PREFIX-FREE GENERATING SETS OF FORMAL LANGUAGES AND LEARNING*

MIKIHARU TERADA

Received February 22, 2001

ABSTRACT. This paper deals with a particular type of generating set, called *prefix-free*, for a language. Given a language L over an alphabet Σ, a set G is a *generating set* of L, denoted by $L \sqsubseteq G$, if $L \subseteq G^+$. It is well known that a prefix-free set G has the property of *unique decipherability* for all strings in G^+.

We first show that for a language L, the class $\mathcal{PFG}_L$ of all prefix-free and reduced generating sets for L is a complete lattice under the relation $\sqsubseteq$ and give explicitly the least element $G_L^{\inf}$ and the greatest element $G_L^{\sup}$. Especially we are concerned with the least element $G_L^{\inf}$ of the lattice. $G_L^{\inf}$ has a *good* property in the sense that every string in L can be represented by the minimum number of strings in $G_L^{\inf}$ among $\mathcal{PFG}_L$. We give a necessary and sufficient condition for $G_L^{\inf}$ to be finite. Moreover, we present a polynomial time algorithm for computing $G_L^{\inf}$ with respect to the sum of lengths of strings in L for a finite language L. For an infinite language, we consider the problem of identifying $G_L^{\inf}$ in the framework of *identification in the limit* proposed by Gold for language learning, and give a polynomial time learning algorithm for computing $G_L^{\inf}$, provided that the target $G_L^{\inf}$ is finite.

1 Introduction. In this paper, we consider particular sets of strings that generate every string in a formal language L over an alphabet Σ. A set G of strings is a *generating set* of a language L, denoted by $L \sqsubseteq G$, if $L \subseteq G^+$, that is, every string of L can be represented as a concatenation of strings in G. For instance, let $\Sigma = \{a, b\}$ and $L = \{(aab)^i b(aa)^j \mid i, j \in N\}$ be a regular language. Then the sets $G_1 = \{aab, b, aa\}$ and $G_2 = \{aa, b\}$ of strings are generating sets of L. Clearly the alphabet Σ is always a generating set of any language L consisting of nonempty strings, i.e., $L \sqsubseteq \Sigma$, and moreover the language L itself is a generating set of L.

A set G of strings is said to be *prefix-free*, if for any string of G, there is no proper prefix in G of the string. In the above example, the string aa in G_1 is a proper prefix of the string aab in G_1, and thus G_1 is not prefix-free. On the other hand, G_2 is prefix-free. For a string u, a sequence $u_1, u_2, \cdots, u_n$ of strings in G is a *factorization* of u in G and n is the *length* of u with respect to G, if $u = u_1 u_2 \cdots u_n$. A prefix-free set G has the property of so-called *unique decipherability* in terminology of coding theory in the sense that any string in G^+ has a unique factorization in G. In coding theory, such a set is called *code* and has been discussed in connection with unique decipherability. Refer in detail to e.g. [3].

In general, there are many prefix-free generating sets of a given language L. In this paper, we investigate the lattice structure of the class $\mathcal{PFG}_L$ of such prefix-free generating sets under the binary relation $\sqsubseteq$ mentioned above. We first show that the class $\mathcal{PFG}_L$ is a complete lattice and give explicitly both the least element $G_L^{\inf}$ and the greatest element $G_L^{\sup}$ of the class. In particular, we are interested in the least prefix-free generating set $G_L^{\inf}$.

2000 *Mathematics Subject Classification.* 03G10, 06B23.

Key words and phrases. Prefix-free, code, generating set, formal language, learning.

*Supported in part by Grant-in-Aid for Scientific Research on Priority Areas No. 10143104 from the Ministry of Education, Science and Culture, Japan.

The set $G_L^{\inf}$ has a *good* property in a sense that every string in L has the shortest length w.r.t. $G_L^{\inf}$ among $\mathcal{PFG}_L$. For instance, $w = aabbaa$ has the length 4 w.r.t. G_2 because of $w = (aa)(b)(b)(aa)$, while the (usual) length 6 w.r.t. Σ.

Secondly, we give a characterization theorem of $G_L^{\inf}$ to be finite, and a subclass of regular languages each of whose least prefix-free generating set is finite.

In terms of the result, we finally present a polynomial time algorithm for computing the least prefix-free generating set of a finite language. Furthermore, we consider inductive inference of the least prefix-free generating set of an infinite language. Inductive learning is the process of hypothesizing a general rule from examples. Gold[4] proposed a mathematical model of inductive inference of recursive functions or formal languages based on the criterion of success of inference in the limit. We give an efficient algorithm for identifying the least prefix-free generating set $G_L^{\inf}$ *in the limit*.

Yokomori[5] presented an efficient algorithm of *strict deterministic automaton* defined over a particular finite set of strings, called *strict prefix*, from positive examples in the limit. The strict prefix set of strings is a special kind of prefix-free generating sets introduced here. The efficient algorithm given by Yokomori contained an efficient procedure to compute the least strict prefix set in the limit. Watanabe [6] has shown that the set of strict prefix sets has a finite lattice structure.

2 Prefix-Free Generating Sets of Formal Languages.

2.1 Preliminaries. We start with some basic definitions and notations used in this paper.

Let Σ be a finite *alphabet*. Let Σ^* be the set of all finite strings over Σ, and Σ^+ be the set of all finite *nonempty* strings over Σ. The *empty* string is denoted by λ. For a string $w \in \Sigma^*$, $|w|$ denotes the *length* of w. In particular, the length of λ is 0.

The concatenation of strings u and v is denoted by uv. For a string w, a string $u \in \Sigma^*$ is a *prefix* of w, if there is a string $v \in \Sigma^*$ such that $w = uv$, particularly when $v \in \Sigma^+$, u is a *proper* prefix of w. Let N be the set of nonnegative integers.

For subsets S_1 and S_2 of Σ^*, let us define the product $S_1S_2 = \{uv \mid u \in S_1, v \in S_2\}$. Define $S^{k+1} = SS^k$ for a set $S \subseteq \Sigma^*$ and $k \in N$, where $S^0 = \{\lambda\}$. Let $S^* = \bigcup_{k \in N} S^k$, and $S^+ = S^* - \{\lambda\}$.

For sets $S_1, S_2 \subseteq \Sigma^+$, we define the following relation:

$$S_1 \sqsubseteq S_2 \text{ if and only if } S_1 \subseteq S_2^+.$$

Clearly $S \sqsubseteq \Sigma$ for any set $S \subseteq \Sigma^+$. As easily seen, the relation $\sqsubseteq$ is reflexive and transitive but not antisymmetric. Indeed, for $S_1 = \{a, b\}$ and $S_2 = \{a, b, ab\}$, we have $S_1 \sqsubseteq S_2$ and $S_2 \sqsubseteq S_1$ but $S_1 \neq S_2$. As shown below, the relation $\sqsubseteq$ is antisymmetric for prefix-free sets.

2.2 Prefix-Free Sets.

Definition 2.1 *A set $S \subseteq \Sigma^*$ is* prefix-free, *if any string in S is not a proper prefix of another string in S. By $\mathcal{PF}$ we denote the set of all prefix-free sets.*

As well known in coding theory, a prefix-free set S has the property of *unique decipherability*. That is, for every string $w \in S^+$, if there are strings $u_1, \cdots, u_n, v_1, \cdots, v_m \in S$ such that $w = u_1 \cdots u_n$ and $w = v_1 \cdots v_m$, then $m = n$ and $u_i = v_i$ for all $i = 1, \cdots, n$. We refer $u_1, \cdots, u_n$ as the unique factorization of w. Using this property, it is easily shown that the set $(\mathcal{PF}, \sqsubseteq)$ is a partially ordered set.

For a set $S \subseteq \Sigma^+$, we define

$$\mathrm{Pre}(S) = \{u \in S \mid \text{ there is no proper prefix of } u \text{ in } S\}.$$

By the above definition, clearly $\mathrm{Pre}(S) \subseteq S$, $\mathrm{Pre}(S) \in \mathcal{PF}$ and for every string $u \in S$, there is a prefix $v \in \mathrm{Pre}(S)$ of u.

Definition 2.2 *We define a binary operation O over strings as follows. For all strings $u, v \in \Sigma^+$, let*

$$O(u,v) = \begin{cases} \{u\}, & \text{if } u = v, \\ \{u, v'\}, & \text{if } \exists v' \in \Sigma^+ \text{ s.t. } v = uv', \\ \{u', v\}, & \text{if } \exists u' \in \Sigma^+ \text{ s.t. } u = vu', \\ \{u, v\}, & \text{otherwise}, \end{cases}$$

For a string w, a set $\{u, v\}$ of two strings is a direct ancestor of w, if $w \in O(u,v)$ and $w \neq u, v$. Using the binary operation O, we next define a set operation $\widetilde{O}$ over sets $S \subseteq \Sigma^+$ by

$$\widetilde{O}(S) = \bigcup_{(u,v) \in S \times S} O(u,v).$$

By the above definition, clearly $S \subseteq \widetilde{O}(S)$ and $S \sqsubseteq \widetilde{O}(S)$. For all $n \in N$, we define $\widetilde{O}^{n+1}(S) = \widetilde{O}(\widetilde{O}^n(S))$ with $\widetilde{O}^0(S) = S$, as well as the closure $\widetilde{O}^*$ by

$$\widetilde{O}^*(S) = \bigcup_{n \in N} \widetilde{O}^n(S).$$

Clearly $S \subseteq \widetilde{O}^n(S) \subseteq \widetilde{O}^{n+1}(S) \subseteq \widetilde{O}^*(S)$ and $S \sqsubseteq \widetilde{O}^n(S) \sqsubseteq \widetilde{O}^{n+1}(S) \sqsubseteq \widetilde{O}^*(S)$ for all $n \in N$. Furthermore, $\widetilde{O}(\widetilde{O}^*(S)) = \widetilde{O}^*(S)$ and if $\widetilde{O}^n(S) = \widetilde{O}^{n+1}(S)$ for some $n \in N$, then $\widetilde{O}^*(S) = \widetilde{O}^m(S)$ for all $m \geq n$.

Let $S \subseteq \Sigma^*$ and $w \in \widetilde{O}^*(S) - S$. A set $\{u, v\} \subseteq \widetilde{O}^*(S)$ is a *direct ancestor* of w, if $w \in O(u,v)$ and $w \neq u, v$.

Note that for every $n \geq 1$ and for every $w \in \widetilde{O}^n(S) - S$, there is a direct ancestor $\{u, v\} \subseteq \widetilde{O}^{n-1}(S)$ of w.

Lemma 2.1 *Let S and S' be subsets of Σ^+ such that $S \sqsubseteq S'$. If S' is prefix-free, then $\widetilde{O}^*(S) \sqsubseteq S'$ and $\mathrm{Pre}(\widetilde{O}^*(S)) \sqsubseteq S'$.*

Proof. We first prove $\widetilde{O}^*(S) \sqsubseteq S'$. By the assumption, $\widetilde{O}^0(S) = S \sqsubseteq S'$ and thus it suffices to show that $\widetilde{O}^n(S) \sqsubseteq S'$ implies $\widetilde{O}^{n+1}(S) \sqsubseteq S'$ for all $n \in N$.

Assume that $\widetilde{O}^n(S) \sqsubseteq S'$. This is obvious if $\widetilde{O}^n(S) = \widetilde{O}^{n+1}(S')$. Let $\widetilde{O}^n(S) \subsetneq \widetilde{O}^{n+1}(S)$ and w be any string of $\widetilde{O}^{n+1}(S) - \widetilde{O}^n(S)$. Then there is a direct ancestor $\{u, v\} \subseteq \widetilde{O}^n(S)$ of w such that $u = vw$. By our assumption of $\widetilde{O}^n(S) \sqsubseteq S'$, $u, v \in S'^+$. Since S' is prefix-free, it follows that $w \in S'^+$. Hence $\widetilde{O}^{n+1}(S) \sqsubseteq S'$. Consequently $\widetilde{O}^*(S) \sqsubseteq S'$.

Since $\mathrm{Pre}(\widetilde{O}^*(S))$ is a subset of $\widetilde{O}^*(S)$, the above implies $\mathrm{Pre}(\widetilde{O}^*(S)) \sqsubseteq S$. ■

Lemma 2.2 *For every set $S \subseteq \Sigma^+$, $\widetilde{O}^*(S) \sqsubseteq \mathrm{Pre}(\widetilde{O}^*(S))$.*

Proof. Let $T = \widetilde{O}^*(S)$. Suppose the converse, i.e., $T \not\sqsubseteq \mathrm{Pre}(T)$, and let $w \in T$ be one of the shortest strings in T but not contained in $(\mathrm{Pre}(T))^+$. By $w \in T$, there is a prefix $u \in \mathrm{Pre}(T)$ such that $w = uv$ for some string v. If $v = \lambda$, then $w \in \mathrm{Pre}(T)$ and a contradiction. Thus $v \neq \lambda$, and so $v \in T$ because of $v \in O(w,u)$. Since $|v| < |w|$, we have $v \in (\mathrm{Pre}(T))^+$, and thus $w \in (\mathrm{Pre}(T))^+$. This contradicts the choice of w. ■

Hereafter, we investigate the lattice structure of $\mathcal{PF}$. For a subset $\mathcal{P}$ of $\mathcal{PF}$, $\sup(\mathcal{P})$ and $\inf(\mathcal{P})$ denote the least upper bound and the greatest lower bound of $\mathcal{P}$ in $\mathcal{PF}$ under the partially ordered relation $\sqsubseteq$, respectively.

Lemma 2.3 *For every* $\mathcal{P} \subseteq \mathcal{PF}$, $\sup(\mathcal{P}) = \mathrm{Pre}(\widetilde{O}^*(\bigcup_{S\in\mathcal{P}} S))$.

Proof. Put $T = \widetilde{O}^*(\bigcup_{S\in\mathcal{P}} S)$. By Lemma 2.2, $T \sqsubseteq \mathrm{Pre}(T)$, and thus $S \sqsubseteq \mathrm{Pre}(T)$ for every $S \in \mathcal{P}$. Hence $\mathrm{Pre}(T)$ is an upper bound of $\mathcal{P}$.

Let $S' \in \mathcal{PF}$ be any upper bound of $\mathcal{P}$. Then $\bigcup_{S\in\mathcal{P}} S \sqsubseteq S'$. Since S' is prefix-free, Lemma 2.1 implies $\mathrm{Pre}(T) \sqsubseteq S'$. Hence $\mathrm{Pre}(T)$ is the least upper bound of $\mathcal{P}$ in $\mathcal{PF}$. ■

As mentioned before, Σ is a prefix-free set, and $S \sqsubseteq \Sigma$ for any $S \in \mathcal{PF}$. Hence Σ is the greatest element of $\mathcal{PF}$, i.e., $\sup(\mathcal{PF}) = \Sigma$.

Lemma 2.4 *For any* $\mathcal{P} \subseteq \mathcal{PF}$, $\inf(\mathcal{P}) = \mathrm{Pre}(\widetilde{O}^*(\bigcap_{S\in\mathcal{P}} S^+))$.

Proof. Put $T = \widetilde{O}^*(\bigcap_{S\in\mathcal{P}} S^+)$. By Lemma 2.2, $T \sqsubseteq \mathrm{Pre}(T)$.

We first prove that $\mathrm{Pre}(T)$ is a lower bound of $\mathcal{P}$, i.e., $\mathrm{Pre}(T) \sqsubseteq S$ for any $S \in \mathcal{P}$. Assume that $\bigcap_{S\in\mathcal{P}} S^+ \subsetneq \widetilde{O}(\bigcap_{S\in\mathcal{P}} S^+)$ and let $w \in \widetilde{O}(\bigcap_{S\in\mathcal{P}} S^+) - \bigcap_{S\in\mathcal{P}} S^+$. Then there is a direct ancestor $\{u, v\} \subseteq \bigcap_{S\in\mathcal{P}} S^+$ such that $u = vw$. Since each $S \in \mathcal{P}$ is prefix-free, we have $w \in \bigcap_{S\in\mathcal{P}} S^+$, and a contradiction. Hence we have $\widetilde{O}(\bigcap_{S\in\mathcal{P}} S^+) = \bigcap_{S\in\mathcal{P}} S^+$, and thus $T = \widetilde{O}^*(\bigcap_{S\in\mathcal{P}} S^+) = \bigcap_{S\in\mathcal{P}} S^+$. It means that $\mathrm{Pre}(T) \subseteq \bigcap_{S\in\mathcal{P}} S^+$. Consequently $\mathrm{Pre}(T) \sqsubseteq S$ for any $S \in \mathcal{P}$, i.e., $\mathrm{Pre}(T)$ is a lower bound of $\mathcal{P}$.

Next, we prove that $\mathrm{Pre}(T)$ is the greatest lower bound of $\mathcal{P}$. Let $S' \in \mathcal{PF}$ be any lower bound of $\mathcal{P}$. Then $S' \subseteq S^+$ for any $S \in \mathcal{P}$, and thus $S' \subseteq \bigcap_{S\in\mathcal{P}} S^+$. Since $\bigcap_{S\in\mathcal{P}} S^+ \sqsubseteq \mathrm{Pre}(T)$, we get $S' \sqsubseteq \mathrm{Pre}(T)$. Therefore $\mathrm{Pre}(T)$ is the greatest lower bound of $\mathcal{P}$. ■

By Lemma 2.3 and Lemma 2.4, the next result on $\mathcal{PF}$ is immediately given as follows:

Theorem 2.5 *The set* $(\mathcal{PF}, \sqsubseteq)$ *is a complete lattice and* Σ *is the greatest element of* $\mathcal{PF}$.

2.3 Prefix-Free Generating Sets. A language over Σ is a subset of Σ^+. In this subsection, we consider a particular set of strings generating all strings in a given language.

Definition 2.3 *Let* $G \subseteq \Sigma^+$ *and* L *be a language.* G *is a generating set of* L *if* $L \sqsubseteq G$. *A generating set* G *of* L *is reduced (with respect to* L*) if* $L \not\sqsubseteq G'$ *for any proper subset* G' *of* G. *By* $\mathcal{PFG}_L$ *we denote the set of all reduced and prefix-free generating sets of* L.

As mentioned before, if a generating set G of a language L is prefix-free, each string of L has a unique factorization of strings in G. Thus for any prefix-free generating set G of L, we have a unique *reduced* generating set $G_0 \subseteq G$ by deleting strings of G not used in factorizations of strings of L. Thus we get:

Lemma 2.6 *Let* L *be a language and* G *be a prefix-free generating set of* L. *Then there uniquely exists a reduced and prefix-free generating set* $G_0 \in \mathcal{PFG}_L$ *of* L *such that* $G_0 \subseteq G$.

For any language L, the set $\mathcal{PFG}_L$ is a subset of $\mathcal{PF}$ and a partially ordered set under $\sqsubseteq$.

In what follows, we investigate a lattice structure of $\mathcal{PFG}_L$ and a characterization of the least element of $\mathcal{PFG}_L$ to be finite.

Lemma 2.7 *Let* $L \subseteq \Sigma^+$ *be a language and* $w \in \mathrm{Pre}(\widetilde{O}^*(L))$. *Then there exists a finite set* $S_w \subseteq L$ *and a string* $u_w \in S_w$ *such that (i)* $w \in \mathrm{Pre}(\widetilde{O}^*(S_w))$ *and (ii)* $u_w = vw$ *for some string* $v \in (\mathrm{Pre}(\widetilde{O}^*(S_w)))^*$.

Proof. We prove by induction on $n \in N$ that for every $w \in \widetilde{O}^n(L)$, there exists a finite set $S_w \subseteq L$ and a string $u_w \in S_w$ satisfying the conditions (i)' $w \in \widetilde{O}^*(S_w)$ and (ii) given in our lemma.

For case of $n = 0$. Let $w \in L$. Then the finite set $S_w = \{w\}$ and the string $u_w = w$ hold the conditions (i)' and (ii).

Suppose that it is valid for $n \leq k$ $(k \geq 1)$. For $n = k+1$, let $w \in \widetilde{O}^{k+1}(L)$. Then there is a direct ancestor $\{w_1, w_2\} \subseteq \widetilde{O}^k(L)$ of w such that $w_1 = w_2 w$. By our induction hypothesis, for $i = 1, 2$ there are a *finite* subsets $S_{w_i} \subseteq L$ and and a string $u_{w_i} \in S_{w_i}$ such that $w_i \in \widetilde{O}^*(S_{w_i})$ and $u_{w_i} = v_i w_i$ for some $v_i \in (\mathrm{Pre}(\widetilde{O}^*(S_{w_i})))^*$. Put $u_w = u_{w_1}$ and $S_w = S_{w_1} \cup S_{w_2}$. Then $u_w \in S_w$ and

$$\mathrm{Pre}(\widetilde{O}^*(S_{w_i})) \subseteq \widetilde{O}^*(S_{w_i}) \subseteq \widetilde{O}^*(S_w) \sqsubseteq \mathrm{Pre}(\widetilde{O}^*(S_w))$$

for $i = 1, 2$. Thus $(\mathrm{Pre}(\widetilde{O}^*(S_{w_i})))^* \sqsubseteq \mathrm{Pre}(\widetilde{O}^*(S_w))$. It means $v_1, w_2 \in (\mathrm{Pre}(\widetilde{O}^*(S_w)))^*$, and so $v_1 w_2 \in (\mathrm{Pre}(\widetilde{O}^*(S_w)))^+$. By $u_w \in S_w$ and $u_w = (v_1 w_2)w$, these imply (i)'$w \in \widetilde{O}^*(S_w)$. The proof by induction completes.

If $w \in \mathrm{Pre}(\widetilde{O}^*(L))$, $w \in \widetilde{O}^n(L)$ for some $n \in N$. By the above, $w \in \widetilde{O}^*(S_w)$ for some finite set $S_w \subseteq L$. Clearly $w \in \widetilde{O}^*(S_w)$ implies $w \in \mathrm{Pre}(\widetilde{O}^*(S_w))$ because of $\widetilde{O}^*(S_w) \subseteq \widetilde{O}^*(L)$. ■

Theorem 2.8 *$(\mathcal{PFG}_L, \sqsubseteq)$ is a complete lattice of $\mathcal{PF}$ for every language L.*

Proof. It suffices to show that for any nonempty subset $\mathcal{G} \subseteq \mathcal{PFG}_L$, there exist the least upper bound and the greatest lower bound of $\mathcal{G}$ in our set $\mathcal{PFG}_L$.

Let $\mathcal{G}$ be any nonempty subset of $\mathcal{PFG}_L$. Put $T = \widetilde{O}^*(\bigcup_{G \in \mathcal{G}} G)$. By Lemma 2.3, $\sup(\mathcal{G}) = \mathrm{Pre}(T)$ in the lattice $\mathcal{PF}$. It is enough to show that $\mathrm{Pre}(T) \in \mathcal{PFG}_L$. By $L \sqsubseteq G \sqsubseteq \sup(\mathcal{G})(= \mathrm{Pre}(T))$ for every $G \in \mathcal{G}$, clearly $\mathrm{Pre}(T)$ is a prefix-free generating set of L. Since $\mathrm{Pre}(T)$ has the property of unique decipherability, this set is reduced w.r.t. L if every string in $\mathrm{Pre}(T)$ occurs in unique factorization of at least one string in L. Let w be any string in $\mathrm{Pre}(T)$. Then Lemma 2.7 implies that there is a string $u_w \in \bigcup_{G \in \mathcal{G}} G$ such that $u_w = vw$ for some $v \in (\mathrm{Pre}(T))^*$. Let $G_w \in \mathcal{G}$ be a set such that $u_w \in G_w$. Then since the set G_w is reduced w.r.t. L, u_w occurs in the factorization of some string in L, i.e., there is a string $x_w \in L$ such that $x_w = y u_w z$ for some $y, z \in G_w^*$. Hence we have $x_w = yvwz$ and $y, v, z \in (\mathrm{Pre}(T))^*$. This means that $\mathrm{Pre}(T)$ is reduced w.r.t. L.

Next, we show the existence of the greatest lower bound of $\mathcal{G}$ in $\mathcal{PFG}_L$. By Lemma 2.4, $\inf(\mathcal{G}) = \mathrm{Pre}(T_0)$ in $\mathcal{PF}$, where $T_0 = \widetilde{O}^*(\bigcap_{G \in \mathcal{G}} G^+)$. Similarly to the above, it can be shown that $\mathrm{Pre}(T_0)$ is a generating set of L. Appealing to Lemma 2.6, there uniquely exists a reduced and prefix-free generating set $G_0 \in \mathcal{PFG}_L$ of L such that $G_0 \subseteq \mathrm{Pre}(T_0)$. By $(\mathrm{Pre}(T_0) =) \inf(\mathcal{G}) \sqsubseteq G$ for every $G \in \mathcal{G}$, clearly the subset G_0 of $\mathrm{Pre}(T_0)$ is a lower bound of $\mathcal{G}$.

Now we show that G_0 is the *greatest* lower bound of $\mathcal{G}$ in $\mathcal{PFG}_L$. Suppose the converse, that is, there is a lower bound $G \in \mathcal{PFG}_L$ of $\mathcal{G}$ such that $G \not\sqsubseteq G_0$. Let $w \in G - G_0^+$. Similarly to the above, by $G \in \mathcal{PFG}_L$, there is a string $x_w \in L$ such that $x_w = uwv$ for some $u, v \in G^*$. $G \sqsubseteq \mathrm{Pre}(T_0)$ implies $u, v \in (\mathrm{Pre}(T_0))^*$ and so $w \in (\mathrm{Pre}(T_0))^+$. By the choice of G_0, the string x_w has the same factorization in $\mathrm{Pre}(T_0)$ and G_0. Thus $w \in G_0^+$ must hold. This contradicts $w \notin G_0^+$.

Therefore G_0 is the *greatest* lower bound of $\mathcal{G}$ in $\mathcal{PFG}_L$. ■

In general, for a subset $\mathcal{G} \subseteq \mathcal{PFG}_L$, $\inf(\mathcal{G})$ is not always the greatest lower bound of $\mathcal{G}$ in $\mathcal{PFG}_L$. In fact, let us consider a language $L = \{w\}^+$, where $w = abcdacdaab$.

Let $G_1 = \{abcd, aab, acd\}$ and $G_2 = \{ab, cda\}$. As easily seen, $G_1, G_2 \in \mathcal{PFG}_L$, and $\inf\{G_1, G_2\} = \{abcdaab, w\}$. Clearly $\inf\{G_1, G_2\}$ is not reduced although $L \sqsubseteq \inf\{G_1, G_2\}$. The greatest lower bound of $\{G_1, G_2\}$ is given by $\{w\} \subseteq \inf\{G_1, G_2\}$. Note that the set $\{w\}$ is the greatest lower bound of the whole set $\mathcal{PFG}_L$.

By the above theorem, the lattice $\mathcal{PFG}_L$ has the greatest lower bound of $\mathcal{PFG}_L$, that is, the least element. The least element is given as follows:

Theorem 2.9 *For any language L, $\mathrm{Pre}(\widetilde{O}^*(L))$ is the least element of $\mathcal{PFG}_L$ under $\sqsubseteq$.*

Proof. By Lemma 2.2, $L \subseteq \widetilde{O}^*(L) \sqsubseteq \mathrm{Pre}(\widetilde{O}^*(L))$. Thus $\mathrm{Pre}(\widetilde{O}^*(L))$ is a prefix-free generating set of L. Furthermore, Lemma 2.1 implies $\mathrm{Pre}(\widetilde{O}^*(L)) \sqsubseteq G$ for every $G \in \mathcal{PFG}_L$. This means that $\mathrm{Pre}(\widetilde{O}^*(L))$ is *reduced* w.r.t. L, i.e., $\mathrm{Pre}(\widetilde{O}^*(L)) \in \mathcal{PFG}_L$, and is the least element of $\mathcal{PFG}_L$. ■

In what follows, we denote by $G_L^{\inf}$ and $G_L^{\sup}$ the least element and the greatest element in $\mathcal{PFG}_L$, respectively. Clearly $G_L^{\sup}$ consists of all symbols of Σ appearing in some string of L. That is,

$$G_L^{\inf} = \mathrm{Pre}(\widetilde{O}^*(L)), \quad G_L^{\sup} = \{a \in \Sigma \mid a \text{ appears in some string in } L\},$$

and $G_L^{\inf} \sqsubseteq G \sqsubseteq G_L^{\sup}$ for any $G \in \mathcal{PFG}_L$.

As a direct result of the above theorem, it follows that:

Corollary 2.10 *Let L_1 and L_2 be languages. If $L_1 \subseteq L_2$, then $G_{L_1}^{\inf} \sqsubseteq G_{L_2}^{\inf}$.*

Let L be a language and $w \in L$. For a prefix-free generating set G, the *length* of w w.r.t. G, denoted by $|w|_G$, is defined by the number of strings in G (allowing repetitions) appearing in the unique factorization for w, i.e.,

$$|w|_G = m, \text{ if } w = u_1 u_2 \cdots u_m \text{ for some } u_i \in G \ (i = 1, \cdots, m).$$

Since G is prefix-free, every string $w \in L$ has a unique factorization of strings in G. Thus the above length function $|\ldots|_G$ is well-defined.

As easily seen, $|w|_{G_L^{\inf}} \leq |w|_G$ for any $w \in L$ and for any $G \in \mathcal{PFG}_L$. In this sense, the least element $G_L^{\inf}$ is a *good* generating set for the language L.

In what follows, we consider the case that $G_L^{\inf}$ is finite.

If a language L is finite, so is $G_L^{\inf}$. In fact, the length of each string in $G_L^{\inf}$ is less than or equal to that of the longest strings in L. Thus the cardinality of $G_L^{\inf}$ is at most finite.

Theorem 2.11 *Let L be a language. $G_L^{\inf}$ is finite if and only if there is a finite subset S of L such that $L \sqsubseteq G_S^{\inf}$.*

Proof. *If part.* Let S be a finite subset of L satisfying $L \sqsubseteq G_S^{\inf}$. Since $G_S^{\inf}$ is reduced w.r.t. S, it must be reduced w.r.t. the superset L of S. Thus $G_S^{\inf} \in \mathcal{PFG}_L$. By Theorem 2.9, $G_L^{\inf} \sqsubseteq G_S^{\inf}$. On the other hand, by $S \subseteq L$, Corollary 2.10 implies $G_S^{\inf} \sqsubseteq G_L^{\inf}$. Hence $G_S^{\inf} = G_L^{\inf}$. Since S is finite, and thus so is $G_L^{\inf}$.

Only if part. Assume that $G_L^{\inf}$ is finite. Let w be any string of $G_L^{\inf}$, i.e., $w \in \mathrm{Pre}(\widetilde{O}^*(L))$. Then by Lemma 2.7, there is a *finite* set $S_w \subseteq L$ such that $w \in \mathrm{Pre}(\widetilde{O}^*(S_w))$. Put $S = (\bigcup_{w \in G_L^{\inf}} S_w)$. Since $G_L^{\inf}$ is finite, so is S. Furthermore, it can be easily seen that

$$G_L^{\inf} \subseteq \bigcup_{w \in G_L^{\inf}} \mathrm{Pre}(\widetilde{O}^*(S_w)) \subseteq \bigcup_{w \in G_L^{\inf}} \widetilde{O}^*(S_w) \subseteq \widetilde{O}^*(S) \sqsubseteq G_S^{\inf}.$$

Hence $G_L^{\inf} \sqsubseteq G_S^{\inf}$. The converse is trivial since $S \subseteq L$ holds. Consequently $G_L^{\inf} = G_S^{\inf}$ for the finite set $S \subseteq L$. ■

The class of regular languages are known to be an important class located in the lowest in Chomsky hierarchy, but the least elements for regular languages considered are not always finite. In fact, the language ab^*a over the alphabet $\Sigma = \{a, b\}$ is regular but $G_L^{\inf} = L$ to be infinite.

For a string $w \in \Sigma^+$, head(w) represents the first letter of w. We consider a particular subclass of $\mathcal{PFG}_L$ introduced by Yokomori[5]:

A generating set G of L is *simple* if head$(u) \neq$ head(v) for any $u, v \in G$ with $u \neq v$. Clearly a simple generating set is prefix-free and the cardinality of any simple prefix-free set is less than or equal to that of Σ.

By $\mathcal{SPFG}_L$ be the subclass of $\mathcal{PFG}_L$ consisting of all simple reduced and prefix-free generating sets of L.

Watanabe[6] has shown the next result on $\mathcal{SPFG}_L$:

Theorem 2.12 (Watanabe[6]) *For every language L, the set $\mathcal{SPFG}_L$ is a finite lattice.*

Yokomori[5] gave an algorithm of finding the least element $G_L^{\inf}$ in the subclass $\mathcal{SPFG}_L$ for a given finite set $L \subseteq \Sigma^+$ in polynomial time. We deal with the problem of finding the least element in $\mathcal{PFG}_L$ in the next paragraph.

3 Polynomial Time Algorithms for Computing $G_L^{\inf}$. In this section, we first present an efficient algorithm for computing a reduced and prefix-free generating set $G_L^{\inf}$ of a finite given language L. For an infinite language, we give an efficient learning algorithm for $G_L^{\inf}$ in the framework of *identification in the limit* due to Gold[4], provided that $G_L^{\inf}$ is finite.

3.1 An Algorithm for a Finite Language. We first consider $G_L^{\inf}$ for a finite language L. As shown in the previous section, $G_L^{\inf} = \mathrm{Pre}(\widetilde{O}^*(L))$. If L is finite, $\widetilde{O}^n(L) = \widetilde{O}^{n+1}(L)$ for some n, and $\widetilde{O}^*(L) = \widetilde{O}^n(L)$. Thus it is easy to compute the set $G_L^{\inf}$, but the number n of operations $\widetilde{O}$ may be exponential even if L is finite. In order to avoid it, we introduce another operations instead of O and $\widetilde{O}$ as follows: For $u, v \in \Sigma^+$ and $S \subseteq \Sigma^+$,

$$
\begin{aligned}
O'(u,v) &= \begin{cases} \{w\}, & \text{if } \exists w \in \Sigma^+ \text{ s.t. } v = uw, \\ \phi, & \text{otherwise,} \end{cases} \\
\widetilde{O}'(S) &= \bigcup_{\substack{u \in \mathrm{Pre}(S) \\ v \in S - \mathrm{Pre}(S)}} O'(u,v) \cup \mathrm{Pre}(S).
\end{aligned}
$$

Similarly to the definition of the operation $\widetilde{O}$, we define $\widetilde{O}'^0(S) = S$ and $\widetilde{O}'^{n+1}(S) = \widetilde{O}'(\widetilde{O}'^n(S))$ for each $n \in N$.

Clearly if $\widetilde{O}'^n(S)$ is prefix-free for some n, $\widetilde{O}'^n(S) = \widetilde{O}'^m(S)$ for any $m \geq n$, and we denote it by $\widetilde{O}'^*(S)$.

Lemma 3.1 *Let $S \subseteq \Sigma^+$ be nonempty set and $n \in N$. Then*
(i) $\widetilde{O}'^n(S) \sqsubseteq \widetilde{O}'^{n+1}(S)$, (ii) $\widetilde{O}'^n(S) \subseteq \widetilde{O}^n(S)$.

Proof. By the definition of the operation $\widetilde{O}'$, (i) is clearly valid. Thus we show (ii) only.

We prove (ii) by induction on $n \in N$. It is clear for $n = 0$. Assume that it is valid for any $n \leq k$ ($k \in N$). Let w be any string of $\widetilde{O}'^{k+1}(S)$. If $w \in \widetilde{O}'^k(S)$, our induction hypothesis

yields $w \in \widetilde{O}^k(S)$, and thus $w \in \widetilde{O}^{k+1}(S)$. Otherwise there are strings $u \in \mathrm{Pre}(\widetilde{O}'^k(S))$ and $v \in \widetilde{O}'^k(S) - \mathrm{Pre}(\widetilde{O}'^k(S))$ such that $v = uw$. By the induction hypothesis, $u, v \in \widetilde{O}^k(S)$. Since $w \in O(u, v)$, we have $w \in O^{k+1}(S)$. Hence $\widetilde{O}'^{k+1}(S) \subseteq \widetilde{O}^{k+1}(S)$. Consequently (ii) is valid for any $n \in N$. ■

For a finite set S of strings, we denote by $\sharp S$ and $||S||$ the number and the sum of lengths of strings contained in S, respectively.

Lemma 3.2 *Let $S \subseteq \Sigma^+$ be a finite set and $n \in N$. Then*

(i) $\sharp\widetilde{O}'^n(S) \geq \sharp\widetilde{O}'^{n+1}(S)$, where the equality is valid if $\widetilde{O}'^n(S)$ is prefix-free.

(ii) $||\widetilde{O}'^n(S)|| \geq ||\widetilde{O}'^{n+1}(S)||$, where the equality is valid if and only if $\widetilde{O}'^n(S)$ is prefix-free.

Proof. As noted above, if $\widetilde{O}'^n(S)$ is prefix-free, then the equalities of (i) and (ii) are valid. Suppose that $\widetilde{O}'^n(S)$ is not prefix-free.

By the definition of $\widetilde{O}'$, $\mathrm{Pre}(\widetilde{O}'^n(S)) \subseteq \widetilde{O}'^{n+1}(S)$. Thus it is enough to show that (i)' $\sharp S_n \geq \sharp S'_{n+1}$ and (ii)' $||S_n|| > ||S'_{n+1}||$, where

$$S_n = \widetilde{O}'^n(S) - \mathrm{Pre}(\widetilde{O}'^n(S)), \qquad S'_{n+1} = \widetilde{O}'^{n+1}(S) - \mathrm{Pre}(\widetilde{O}'^n(S)).$$

As easily seen, for each $w \in S'_{n+1}$, there are two strings $u_w \in \mathrm{Pre}(\widetilde{O}'^n(S))$ and $v_w \in S_n$ such that $v_w = u_w w$.

Since $u_w \in \mathrm{Pre}(\widetilde{O}'^n(S))$, it follows that $v_w \neq v_{w'}$ for any $w' \in S'_{n+1}$ with $w' \neq w$. Hence $\sharp S_n \geq \sharp S'_{n+1}$, i.e., (i)' is valid.

The inequality $||S_n|| > ||S'_{n+1}||$ can be derived from $|v_w| > |w|$ for each $w \in S'_{n+1}$. ■

Theorem 3.3 *For every finite language L, $\widetilde{O}'^*(L) = G_L^{\inf}$.*

Proof. Since L is finite, by Lemma 3.2 $\widetilde{O}'^n(L)$ eventually becomes prefix-free for some n, say n_0. Thus $\widetilde{O}'^n(L) = \widetilde{O}'^{n_0}(L)$ for any $n \geq n_0$. It means that $\widetilde{O}'^*(L) = \widetilde{O}'^{n_0}(L)$. By Lemma 3.1, $\widetilde{O}'^*(L) \subseteq \widetilde{O}^{n_0}(L) \subseteq \widetilde{O}^*(L) \sqsubseteq G_L^{\inf}$. This implies $\widetilde{O}'^*(L) \sqsubseteq G_L^{\inf}$.

In order to show the converse, we prove $\widetilde{O}^n(L) \sqsubseteq \widetilde{O}'^*(L)$ by induction on $n \in N$. It is clear for $n = 0$. Assume that $\widetilde{O}^k(L) \sqsubseteq \widetilde{O}'^*(L)$. Let $w \in \widetilde{O}^{k+1}(L)$. If $w \in \widetilde{O}^k(L)$, our induction hypothesis yields $w \in (\widetilde{O}'^*(L))^+$. Otherwise there is a direct ancestor $\{u, v\} \subseteq \widetilde{O}^k(L)$ of w such that $u = vw$. By the induction hypothesis, $u, v \in (\widetilde{O}'^*(L))^+$. Since $\widetilde{O}'^*(L)$ is prefix-free, we obtain $w \in (\widetilde{O}'^*(L))^+$. Thus $\widetilde{O}^n(L) \sqsubseteq \widetilde{O}'^*(L)$ for any $n \in N$. It implies that $\widetilde{O}^*(L) \sqsubseteq \widetilde{O}'^*(L)$. By Lemma 2.1, $G_L^{\inf} \sqsubseteq \widetilde{O}'^*(L)$.

Consequently we have $G_L^{\inf} = \widetilde{O}'^*(L)$. ■

We first present a procedure for computing $\widetilde{O}'(S)$ for a given finite set S:

Algorithm $\widetilde{O}'(S)$
Input: a finite set S of strings;
Output: the set $\widetilde{O}'(S)$;

begin
  $T := \phi$;
  for each $(u, v) \in \mathrm{Pre}(S) \times (S - \mathrm{Pre}(S))$ **do** $T := T \cup O'(u, v)$;
  output $T \cup \mathrm{Pre}(S)$
end.

Let $n = \sharp S$ and $m = \max\{|w| \mid w \in S\}$. In the above procedure, $\mathrm{Pre}(S)$ can be computed in time $O(n^2m)$, and for each pair (u, v), $O'(u, v)$ can be computed in $O(m)$. Thus the procedure for $\widetilde{O}'(S)$ correctly outputs $\widetilde{O}'(S)$ in time $O(n^2m)$.

Now we give a polynomial time algorithm for computing $G_L^{\inf}$ as follows:

Algorithm $G_L^{\inf}$
Input: a finite language L;
Output: the least element $G_L^{\inf}$;

begin
$T := L$;
repeat
$T' := T$; $T := \widetilde{O}'(T)$
until $T = T'$;
output T
end.

Theorem 3.4 *Let L be a finite language. Then Algorithm $G_L^{\inf}$ correctly computes $G_L^{\inf}$ in time $O(n^3m^2)$, where $n = \sharp L$ and $m = \max\{|w| \mid w \in L\}$.*

Proof. As mentioned above, each $\widetilde{O}'(T)$ can be computed in time $O(n^2m)$. By Lemma 3.2(ii), the number of the repetitions until $T = T'$ is at most nm. Hence the time of complexity is given by $O(n^3m^2)$. ■

3.2 Identification of $G_L^{\inf}$ in the Limit. In this subsection, we consider the problem of identifying $G_L^{\inf}$ in the frame-work of inductive inference based on *identification in the limit* introduced by Gold[4] for language learning, provided $G_L^{\inf}$ is finite.

Inductive inference is a process to guess an unknown general rule from given examples. Gold[4] proposed a mathematical model of inductive inference based on a criterion called *identification in the limit* as follows: A *positive presentation* σ of a language L is an infinite sequence $w_1, w_2, \cdots$ of strings such that $\{w_n \mid n \geq 1\} = L$. An *inference machine* M is an effective procedure that requests a string and produces a *hypothesis* at a time. Given a positive presentation $\sigma = w_1, w_2, \cdots$, M generates an infinite sequence $g_1, g_2, \cdots$ of hypotheses. In language identification, hypotheses mean some devices defining languages such as automata, formal grammars and so on. The inference machine M identifies the target language from positive examples, if for any positive presentation σ the sequence of the hypotheses $g_1, g_2, \cdots$ generated by M converges to some hypothesis g which defines the target language. A language L is inferable from positive examples, if there exists an inference machine which identifies L from positive examples. Refer in detail to Gold[4].

In this paper, the goal of the learning process is the least prefix-free generating set $G_L^{\inf}$ but not usual devices defining the target language L. Thus the space of hypotheses is the class of all prefix-free sets to be finite. According to the above, an inference machine M identifies the least prefix-free generating set $G_L^{\inf}$ of L from positive examples in the limit, if there is an integer n_0 such that $g_n = G_L^{\inf}$ for any $n \geq n_0$.

Let $G_1, G_2, \cdots$ be an infinite sequence of sets of strings. The sequence $G_1, G_2, \cdots$ converges to a set $G \subseteq \Sigma^+$, denoted by $\lim_{n\to\infty} G_n = G$, if there exists an integer n_0 such that $G_n = G$ for any $n \geq n_0$.

Let $\sigma = w_1, w_2, \cdots$ be a positive presentation of L, and let $S_n = \{w_1, w_2, \cdots, w_n\}$ for each $n \in N$.

Lemma 3.5 *Let L be a language. If $G_L^{\inf}$ is finite, then*

$$\lim_{n\to\infty} G_{S_n}^{\inf} = G_L^{\inf}.$$

Proof. By Theorem 2.11, there exists a finite subset S of L such that $G_S^{\inf} = G_L^{\inf}$. Since S is finite, $S \subseteq S_{n_0}$ for some $n_0 \in N$. Using Corollary 2.10, $G_S^{\inf} \sqsubseteq G_{S_n}^{\inf} \sqsubseteq G_L^{\inf}$ for any $n \geq n_0$. Hence $G_S^{\inf} = G_L^{\inf}$ implies $G_{S_n}^{\inf} = G_L^{\inf}$ for any $n \geq n_0$. ■

Lemma 3.6 *For a finite set $S \subseteq \Sigma^+$ and a string $w \in \Sigma^+$,*

$$G_{S\cup\{w\}}^{\inf} = G_{G_S^{\inf}\cup\{w\}}^{\inf}.$$

Proof. Clearly $S \cup \{w\} \sqsubseteq G_S^{\inf} \cup \{w\}$. By Theorem 2.9, $G_{S\cup\{w\}}^{\inf}$ is the least element of $\mathcal{PFG}_{S\cup\{w\}}$, and so $G_{S\cup\{w\}}^{\inf} \sqsubseteq G_{G_S^{\inf}\cup\{w\}}^{\inf}$.

Next, we prove the converse. Since $G_S^{\inf} \cup \{w\} \subseteq O^*(S \cup \{w\})$, $G_{S\cup\{w\}}$ is a prefix-free generating set of $G_S^{\inf} \cup \{w\}$. Appealing to Theorem 2.9, we get $G_{G_S^{\inf}\cup\{w\}}^{\inf} \sqsubseteq G_{S\cup\{w\}}^{\inf}$. ■

Now we present an inference algorithm as follows:

Algorithm LA
Input: a positive presentation of a language L;
Output: a sequence of reduced and prefix-free generating sets;

begin
  $G_0 := \phi$; $n := 1$;
  repeat
    read the next data w_n;
    $G_n := G_{G_{n-1}\cup\{w_n\}}^{\inf}$;
    output G_n as the n-th conjecture;
    $n := n + 1$
  forever
end.

For each n, let $S_n = \{w_1, \cdots, w_n\}$ be a sample set of a target language L, and G_n be the n-th hypothesis of the above algorithm.

Theorem 3.7 *Let L be a language. If $G_L^{\inf}$ is finite, the algorithm LA identifies $G_L^{\inf}$ from positive examples in the limit, and may be implemented to update the n-th hypothesis in time $O(n^3m^2)$, where $m = \max\{|w_i| \mid i = 1, 2, \cdots, n\}$.*

Proof. By Lemma 3.6, it is easy to show that $G_n = G_{S_n}^{\inf}$ for any n. Appealing to Lemma 3.5, we obtain $\lim_{n\to\infty} G_n = G_L^{\inf}$ because $G_L^{\inf}$ is finite. Thus the algorithm identifies $G_L^{\inf}$ in the limit.

Using Theorem 3.3, $\sharp G_{S_n}^{\inf} \leq n$ and the length of the longest strings in $G_{S_n}^{\inf}$ is less than or equal to that in S_n. Thus by Theorem 3.4, the algorithm LA may be implemented to update the n-th conjecture $G_n = G_{S_{n-1}\cup\{w_n\}}^{\inf}$ in time $O(n^3m^2)$, where $m = \max\{|w_i| \mid i = 1, 2, \cdots, n\}$. ■

Acknowledgements

The author would like to give his thanks to Prof. Masako Sato and Dr. Yasuhito Mukouchi of Osaka prefecture university for their valuable suggestions and comments.

References

[1] Angluin, D.: *Inductive Inference of Formal Languages from Positive Data*. Information and Control, **45** (1980), 117–135.

[2] Ash, R.: "Information Theory." Interscience Publishers, (1965).

[3] Capocelli, R.M.: *A Decision Procedure for Finite Decipherability and Synchronizability of Multivalued Encodings*. IEEE Transactions on Information Theory, **IT-28**(2) (1982), 307–318.

[4] Gold, E.M.: *Language Identification in the Limit*. Information and Control, **10** (1967), 447–474.

[5] Yokomori, T.: *On Polynomial-Time Learnability in the Limit of Strictly Deterministic Automata*. Machine Learning, **19** (1995), 153–179.

[6] Watanabe, N.: *Polynomial-Time Inductive Inference of Simple Regular Automata*. Master thesis, Osaka Prefecture University, (1996).

Department of Mathematics and Information Sciences,
Graduate School, Osaka Prefecture University, Sakai, Osaka, 599-8531, Japan