4.3 *Let Q be a linear subset of $\mathcal{N}^k$, $Q(\mathbf{c}, P)$ be a canonical description of Q, and $C(Q)$ be the characteristic set of Q. For any element $\mathbf{q}$ of Q such that $\mathbf{q} \notin C(Q)$, there exist periods $\mathbf{p}_1, \ldots, \mathbf{p}_m \in P$ such that for each i, $|\mathbf{q}|_c > |\mathbf{p}_i|_c$ and $\mathbf{q} = \mathbf{c} + n_1\mathbf{p}_1 + \cdots + n_m\mathbf{p}_m$, where each $n_i \geq 1$.*

Lemma 4.4 *Let Q be a linear subset of $\mathcal{N}^k$ and $C(Q)$ be the characteristic set of Q. Then, for any linear subset Q' of $\mathcal{N}^k$, if $C(Q) \subseteq Q'$ then $Q \subseteq Q'$.*

Proof. Let $Q = Q(\mathbf{c}, P)$ be a linear subset of $\mathcal{N}^k$ and $C(Q)$ the characteristic set of Q. Suppose that $Q' = Q(\mathbf{c}', \{\mathbf{p}'_1, \ldots, \mathbf{p}'_r\})$ is a linear subset of $\mathcal{N}^k$ such that $C(Q) \subseteq Q'$. Since $C(Q) \subseteq Q'$, for each $\mathbf{q}_j$ of $C(Q)$, $\mathbf{q}_j = \mathbf{c}' + n_1^j\mathbf{p}'_1 + \cdots + n_r^j\mathbf{p}'_r$. Therefore, for each period $\mathbf{p}_i$ of Q, $\mathbf{p}_i = \mathbf{q}_i - \mathbf{c} = (n_1^i - n_1^c)\mathbf{p}'_1 + \cdots + (n_r^i - n_r^c)\mathbf{p}'_r$. Hence, for each $\mathbf{q} \in Q$, there exist $m_1, \ldots, m_r \in \mathcal{N}$ such that $\mathbf{q} = \mathbf{c}' + m_1\mathbf{p}'_1 + \cdots + m_r\mathbf{p}'_r$. $\square$

Lemma 4.5 *The family of linear subsets of $\mathcal{N}^k$ is learnable from positive examples.*

Proof. Let $Q(\mathbf{c}_1, P_1), Q(\mathbf{c}_2, P_2), Q(\mathbf{c}_3, P_3), \ldots,$ be an effective enumeration of all descriptions of linear sets. It is obvious that there exists an effective procedure which on any input $i \geq 1$ enumerates a characteristic set C_i of a linear set $Q(\mathbf{c}_i, P_i)$. By definition of characteristic sets of linear sets, C_i is finite and $C_i \subseteq Q(\mathbf{c}_i, P_i)$. Moreover, by Lemma 4.4, for all $j \geq 1$, if $C_i \subseteq Q(\mathbf{c}_j, P_j)$ then $Q(\mathbf{c}_j, P_j)$ is not a proper subset of $Q(\mathbf{c}_i, P_i)$. Therefore, the family satisfies Condition 1 and by Theorem 4.1 the proof is completed. $\square$

Corollary 4.6 *The family of simple subsets of $\mathcal{N}^k$ is learnable from positive examples.*

Since for each $Q(\mathbf{c}_i, P_i)$ there exists an effective enumeration $\psi_{i1}, \psi_{i2}, \psi_{i3}, \ldots$, of all Parikh mapping, by an obvious dovetailing, $L_{11}, L_{21}, L_{21}, \ldots, L_{ij}, \ldots$, is an indexed family of 1-linear *SBEMLs*, where $Q(\mathbf{c}_i, P_i) = \psi_{ij}(L_{ij})$. Therefore, from Theorem 3.1 and Lemma 4.5, we have the following theorem.

Theorem 4.7 *The family of* 1*-linear SBEMLs is learnable from positive examples.*

On the other hand, for $n \geq 2$, the family of n-linears *SBEMLs* is not learnable from positive examples, as shown in the followings:

Lemma 4.8 *The family of semilinear subsets of $\mathcal{N}^k$ consisting of two linear sets is not learnable from positive examples.*

Proof. Consider the semilinear set $Q = Q_1 \cup Q_2$, where $Q_1 = Q((0,0), \emptyset)$ and $Q_2 = Q((1,1), \{(1,0),(0,1)\})$. Let $T = \{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ be any nonempty finite subset of Q. Consider the semilinear set $Q^T = Q_1^T \cup Q_2^T$, where

$$\begin{aligned} Q_1^T &= Q((1,1), \{\mathbf{q}_i - (1,1) \mid \mathbf{q}_i = (1,m) \in T\}) \\ Q_2^T &= Q((0,0), \{\mathbf{q}_j \in T \mid \mathbf{q}_j = (n_1, n_2),\ n_1 \neq 1\}). \end{aligned}$$

Clearly, $T \subseteq Q^T$ and it is easy to verify that $Q^T \subseteq Q$. For each $\mathbf{q}_i \in T$ let $\mathbf{q}_i = (n_1^i, n_2^i)$. Let n_1^m be the maximum integer of $n_1^1, \ldots, n_1^n$. Then, $\mathbf{q}_n = (n_1^m + 1, 1)$ is in Q but not in Q^T, so Q^T is a proper subset of Q. Thus Condition 2 fails. □

The following lemma is proved by the trivial extension of the proof of Lemma 4.8.

Lemma 4.9 *For each $n \geq 2$, the family of semilinear subsets of $\mathcal{N}^k$ consisting of n linear sets is not learnable from positive examples.*

Proof. Let n be an integer greater than 2. Consider the semilinear subset $Q = Q_1 \cup \cdots \cup Q_n$ of $\mathcal{N}^2$, where for l $(1 \leq l \leq n-1)$, $Q_l = Q((l-1,0), \emptyset)$ and $Q_n = Q((n-1,1), \{(1,0),(0,1)\})$. Let $T = \{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$ be any nonempty finite subset of Q. Consider the semilinear set $Q^T = Q_1^T \cup \cdots \cup Q_n^T$, where

$$\begin{aligned} Q_l^T &= Q((l-1,0), \emptyset) \quad \text{for } 1 \leq l \leq n-2 \\ Q_{n-1}^T &= Q((n-1,1), \{\mathbf{q}_i - (n-1,1) \mid \mathbf{q}_i = (n-1,m) \in T\}) \\ Q_n^T &= Q((n-2,0), \{\mathbf{q}_i \in T \mid \mathbf{q}_i = (n_1,n_2),\ n_1 \neq n-1\}) \end{aligned}$$

From the proof of Lemma 4.8, it is easy to verify that $T \subseteq Q^T$ and Q^T is a proper subset of Q. Thus Condition 2 fails. □

The next theorem follows from Theorem 3.1 and Lemma 4.9.

Theorem 4.10 *For each $n \geq 2$, the family of n-linears SBEMLs is not learnable from positive examples.*

Corollary 4.11 *The family of SBEMLs is not learnable from positive examples.*

Procedure *ID1*
Input: A positive presentation $s_1, s_2, s_3, \ldots,$ of a 1-linear *SBEML* L.
Output: A sequence $G_1, G_2, G_3, \ldots,$ of *EMG*s.

Let $E_0 := \emptyset$ and $Q_0 := Q(0^k, \emptyset)$;
For each $i \geq 1$ **do**
  Read $(+, w_i)$;
  $E_i := E_{i-1} \cup \{\psi(w_i)\}$;
  If Q_{i-1} is consistent with E_i
  then $G_i := G_{i-1}$, $Q_i := Q_{i-1}$, output G_i and go to $i+1$ step;
  If found a unique minimum element $\mathbf{q}$ of E_i with respect to $\preceq$
  then let $\mathbf{q}$ be a constant of Q_i
  else let 0^k be a constant of Q_i;
  While Q_i is not consistent with E_i **do**
    find $\mathbf{q} \in E_i$ such that $\mathbf{q} \notin Q_i$ and $|\mathbf{q}|_c$ is minimum;
    add new period $\mathbf{q} - \mathbf{c}$ to Q_i;
  Construct an *EMG* G_i from Q_i and output G_i;
  go to $i+1$ step;

Figure 1: The learner *ID1*

5 A Simple Learning Method for 1-linear SBEMLs

Let L be an unknown 1-linear *SBEML* over an alphabet Σ. As described in the previous sections, if the characteristic set of a linear set $\psi(L)$ is found, then an *EMG* which generates L is effectively found. Therefore, the learner *ID1*, illustrated in Figure 1, tries to find the characteristic set from the given examples. *ID1* outputs the same *EMG* as a conjecture while it is consistent with the given examples. When a conjecture is not consistent with the examples, *ID1* constructs a new conjecture.

Definition Let L be a 1-linear *SBEML*. A *representative sample* $R(L)$ of L is a finite subset of L such that $\psi(R(L))$ contains the characteristic set of the linear set $\psi(L)$.

Lemma 5.1 *Let L be a 1-linear SBEML. Given a representative sample of L, the learner ID1 constructs an EMG G which generates L.*

Proof. We shall show that, given a representative sample of L, *ID1* constructs a description of a linear set $Q = \psi(L)$. Since $\psi(R(L))$ contains the characteristic set of Q, *ID1* finds a unique minimum element of it with respect to $\preceq$, which is precisely a constant $\mathbf{c}$ of a description of Q. Also, Lemma 4.3 and the construction of *ID1* ensure that *ID1* finds each period $\mathbf{p}_i$ of a canonical description of Q in order of smaller size of $|\mathbf{p}_i|_c$. $\square$

Since for any positive presentation $\sigma = s_1, s_2, s_3, \ldots$, there exists a positive integer i such that the set of strings appearing in $s_1, s_2, \ldots, s_i$ is a representative sample of L, by Lemma 5.1, we have the following theorem:

Theorem 5.2 *The learner ID1 identifies any 1-linear SBEML in the limit from positive examples.*

Unfortunately, *ID1* uses membershipness of examples, which is an NP-complete problem, so *ID1* is time-consuming. If there is a polynomial-time algorithm to solve the problem of finding a canonical description of a linear set consistent with the given examples, then we could have a learner which makes a conjecture in polynomial time for each time and identifies any 1-linear *SBEML* in the limit. However, we give some partial evidence for the difficulty of the case.

Theorem 5.3 *If $P \neq NP$, then there is no polynomial-time algorithm to solve the following problem: given a finite subset E of $\mathcal{N}^k$, find a canonical description $Q(\mathbf{c}, P)$ of a linear subset of $\mathcal{N}^k$ which contains all elements of E.*

Proof. Suppose that there exists an algorithm A that runs in polynomial time and is such that for any subset E of $\mathcal{N}^k$, A on input E outputs a canonical description $Q(\mathbf{c}, P)$ of a linear subset of $\mathcal{N}^k$ which contains all elements of E. We shall use A to construct a polynomial-time algorithm to decide whether $\mathbf{q} \in Q(\mathbf{c}, P)$ for an arbitrary element $\mathbf{q} \in \mathcal{N}^k$ and a canonical description $Q(\mathbf{c}, P)$. Since this latter problem is NP-complete, this will imply $P = NP$, proving the theorem.

Let $\mathbf{q}$ be an element in $\mathcal{N}^k$ and $Q(\mathbf{c}, P)$ be a canonical description of a linear subset of $\mathcal{N}^k$. We may construct the characteristic set C of $Q(\mathbf{c}, P)$ in polynomial time. Run A on input $C \cup \{\mathbf{q}\}$ and denote the output by $Q(\mathbf{c}', P')$. Since a canonical description is unique for any linear set, if $\mathbf{c}' = \mathbf{c}$ and $P = P'$ then $\mathbf{q} \in Q(\mathbf{c}, P)$, otherwise, $\mathbf{q} \notin Q(\mathbf{c}, P)$. We may test whether $\mathbf{c} = \mathbf{c}'$ and $P = P'$ in polynomial time, we complete the proof. □

Thus, as far as based on linear sets, it seems that the learning problem for 1-linear *SBEMLs* is computationally intractable.

Remark It is easy to verify that all processes of *ID1* other than the consistency check are done in polynomial time of the size of inputs.

Consider the family of *SBEMLs* such that the Parikh sets of any language in the family is a simple set. This family is also learnable from positive examples by Corollary 4.6. Since the membership problem of simple sets is solvable in polynomial time, for each time i, *ID1* constructs an *EMG* in polynomial time of i, k, and m. Therefore, from the above remark, we have the following:

Procedure *ID1S*
Input: A positive presentation $s_1, s_2, s_3, \ldots$, of a 1-linear *SBEML* L.
Output: A sequence $G_1, G_2, G_3, \ldots$, of *EMG*s.

Let $E_0 := \emptyset$ and $Q_0 := Q(0^k, \emptyset)$;
For each $i \geq 0$ **do**
  Construct an *EMG* G_i from Q_i;
  Ask the ideal teacher whether $L \subseteq L(G_i)$;
  If the teacher replies *yes*
  then output G_i and halt
  Read $(+, w_i)$;
  $E_i := E_{i-1} \cup \{\psi(w_i)\}$;
  If found a unique minimum element $\mathbf{q}$ of E_i with respect to $\preceq$
  then let $\mathbf{q}$ be a constant of Q_i
  else let 0^k be a constant of Q_i;
  For each element $\mathbf{q}$ in E_i **do**
    let $\mathbf{q} - \mathbf{c}$ be a new period of Q_i;
  go to $i + 1$ step;

Figure 2: The learner *ID1S*

Theorem 5.4 *For the family of SBEMLs such that the Parikh set of any language in the family is a simple set, there exists a learner which, for each time* i $(i \geq 1)$, *constructs an EMG* G *in polynomial time of* i, k *and* m, *where* k *is the cardinality of* Σ *and* m *is the maximum length of the given examples.*

6 Learning 1-linear SBEMLs with an Ideal Teacher

In the previous section, we had no assumption on presentations of examples. In this time, we assume that there exists a teacher who can answer questions of a learner and the learner get informations from the teacher.

Let L be an unknown *SBEML*. An *ideal teacher* gives informations to a learner on the following conditions: (1) for any question whether $L \subseteq L(G)$, the ideal teacher answers *yes* if $L \subseteq L(G)$ and *no* otherwise. In addition, if the answer is *no*, the teacher gives an element $s \in L - L(G)$ to the learner. (2) Eventually, the set of examples given by the ideal teacher constitutes a representative sample of L. Note that an ideal teacher gives only positive examples.

For each time i $(i \geq 0)$, the learner *ID1S*, illustrated in Figure 2, asks whether $L \subseteq L(G_i)$ to the teacher. If the answer is *yes*, then *ID1S* outputs G_i and halts. Otherwise, *ID1S* reads a new example and reconstructs a description from the given examples.

The learner *ID1* constructs a new conjecture only if a current conjecture is not consistent with the examples, while the learner *ID1S* does so each time when an ideal teacher gives a new example. *ID1S* constructs a conjecture in the same way as *ID1* does. Therefore, as we have shown in Section 5, given a representative sample of L, *ID1S* constructs an *EMG* G which generates L. Therefore, when all given examples consists of a representative sample of L, the teacher should answer *yes*, so the learner halts. From these observations, we have the following theorem.

Theorem 6.1 *Given any ideal teacher, then for any 1-linear SBEML L, ID1S eventually outputs an EMG G such that $L = L(G)$ and halts.*

We note that an identified description of a linear set is not always canonical.

The condition (2) on an ideal teacher is crucial. If examples are provided by a teacher satisfying only the condition (1), *ID1S* might not identify a linear set. For example, consider a linear subset $Q((0,0), \{(1,0),(0,1)\})$ of $\mathcal{N}^k$. If the teacher always gives examples from the set $\{(n,1) \mid n > 0\}$, then *ID1S* never identifies the linear set.

Next, we show the time complexity of learning. As we have remarked in Section 5, all processes of *ID1* other than the consistency check are done in polynomial time of i, k, and m, where i is a time, k is the cardinality of Σ, and m is the maximum length of the given examples. Since the learner *ID1S* never checks whether a conjecture is consistent with the examples, we have the following theorem.

Theorem 6.2 *Given any ideal teacher, then for any 1-linear SBEML, the total running time of ID1S is bounded by a polynomial in k, n, and m, where k is the cardinality of an alphabet Σ, n is the number of all examples given by the teacher, and m is the maximum length of the examples.*

7 Concluding Remarks

Intrinsically, our methods are based on semilinear subsets of $\mathcal{N}^k$. Therefore, we could apply the methods to families of languages other than *SBEML*s, which have the same properties as *SBEML*s on the Parikh mappings, and also to families of objects closely related to semilinear sets such as Presburger formulas, Petri nets, and so on.

References

[1] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.

[2] S. Ginsburg. *The Mathematical Theory of Context Free Languages.* McGraw-Hill, New York, 1966.

[3] O. H. Ibarra. Simple matrix languages. *Information and Control*, 17:359–394, 1970.

[4] R. Siromoney. On equal matrix languages. *Information and Control*, 14:135–151, 1969.