ICOT Technical Report: TR-452

TR-452

On Learning Equal Matrix Languages

by
Y. Takada (Fujitsu)

February, 1989

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

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

Abstract

We consider the learning problem for languages, called *strongly bounded equal matrix languages*, consisting 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. The languages 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. 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, the subfamily is learnable in polynomial time of the size of inputs.

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

- 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,
- 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.

In Section 2, the family of strongly bounded equal matrix languages is formally defined. In Section 3, we quote the Siromoney's result which connects strongly bounded equal matrix languages with semilinear sets. Meaningful subfamilies of the languages, called *n-linears strongly bounded equal matrix languages*, are introduced. Also, we note some properties of semilinear sets. From these, we show learnabilities from positive examples for the families based on semilinear sets, in Section 4. It is proved that the family of strongly bounded equal matrix languages is not learnable from positive examples, while there exists a meaningful subfamily, called 1-linear strongly bounded equal matrix languages, which is learnable from positive examples. In Section 5, we present a simple learning method for 1-linear strongly bounded equal matrix languages from positive examples. With our method, it seems that the learning problem for 1-linear strongly bounded equal matrix languages is computationally intractable. From this observation, in Section 6, we assume that there exists an *ideal* teacher who presents examples of an unknown language L for the question whether $L \subseteq L(G)$ for any conjecture G and eventually gives sufficient examples for learning. We present a polynomial-time learning method for 1-linear strongly bounded equal matrix languages with an ideal teacher.

Finally, in Section 7, we apply our results to the learning problem for concepts of simple pictures described in string languages. Our results suggest that each single concept of polygons, such as "square", "rectangular", described in string languages is learnable from positive examples and learning for them is accomplished in polynomial time with an ideal teacher, while mixed concepts of them are not so. This matchs with our intuition.

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. u_1u_2 denote the concatenation of strings u_1 and u_2. For sets of strings U_1 and U_2, U_1U_2 denotes the set $\{u_1u_2 \mid u_1 \in U_1 \text{ and } u_2 \in U_2\}$. 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 $\Sigma = \{a_1, \ldots, a_k\}$.

In general, a language over Σ is said to be *bounded* if and

*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.

†*1989 LA Winter Symposium*

only if there exist words $w_1, \ldots, w_k \in \Sigma^*$ such that $L \subseteq w_1^* \cdots w_k^*$.

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 $(A_1, A_2, \ldots, A_k)$, called a *nonterminal*, such that for any pair $(A_1, A_2, \ldots, A_k)$ and $(B_1, B_2, \ldots, B_k)$ of N, $\{A_1, A_2, \ldots, A_k\} \cap \{B_1, B_2, \ldots, B_k\} = \emptyset$.
3. Π is a finite nonempty set consisting of the following types of *matrix rules*;
   (a) *initial matrix rules* of the form
   $$[S \to w_1 A_1 w_2 A_2 \cdots w_k A_k]$$
   (b) *nonterminal equal matrix rules* of the form
   $$\begin{bmatrix} A_1 & \to & w_1 B_1 \\ A_2 & \to & w_2 B_2 \\ & \vdots & \\ A_k & \to & w_k B_k \end{bmatrix}$$
   (c) *terminal equal matrix rules* of the form
   $$\begin{bmatrix} A_1 & \to & w_1 \\ A_2 & \to & w_2 \\ & \vdots & \\ A_k & \to & w_k \end{bmatrix}$$
   where $w_1, w_2, \ldots, w_k \in \Sigma^*$, S is the initial symbol, and $(A_1, A_2, \ldots, A_k)$, $(B_1, B_2, \ldots, B_k)$ are nonterminals.

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

1. x is the initial symbol S and the initial matrix rule $[S \to y]$ is in Π,
2. there exist strings $u_1, \ldots, u_k$, $v_1, \ldots, v_k$ 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
   $$\begin{bmatrix} A_1 & \to & z_1 \\ & \vdots & \\ A_k & \to & z_k \end{bmatrix}$$
   is in Π.

For any $x, y \in V^*$, we write $x \overset{*}{\Longrightarrow} y$ if either $x = y$ or there exist $x_0, \ldots, x_n \in V^*$ such that $x = x_0$, $y = x_n$, and $x_i \Longrightarrow x_{i+1}$ for each i. The sequence $x_0, \ldots, x_n$ is called a *derivation* (from x_0 to x_n) and is denoted by

$$x_0 \Longrightarrow \cdots \Longrightarrow x_n.$$

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.

The family of *EML*s contains context-sensitive languages. For example, the context-sensitive language $\{a^n b^n c^n \mid n \geq 1\}$ is an *EML*. Also, there exists a context-free language which is not an *EML* (Ibarra [5]). For example, the context-free language $\{a^n b^n \mid n \geq 1\}^*$ is not an *EML*.

In this paper, we consider the learning problem for a *strongly bounded equal matrix language* (abbreviated *SBEML*). Again, the family of *SBEML*s contains context-sensitive languages and there exists a context-free language not in the family.

We shall also consider computational complexities of learnings. We use the definitions of deterministic and nondeterministic polynomial time computability and reducibility, of classes P and NP, and of NP-hardness and NP-completeness as described in [3].

3 Algebraic Characterization

In this section, we show an algebraic characterization of *SBEML*s.

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 $n \in \mathcal{N}$, $n^k = (n, \ldots, n)$ (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. Thus for elements $\mathrm{u} = (u_1, \ldots, u_k)$ and $\mathrm{v} = (v_1, \ldots, v_k)$ in $\mathcal{N}^k$ and n in $\mathcal{N}$, $\mathrm{u}+\mathrm{v} = (u_1+v_1, \ldots, u_k+v_k)$, $\mathrm{u}-\mathrm{v} = (u_1-v_1, \ldots, u_k-v_k)$, and $n\mathrm{u} = (nu_1, \ldots, nu_k)$. We may also speak of the linear dependence and the linear independence of elements of $\mathcal{N}^k$.

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

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

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

Definition A subset Q of $\mathcal{N}^k$ is said to be *linear* if and only if there exist an element c and a finite subset P of $\mathcal{N}^k$ such that $Q = Q(\mathrm{c}, P)$. Q is said to be *semilinear* if and only if Q is the union of a finite number of linear sets.

Furthermore, a subset $Q = Q(\mathrm{c}, P)$ of $\mathcal{N}^k$ is said to be *simple* if and only if the elements of P are linearly independent. Furthermore, 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 $Q(\mathrm{c}_1, P_1) \cup \cdots \cup Q(\mathrm{c}_n, P_n)$ of Q.

Definition A description $Q(\mathrm{c}, P)$ 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 $Q(\mathrm{c}_1, P_1) \cup \cdots \cup Q(\mathrm{c}_n, P_n)$ of a semilinear set is said to be *canonical* if and only if each description $Q(\mathrm{c}_i, P_i)$ of a linear set is canonical.

Note that for any linear subset Q of $\mathcal{N}^k$, a canonical description $Q(\mathrm{c}, P)$ is unique because $\mathrm{c} \in \mathcal{N}^k$ 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. In fact, for the semilinear subset $Q = Q(0^2, \{(1,0)\}) \cup Q(1^2, \{(1,0),(0,1)\})$ of $\mathcal{N}^2$, the description $Q(0^2, \{(1,0),(1,1)\}) \cup Q(1^2, \{(0,1),(1,1)\})$ of Q is also canonical.

The Parikh mapping of *EMLs* defined as follows connects *EMLs* 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.

Thus $\psi(\lambda) = 0^k$ and $\psi(w_1 \cdots w_n) = \sum_1^n \psi(w_i)$ for each $w_i \in \Sigma^*$. We call $\psi(L) = \{\psi(w) \mid w \in L\}$ *the Parikh set of an EML* L.

The following theorem is due to Siromoney [9]:

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 $Q(\mathrm{c}, \{\mathrm{p}_1, \ldots, \mathrm{p}_r\})$ be a description of the linear set Q. Also, let $\mathrm{c} = (c_1, \ldots, c_k)$ and $\mathrm{p}_i = (p_i^1, \ldots, p_i^k)$. 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:

$$[S \to a_1^{c_1} A_1 \cdots a_k^{c_k} A_k] \begin{bmatrix} A_1 & \to & \lambda \\ & \vdots & \\ A_k & \to & \lambda \end{bmatrix}$$

$$\begin{bmatrix} A_1 & \to & a_1^{p_i^1} A_1 \\ & \vdots & \\ A_k & \to & a_k^{p_i^k} A_k \end{bmatrix} \quad \text{for each } i$$

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

Corollary 3.2 *Let L_1, L_2 be SBEMLs and ψ be the Parikh mapping. Then, $L_1 \subseteq L_2$ if and only if $\psi(L_1) \subseteq \psi(L_2)$.*

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.

In the rest of this section, we note some basic properties on semilinear subsets of $\mathcal{N}^k$.

At first, we show the time complexity of the membership problem for linear sets. As we will show later, this plays an important role in the learning problem for them. The problem is effectively solvable. However, the following theorem shows that the problem is computationally intractable.

Theorem 3.3 *For any fixed positive integer k, given a canonical description $Q(\mathrm{c}, P)$ of a linear subset of $\mathcal{N}^k$ and an element q of $\mathcal{N}^k$, the problem of deciding whether $\mathrm{q} \in Q(\mathrm{c}, P)$ is NP-complete.*

The proof of this theorem is in Appendix.

Corollary 3.4 *For any fixed positive integer k, given a canonical description $Q(\mathrm{c}_1, P_1) \cup \cdots \cup Q(\mathrm{c}_n, P_n)$ of a semilinear subset of $\mathcal{N}^k$ and an element q of $\mathcal{N}^k$, the problem of deciding whether $\mathrm{q} \in Q(\mathrm{c}_1, P_1) \cup \cdots \cup Q(\mathrm{c}_n, P_n)$ is NP-complete.*

Remark Given a description $Q(\mathrm{c}, P)$ of a simple subset of $\mathcal{N}^k$ and an element q of $\mathcal{N}^k$, the problem of deciding whether $\mathrm{q} \in Q(\mathrm{c}, P)$ is solvable in polynomial time by the famous elimination method. Therefore, for semi simple sets, the problem is also solvable in polynomial time.

Finally, we summarize the closure properties on Boolean operations and the one on containments of semilinear sets. The reader may find formal proofs of them in [4], for example.

Proposition 3.5 *The family of semilinear subsets of $\mathcal{N}^k$ is closed under Boolean operations.*

Corollary 3.6 *It is effectively solvable to determine for arbitrary semilinear sets Q_1 and Q_2, whether (1) $Q_1 \subseteq Q_2$, (2) $Q_1 = Q_2$.*

The next corollary follows from Theorem 3.1, Proposition 3.5, and Corollary 3.6.

Corollary 3.7 *The family of SBEMLs is closed under Boolean operations. It is effectively solvable to determine for arbitrary SBEMLs L_1 and L_2, whether $L_1 = L_2$.*

4 Learnabilities from Positive Examples

In this section, we consider the learnabilities of the families of *SBEMLs* from positive examples. On learning of formal languages, Angluin [1] presented a necessary and sufficient condition for languages to be learnable from positive examples.

Let L be a nonempty language over an alphabet Σ and "+", "−" be symbols not in Σ. A *positive example* of L is a

pair $(+, w)$ such that $w \in L$ and a *negative example* of L is a pair $(-, v)$ such that $v \in \Sigma^* - L$. A *presentation of* L is an infinite sequence $\sigma = s_1, s_2, s_3, \ldots$, of positive and negative examples such that the set of all strings appearing in σ is Σ^*. A *positive presentation of* L is an infinite sequence $\sigma = s_1, s_2, s_3, \ldots$, of positive examples such that the set of all strings appearing in σ is L.

An *indexed family of nonempty languages* is an infinite sequence $L_1, L_2, L_3, \ldots$, where each L_i is a nonempty language. An *indexed family of nonempty recursive languages* is an indexed family of nonempty languages $L_1, L_2, L_3, \ldots$, such that there exists an effective procedure to compute the membership function

$$\begin{aligned} f(i, w) &= 1 \quad \text{if} \quad w \in L_i \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

A learner is defined to be a procedure whose input is a (positive) presentation of a language L and output is an infinite sequence of grammars.

Let σ be a (positive) presentation of L and M be a learner. We denote by $M[\sigma]$ an output sequence $G_1, G_2, G_3, \ldots$ of M for σ. Each G_i is called a *conjecture of M at the time i*. M is said to *identify L in the limit from (positive) examples* if and only if for every (positive) presentation σ of L there exists a positive integer n such that $L = L(G_n)$ and $G_n = G_{n+1} = G_{n+2} \cdots$ in $M[\sigma]$.

Let $L_1, L_2, L_3, \ldots$, be an indexed family of nonempty languages. An indexed family of nonempty languages $L_1, L_2, L_3, \ldots$, is *learnable from (positive) examples* if and only if there exists a learner which identifies L_i in the limit from (positive) examples for every $i \geq 1$.

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 $\mathcal{N}^k$. Thus we may speak of minimal elements in a subset of $\mathcal{N}^k$. The condition for two elements $(u_1, \ldots, u_k)$ and $(v_1, \ldots, v_k)$ in $\mathcal{N}^k$ to be incomparable is the existence of i and j such that $u_i < v_i$ and $u_j > v_j$.

Lemma 4.3 *Every linear subset Q of $\mathcal{N}^k$ has the unique minimum element with respect to $\preceq$.*

Proof. 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 (recall that a canonical description is unique). Since $n_i \in \mathcal{N}$ and $\mathbf{p}_i \in \mathcal{N}^k$, clearly $\mathbf{c}$ is the unique minimum element of Q with respect to $\preceq$. □

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 $Q((\mathbf{c}_1, \ldots, \mathbf{c}_k), P)$ 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 $|\mathbf{q}|_{\mathbf{c}}$. The next lemma immediately follows from definitions Q and $C(Q)$:

Lemma 4.4 *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, $|\mathbf{q}|_{\mathbf{c}} > |\mathbf{p}_i|_{\mathbf{c}}$ and $\mathbf{q} = \mathbf{c} + n_1\mathbf{p}_1 + \cdots + n_m\mathbf{p}_m$, where each $n_i \geq 1$.*

Lemma 4.5 *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 $\mathbf{q} = \mathbf{c}' + m_1\mathbf{p}'_1 + \cdots + m_r\mathbf{p}'_r$. □

Lemma 4.6 *The family of linear sets 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. We have only to consider the inclusions of linear subsets of $\mathcal{N}^k$. 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 4.5, 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. □

Corollary 4.7 *The family of simple sets 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,

$$L_{11}, L_{21}, L_{21}, \ldots, L_{ij}, \ldots,$$

where $(i, j) \in \mathcal{N}^2$ and $Q(\mathbf{c}_i, P_i) = \psi_{ij}(L_{ij})$, is an indexed family of 1-linear *SBEMLs*. Therefore, from Theorem 3.1 and Lemma 4.6, we have the following theorem.

Theorem 4.8 *The family of 1-linear SBEMLs is learnable from positive examples.*

Note that Theorem 4.8 does not depend on alphabets.

Thus, 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 4.9 *The family of semilinear sets 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 $Q_2 = Q((1,1), \{(1,0), (0,1)\})$. In fact, Q is a semilinear subset of $\mathcal{N}^2$ consisting of two linear sets (see [2], for example).

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

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

Then, a canonical description of Q^T is effectively found from the above description (cf, Figure 1). 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 4.9.

Lemma 4.10 *For each $n \geq 2$, the family of semilinear sets 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 $Q_n = Q((n-1,1), \{(1,0), (0,1)\})$. Clearly, Q is a semilinear subset of $\mathcal{N}^2$ consisting of n linear sets.

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

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

Then, a canonical description of Q^T is effectively found from the above description. From the proof of Lemma 4.9, 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 4.10.

Theorem 4.11 *For each $n \geq 2$, the family of n-linears SBEMLs is not learnable from positive examples.*

Remark The proofs show that for any alphabet which contains at least two symbols, even if it is fixed, the family is not learnable from positive examples.

Corollary 4.12 *The family of SBEMLs is not learnable from positive examples.*

5 A Simple Learning Method for 1-linear SBEMLs

In this section, we present a learning method for 1-linear *SBEMLs* based on linear sets. By this method, any 1-linear *SBEML* is identified in the limit from positive examples.

Let $\Sigma = \{a_i \mid 1 \leq i \leq k\}$ be a fixed alphabet and ψ be the Parikh mapping $\psi_{(a_1, \ldots, a_k)}$. Also, let L be an unknown 1-linear *SBEML* such that $L \subseteq a_1^* \cdots a_k^*$. 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 2, 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

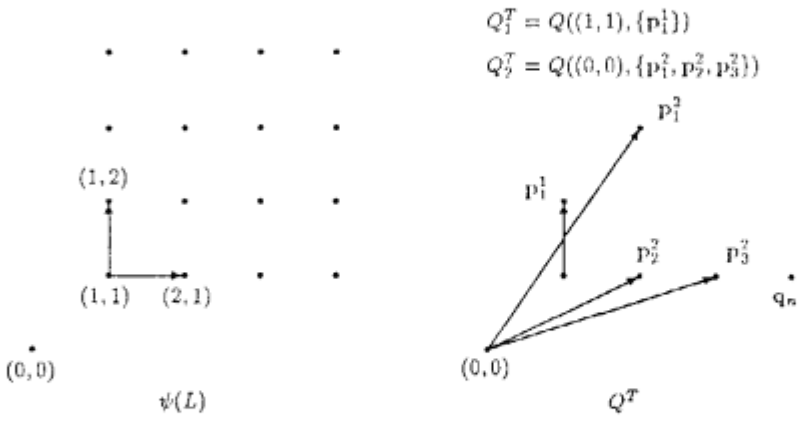

Figure 1: Construction of a description Q^T

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 *EMGs*.

$E_0 := \emptyset$;
$Q_0 := Q(0^k, \emptyset)$;

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 q of E_i with respect to $\prec$
  then let 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 $q \in E_i$ such that $q \notin Q_i$ and $|q|_c$ is minimum;
    add new period $q - c$ to Q_i;

  Construct an *EMG* G_i from Q_i and output G_i;
  go to $i+1$ step;

Figure 2: The learner *ID1*

precisely a constant c of a description of Q. Also, Lemma 4.4 and the construction of *ID1* ensure that *ID1* finds each period p_i of a canonical description of Q in order of smaller size of $|p_i|_c$. □

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.*

Note that since the alphabet and the Parikh mapping are effectively found from examples, the learner *ID1* does not depend on them. Therefore, *ID1* is precisely a learner for the family of 1-linear *SBEMLs*.

Remark A learner is said to *make an overgeneral conjecture* provided that in the process it outputs a grammar which generates a proper superset of the correct language, i.e., the language which should be identified. It is easy to verify that *ID1* never makes overgeneral conjectures. However, *ID1* does not always output conjectures which generate subsets of the correct language. If *ID1* cannot find a constant, then 0^k is assumed to be a constant. Therefore, if 0^k should not be in the Parikh set of the correct language, then the conjecture constructed by *ID1* generates a string not in the correct language.

Unfortunately, *ID1* uses membershipness of examples, which is an *NP*-complete problem as we have described, 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 $Q(\mathrm{c}, P)$ 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 $Q(\mathrm{c}, P)$ 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 $\mathrm{q} \in Q(\mathrm{c}, P)$ for an arbitrary element $\mathrm{q} \in \mathcal{N}^k$ and a canonical description $Q(\mathrm{c}, P)$. Since this latter problem is NP-complete shown in Theorem 3.3, this will imply $P = NP$, proving the theorem.

Let q be an element in $\mathcal{N}^k$ and $Q(\mathrm{c}, P)$ be a canonical description of a linear subset of $\mathcal{N}^k$. We may construct the characteristic set C of $Q(\mathrm{c}, P)$ in polynomial time. Run A on input $C \cup \{\mathrm{q}\}$ and denote the output by $Q(\mathrm{c}', P')$. Since a canonical description is unique for any linear set, if $\mathrm{c}' = \mathrm{c}$ and $P = P'$ then $\mathrm{q} \in Q(\mathrm{c}, P)$, otherwise, $\mathrm{q} \notin Q(\mathrm{c}, P)$. We may test whether $\mathrm{c} = \mathrm{c}'$ 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 All processes of *ID1* other than the consistency check are done in polynomial time of the size of inputs.

For each time $i \geq 1$, let $(+, w_1), \ldots, (+, w_i)$ be a finite subsequence of a positive presentation and $E = \{w_1, \ldots, w_i\}$. Also, let m be the maximum length of the elements in E and k be the cardinality of the alphabet Σ. Then, for each j $(1 \leq j \leq i)$, $\psi(w_j)$ is computable in at most mk^2 steps. Also, since E has at most i elements, a unique minimum element of E with respect to $\preceq$, if there exists, is found in at most i steps. On the other hand, While loop is executed at most $i - 1$ times. In each loop, since for each $\mathrm{q} \in E$, $|\mathrm{q}|_e$ is computable in polynomial time of k, a period is computable in polynomial time of i and k. Therefore, a description of a linear set Q is constructed in polynomial time of i, k, and m. Since G is constructed from the description of Q in an obvious way in polynomial time of i and k, all process other than the consistency check is done in polynomial time of i, k and m.

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 4.7. Since the membership problem of simple sets is solvable in polynomial time as we have noted, 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:

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.*

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 *EMGs*.

$E_0 := \emptyset$;
$Q_0 := Q(0^k, \emptyset)$;

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 q of E_i with respect to $\preceq$
  then let q be a constant of Q_i
  else let 0^k be a constant of Q_i;

  For each element q in E_i do
    let $\mathrm{q} - \mathrm{c}$ be a new period of Q_i;

  go to $i + 1$ step;

Figure 3: The learner *ID1S*

6 Learning 1-linear SBEMLs with an Ideal Teacher

In this section, we show that any 1-linear *SBEML* is *efficiently* learnable 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 3, 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.

We show the correctness of the learner *ID1S*. 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 a 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. Also, the learner *ID1* never makes an overgeneral conjecture and so does *ID1S*. Therefore, the ideal teacher has only to give positive examples, and 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 An Application to Simple Picture Languages

Consider the problem of describing polygons, illustrated in Figure 4, in string languages. One of the most simple answers for the problem is to describe them in sequences of symbols which represent unit lines, as illustrated in Figure 5. Then, these strings have the same form $u_1^{n_1} \cdots u_m^{n_m}$, where each symbol u_i denotes a unit line. For example, a set of squares is described as the language $L_S = \{u_1^n u_2^n u_3^n u_4^n \mid n \geq 1\}$.

On pictures described in words over the symbols, which denote unit lines from the Cartesian plane considered as a square grid, Maurer et al. studied various properties in [7]. In this section, we shall touch the learning problem of such descriptions.

We have shown that the family of *SBEMLs* is not learnable from positive examples, while the family of 1-linear *SBEMLs* is learnable from positive examples. Also, the family of 1-linear *SBEMLs* is efficiently learnable with an ideal teacher. These results suggest that

- each concept of polygons described in string languages is learnable from positive examples, while mixed concepts of them are not so,
- each concept of polygons is efficiently learnable with an ideal teacher.

For example, consider the concept "square" is the language $L_S = \{u_1^n u_2^n u_3^n u_4^n \mid n \geq 1\}$. The Parikh set of L_S is a linear set $\psi_{(u_1,u_2,u_3,u_4)}(L_S) = \{(1,1,1,1) + n(1,1,1,1) \mid n \in N\}$. Therefore, L_S is a 1-linear *SBEML* and efficiently learnable with an ideal teacher. On the other hand, "rectangular in which vertical lines are two or three times longer than horizontal lines" is the language $L_{2,3} = \{u_1^n u_2^{2n} u_3^n u_4^{2n} \mid n \geq 1\} \cup \{u_1^n u_2^{3n} u_3^n u_4^{3n} \mid n \geq 1\}$. The Parikh set of $L_{2,3}$ is a semilinear set $\psi_{(u_1,u_2,u_3,u_4)}(L_{2,3}) = \{(1,2,1,2) + n(1,2,1,2) \mid n \in N\} \cup \{(1,3,1,3) + n(1,3,1,3) \mid n \in N\}$, so it is not learnable from positive examples. This matchs with our intuition.

8 Concluding Remarks

We have shown that the family of *SBEMLs* is not learnable from positive examples, while the family of 1-linear *SBEMLs* is learnable from positive examples. Also, we have presented an efficient learning method for 1-linear *SBEMLs* with an ideal teacher.

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

Acknowledgements

The author is grateful to his colleagues, Takashi Yokomori and Yasubumi Sakakibara who worked through an earlier draft of the paper and many suggestions. The author would also like to thank Toshiro Minami, Kazuhiro Yokoyama, and Kunihiko Hiraishi for their helpful comments and suggestions.

References

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

[2] S. Eilenberg and M. P. Schützenberger. Rational sets in commutative monoids. *Journal of Algebra*, 13:173–191, 1969.

[3] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, New York, 1979.

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

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

[6] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcker, editors,

Figure 4: Polygons

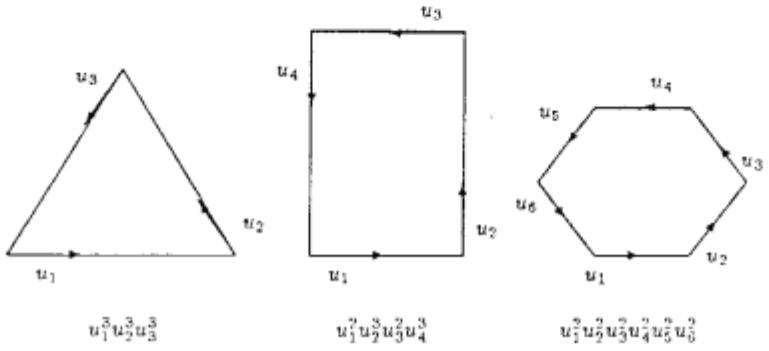

Figure 5: Polygons described in string languages

Complexity of Computer Computations, pages pp.85–103, Plenum Press, New York, 1972.

[7] H. A. Maurer, G. Rozenberg, and E. Welzl. Using string languages to describe picture languages. *Information and Control*, 54:155–185, 1982.

[8] A. Salomaa. *Formal Languages*. Academic Press, Inc., New York, 1973.

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

Appendix: Proof of Theorem 3.3

We denote the membership problem for linear sets in the following way:

> LINEAR SET MEMBERSHIP (LM)
> INSTANCE: A canonical description $Q(\mathbf{c}, P)$ of a linear subset of $\mathcal{N}^k$ and an element $\mathbf{q}$ of $\mathcal{N}^k$.
> QUESTION: Is $\mathbf{q}$ an element of $Q(\mathbf{c}, P)$?

Note that k is a *fixed* positive integer.

Consider the following procedure: Given $Q(\mathbf{c}, P)$ and $\mathbf{q}$,

step 0 let $\mathbf{q}_0 = \mathbf{q} - \mathbf{c}$,

step i choose a period $\mathbf{p}$ of P, nondeterministically,
and let $\mathbf{q}_i = \mathbf{q}_{i-1} - \mathbf{p}$,
if the value of any coordinate of $\mathbf{q}_i$ is 0, then output *TRUE* and halt,
else go to *step i+1*.

Clearly, there exists a nondeterministic Turing machine which executes the procedure in polynomial time of the size of inputs, and it outputs *TRUE* if and only if $\mathbf{q} \in Q(\mathbf{c}, P)$.

To see that the problem LM is NP-hard, consider the following problem:

> EXACT COVER (XC)
> INSTANCE: Set X and a collection C of subsets of X.
> QUESTION: Does C contain an exact cover for X, i.e., a subcollection $C' \subseteq C$ such that every element of X occurs in exactly one member of C'?

This problem is known as an NP-complete problem (see [6]). We exhibit a polynomial time reduction to LM of XC.

Let $X = \{x_1, \ldots, x_n\}$ be a set and $C = \{c_1, \ldots, c_m\}$ be a collection of subsets of X. Without loss of generality, we assume that X and C are ordered sets. Given X and C, we constructs a canonical description $Q(0, P)$ of a linear subset of $\mathcal{N}$ and show that C contains an exact cover for X if and

only if $\mathrm{q} \in Q(0, P)$, where

$$\mathrm{q} = \sum_{j=1}^{m} 2^{n+m(n-1)+j} + \sum_{i=1}^{n} 2^{i+m(i-1)}.$$

For each $x_i \in X$ and each $c_j \in C$, define

$$g(x_i, c_j) = \begin{cases} 1 & \text{if } x_i \in c_j \\ 0 & \text{if } x_i \notin c_j \end{cases}$$

For each $c_j \in C$, define

$$\mathrm{p}_j = 2^{n+m(n-1)+j} + \sum_{i=1}^{n} g(x_i, c_j) 2^{i+m(i-1)}.$$

Also, we define

$$\mathrm{p}_0 = 2^{n+m(n-1)}$$

Then, define

$$P = \{\mathrm{p}_j \mid 1 \leq j \leq m\} \cup \{\mathrm{p}_0\}$$

Clearly, the indicated construction of P from X and C is carried out in polynomial time of the numbers of elements of X and C.

We first show that a description $Q(0, P)$ is canonical. Assume that for some j $(1 \leq j \leq m)$, $\mathrm{p}_j = n_1\mathrm{p}_1 + \cdots + n_{j-1}\mathrm{p}_{j-1} + n_{j+1}\mathrm{p}_{j+1} + \cdots + n_m\mathrm{p}_m$. Let x_i be an element of c_j such that i is a minimum index in c_j. If there exist $i'_1, \ldots, i'_t$ such that $2^{i+m(i-1)} = n'_1 2^{i'_1+m(i'_1-1)} + \cdots + n'_t 2^{i'_t+m(i'_t-1)}$, every $n'_1, \ldots, n'_t$ are greater than 0, and every $i'_1, \ldots, i'_t$ are less than i, then $n'_1 + \cdots + n'_t \geq 2^{m+1}$, so $2^{n+m(n-1)+j} < (n'_1 + \cdots + n'_t)2^{n+m(n-1)}$, contradiction. If there exists another $c_{j'}$ such that i is a minimum index in $c_{j'}$, then $2^{n+m(n-1)+j} \neq 2^{n+m(n-1)+j'}$ and if $2^{n+m(n-1)+j} = n'2^{n+m(n-1)+j'}$ then $n' < m$, so there is no $n'' \in \mathcal{N}$ such that $\mathrm{p}_j = n''\mathrm{p}_{j'}$.

Suppose that C' is an exact cover for X. Then we define the coefficients of periods as follows: For each $j = 1, \ldots, m$, if $c_j \in C'$ then the coefficient n_j of p_j is 1, while if $c_j \notin C'$ then $n_j = 0$. Also, we define the coefficient n_0 of the period p_0 by

$$n_0 = \sum_{c_{j'} \in C - C'} 2^{n+m(n-1)+j'}$$

The construction of $Q(0, P)$ and q ensures that $\mathrm{q} \in Q(0, P)$.

Conversely, suppose that $\mathrm{q} \in Q(0, P)$. As we have shown, for any i $(1 \leq i \leq n)$, $2^{i+m(i-1)} \neq \sum_{l=1}^{i-1} n_l 2^{l+m(l-1)}$ for any $n_1, \ldots, n_l$ less than 2^{m+1}. Therefore, for any i, there exists exactly one period p_j such that p_j is constructed from c_j which contains x_i and the coefficient of p_j is 1. Let C' be a set which contains c_j such that the coefficient of p_j is 1. It is easy to verify that C' is an exact cover for X.

Thus, even if $k = 1$, the problem LM is NP-hard. This completes the proof. □