

# The Strong, Weak, and Very Weak Finite Context and Kernel Properties

Makoto Kanazawa<sup>1</sup> and Ryo Yoshinaka<sup>2</sup>

<sup>1</sup> National Institute of Informatics and SOKENDAI, Tokyo 101–8430, Japan

<sup>2</sup> Graduate School of Information Sciences, Tohoku University, Sendai, Miyagi 980–8579, Japan

**Abstract.** We identify the properties of context-free grammars that exactly correspond to the behavior of the dual and primal versions of Clark and Yoshinaka’s distributional learning algorithm and call them the *very weak finite context/kernel property*. We show that the very weak finite context property does not imply Yoshinaka’s *weak finite context property*, which has been assumed to hold of the target language for the dual algorithm to succeed. We also show that the weak finite context property is genuinely weaker than Clark’s *strong finite context property*, settling a question raised by Yoshinaka.

**Keywords:** grammatical inference and algorithmic learning, distributional learning, finite context property, finite kernel property, context-free languages

## 1 Introduction

Clark [2] and Yoshinaka [8] pioneered an approach to efficient learning of context-free languages under the paradigm of *polynomial-time identification in the limit from positive data and membership queries*, which they called *distributional learning*.<sup>1</sup> The idea of distributional learning was based on the assumption that the target language has a context-free grammar each of whose nonterminals is *characterized* either by a finite set of strings or by a finite set of *contexts* (i.e., pairs of strings). There are strong and weak variants to this notion of characterization: A nonterminal  $X$  is *strongly* characterized by a set  $C$  of contexts if the strings derived from  $X$  are exactly those that can appear in all contexts from  $C$ ;  $X$  is strongly characterized by a set  $K$  of strings if  $X$  is strongly characterized by the set of contexts in which all elements of  $K$  can appear. The *weak* notion of characterization is obtained by replacing “the strings derived from  $X$ ” in this definition by their *closure*, i.e., “the strings that can appear in every context in which all strings derived from  $X$  can appear”. A context-free grammar  $G$  has the *strong* (resp. *weak*) *finite context property (FCP)* if each nonterminal of  $G$  is strongly (resp. weakly) characterized by a finite set of contexts;  $G$  has the *strong*

---

<sup>1</sup> Clark and Yoshinaka have used the term “distributional learning” more loosely in connection with a number of different learning paradims (see [5] for a survey).

(resp. *weak*) *finite kernel property (FKP)* if each nonterminal of  $G$  is strongly (resp. weakly) characterized by a finite set of strings. The *dual* variant of the distributional learner uses finite sets of contexts as nonterminals and succeeds under the assumption that the target grammar has the weak FCP, while the *primal* variant uses finite sets of strings as nonterminals and assumes that the target grammar has the weak FKP. Distributional learning, both in its dual and primal variants, has since been extended to string and tree grammar formalisms that are “context-free” in a broad sense of the term [11,9,5,4].

Despite the naturalness and wide applicability of the approach, there is a notable discrepancy between the assumption of the (weak) FCP/FKP and the behavior of the learning algorithm. For, in hypothesizing a rule, what the algorithm tries to check is the “local” requirement that the rule be *valid* under the presumption that each nonterminal in the rule is strongly characterized by itself. For example, when the dual algorithm constructs a binary rule  $C_0 \rightarrow C_1 C_2$ , where each  $C_i$  is a finite set of contexts, it checks whether the available evidence is consistent with the assumption  $C_0^{\mathcal{L}} \supseteq C_1^{\mathcal{L}} C_2^{\mathcal{L}}$ , where  $C_i^{\mathcal{L}}$  denotes the set of strings that can appear in each context from  $C_i$  in the target language. (Note that the definition of the set  $C_i^{\mathcal{L}}$  refers to the target language, not to any grammar for it.) In contrast, the weak or strong FCP/FKP of a context-free grammar is a “global” property of the grammar, since it refers to the set of strings derived from each nonterminal, something that you cannot determine just by looking at each individual rule in isolation.

The original idea of Clark [2] was to identify each nonterminal with a *closed* set of strings (i.e., a set that is its own closure). Thus, he employed the strong variant of the finite context property. Yoshinaka [8] recognized the possibility that the set of strings derived from a nonterminal in the hypothesized grammar may not be a closed set, and introduced the weak variants of the FCP and FKP. In his later paper [10] on distributional learning of conjunctive grammars, Yoshinaka mentioned as an open problem the question of whether the context-free grammars with the strong FCP generate the same languages as the context-free grammars with the weak FCP. However, the question of whether the weak FCP and FKP are weak enough, i.e., whether the two variants of the distributional learner always converge to a grammar with the weak FCP/FKP, does not seem to have attracted much attention.

In fact, the properties of context-free grammars that exactly correspond to the behavior of the dual and primal variants of the distributional learner are easy to state; we call them the *very weak FCP/FKP*. The dual/primal algorithm converges to a correct grammar for the target language when the latter has a grammar satisfying the very weak FCP/FKP, and the grammar it converges to always satisfies the very weak FCP/FKP. As we show below, the very weak FKP turns out to be equivalent to the weak FKP, and the primal algorithm is guaranteed to converge to a grammar with the weak FKP. In contrast, the very weak FCP is genuinely weaker than the weak FCP. We exhibit a language that has a context-free grammar with the very weak FCP but has no context-free grammar with the weak FCP. The fact that the dual algorithm succeeds on such

languages has not been recognized so far. We also show that a similar separation holds between the weak and strong variants of both the FCP and the FKP. This negatively settles the above-mentioned open problem of Yoshinaka [10].

## 2 Preliminaries

We adopt the standard definition of context-free grammars (as in [1]), except that we allow them to have multiple initial nonterminals. Thus, a context-free grammar (CFG) is a 4-tuple  $G = (N, \Sigma, P, I)$ , where  $I \subseteq N$  is the set of initial nonterminals. If  $A$  is a nonterminal of  $G$ , we write  $L(G, A)$  for  $\{w \in \Sigma^* \mid A \Rightarrow_G^* w\}$ , the set of terminal strings derived from  $A$ . Then the language of  $G$  is  $L(G) = \bigcup_{A \in I} L(G, A)$ . We write  $C(G, A)$  for  $\{(u, v) \in \Sigma^* \times \Sigma^* \mid S \Rightarrow_G^* uAv \text{ for some } S \in I\}$ .

Let  $G = (N, \Sigma, P, I)$  be a CFG. It is well known that  $G$  can be thought of as a system of equations where the nonterminals are viewed as variables ranging over  $\mathcal{P}(\Sigma^*)$ . This system contains the equation  $A = \alpha_1 \cup \dots \cup \alpha_k$  for each nonterminal  $A$ , where  $\alpha_1, \dots, \alpha_k$  list the right-hand sides of the productions with  $A$  on the left-hand side. A sequence of sets  $(X_A)_{A \in N}$  is a *fixed point* of  $G$  if assigning the value  $X_A$  to each  $A$  satisfies all these equations. The sequence  $(X_A)_{A \in N}$  is a *pre-fixed point* of  $G$  if it instead satisfies  $A \supseteq \alpha_1 \cup \dots \cup \alpha_k$ . Equivalently,  $(X_A)_{A \in N}$  is a pre-fixed point of  $G$  if for each production  $A \rightarrow w_0 A_1 w_1 \dots A_n w_n$  in  $P$  (with  $w_i \in \Sigma^*$  and  $A_i \in N$ ), it holds that  $X_A \supseteq w_0 X_{A_1} w_1 \dots X_{A_n} w_n$ . It is known that the sequence  $(L(G, A))_{A \in N}$  is the *least fixed point* as well as the *least pre-fixed point* of  $G$  under the partial order of componentwise inclusion.

## 3 Three Variants of the Finite Context and Kernel Properties

We begin by reviewing some important notions from Clark's *syntactic concept lattice* [3]. Let  $L \subseteq \Sigma^*$  be given. For  $C \subseteq \Sigma^* \times \Sigma^*$  and  $K \subseteq \Sigma^*$ , we put

$$C^{\langle L \rangle} = \{x \in \Sigma^* \mid uxv \in L \text{ for all } (u, v) \in C\},$$

$$K^{\langle L \rangle} = \{(u, v) \in \Sigma^* \times \Sigma^* \mid uKv \subseteq L\}.$$

When the language  $L$  is understood from context, these are simply written  $C^{\langle \cdot \rangle}$  and  $K^{\langle \cdot \rangle}$ . For a set  $K \subseteq \Sigma^*$ , its *closure* is  $K^{\langle \cdot \rangle}$ . The function  $(\cdot)^{\langle \cdot \rangle}$  is indeed a closure operator in the sense that (i)  $K \subseteq K^{\langle \cdot \rangle}$ , (ii)  $K_1 \subseteq K_2$  implies  $K_1^{\langle \cdot \rangle} \subseteq K_2^{\langle \cdot \rangle}$ , and (iii)  $K^{\langle \cdot \rangle \langle \cdot \rangle} = K^{\langle \cdot \rangle}$ . A set  $K \subseteq \Sigma^*$  is *closed* if  $K = K^{\langle \cdot \rangle}$ ; equivalently,  $K$  is closed if and only if there exists a  $C \subseteq \Sigma^* \times \Sigma^*$  such that  $K = C^{\langle \cdot \rangle}$ . (These notions are all relative to the given language  $L$ .) Note that  $L$  is always closed relative to  $L$ , since  $L^{\langle \cdot \rangle}$  always contains  $(\varepsilon, \varepsilon)$ . (We write  $\varepsilon$  for the empty string.)

An important property of the closure operator  $(\cdot)^{\langle \cdot \rangle}$  is the following [3]: for  $X, Y \subseteq \Sigma^*$ ,

$$(XY)^{\langle \cdot \rangle} = (X^{\langle \cdot \rangle} Y^{\langle \cdot \rangle})^{\langle \cdot \rangle}. \quad (1)$$

Let  $G = (N, \Sigma, P, I)$  be a CFG, and let the operators  $\langle \cdot \rangle$  and  $\triangleright$  be understood relative to  $L(G)$ . A pre-fixed point  $(X_A)_{A \in N}$  of  $G$  is *sound* if  $\bigcup_{A \in I} X_A = L(G)$ .

We abbreviate “sound pre-fixed point” to “SPP”. It is easy to see that if  $(X_A)_{A \in N}$  is an SPP of  $G$ , then  $uX_Av \subseteq L(G)$  for all  $(u, v) \in C(G, A)$ . Since  $(L(G, A))_{A \in N}$  is the least pre-fixed point of  $G$ , it is the least SPP.

**Proposition 1.** *If  $(X_A)_{A \in N}$  is an SPP, then so is  $(X_A^{\triangleright\triangleleft})_{A \in N}$ .*

*Proof.* This easily follows from the fact that  $(w_0X_{A_1}^{\triangleright\triangleleft}w_1 \dots X_{A_n}^{\triangleright\triangleleft}w_n)^{\triangleright\triangleleft}$  equals  $(w_0X_{A_1}w_1 \dots X_{A_n}w_n)^{\triangleright\triangleleft}$ , which in turn is a consequence of (1).  $\square$

Since  $(L(G, A)^{\triangleright\triangleleft})_{A \in N}$  is the least pre-fixed point consisting entirely of closed sets, it is the least such SPP. It need not be the greatest SPP of  $G$ . In fact, it is not hard to see that the greatest SPP may not exist.

Let  $k$  be a natural number, and let  $L \subseteq \Sigma^*$ . A set  $X \subseteq \Sigma^*$  is *k-context-generated* relative to  $L$  if  $X = C^{\langle L \rangle}$  for some  $C \subseteq \Sigma^* \times \Sigma^*$  such that  $|C| \leq k$ . We say that  $X$  is *k-kernel-generated* relative to  $L$  if  $X = K^{\langle L \rangle \langle L \rangle}$  for some  $K \subseteq \Sigma^*$  such that  $|K| \leq k$ . A sequence of sets is *k-context-generated* (*k-kernel-generated*) if each of its component sets is *k-context-generated* (*k-kernel-generated*).

We say that  $G$  has

- the *strong k-finite context property* (resp. *strong k-finite kernel property*) if  $(L(G, A))_{A \in N}$  is *k-context-generated* (resp. *k-kernel-generated*) relative to  $L(G)$ ;
- the *weak k-finite context property* (resp. *weak k-finite kernel property*) if  $(L(G, A)^{\triangleright\triangleleft})_{A \in N}$  is *k-context-generated* (resp. *k-kernel-generated*) relative to  $L(G)$ ;
- the *very weak k-finite context property* (resp. *very weak k-finite kernel property*) if  $G$  has an SPP that is *k-context-generated* (resp. *k-kernel-generated*) relative to  $L(G)$ .

We abbreviate “finite context property” to “FCP” and “finite kernel property” to “FKP”. Clearly, a CFG  $G$  has the strong *k*-FCP (*k*-FKP) if and only if  $G$  has the weak *k*-FCP (*k*-FKP) and in addition  $L(G, A)$  is a closed set relative to  $L(G)$  for each nonterminal  $A$  of  $G$ . It is also obvious that the weak *k*-FCP (*k*-FKP) implies the very weak *k*-FCP (*k*-FKP). We say that  $G$  has the strong/weak/very weak FCP/FKP when  $G$  has the strong/weak/very weak *k*-FCP/*k*-FKP for some  $k$ .

What Clark [2] called the finite context property was what we here call the strong finite context property. The weak finite context property was introduced by Yoshinaka [8] and adopted by Leiß [6].

- Proposition 2.** (i) *There is a language that has a CFG with the strong 1-FKP but has no CFG with the very weak FCP.*  
(ii) *There is a language that has a CFG with the strong 1-FCP but has no CFG with the very weak FKP.*

*Proof.* (i). Let

$$L = L_1 \cup L_2, \quad L_1 = \{a^mcb^{2m} \mid m \in \mathbb{N}\}, \quad L_2 = \{a^mdb^n \mid n \leq m \leq 2n\}.$$

Let  $G$  be the following CFG with initial nonterminals  $S_1$  and  $S_2$ :

$$S_1 \rightarrow aS_1bb \mid c, \quad S_2 \rightarrow aS_2b \mid aaS_2b \mid d.$$

Then  $L(G, S_1) = L_1 = \{c\}^{\triangleright\triangleleft}$  and  $L(G, S_2) = L_2 = \{d\}^{\triangleright\triangleleft}$ . So  $L(G) = L$  and  $G$  has the strong 1-FKP.

Now let  $G'$  be any CFG for  $L$ . Applying the pumping lemma to  $a^p cb^{2p}$  for a sufficiently large  $p$ , we get

$$S \Rightarrow_{G'}^* a^{i_1} E b^{j_1}, \quad E \Rightarrow_{G'}^+ a^i E b^j, \quad E \Rightarrow_{G'}^* a^{i_2} c b^{j_2}$$

with  $i + j > 0$ ,  $i_1 + i + i_2 = p$ , and  $j_1 + j + j_2 = 2p$ , for some nonterminal  $E$  and initial nonterminal  $S$ . Since  $a^{i_1+ni+i_2} c b^{j_1+nj+j_2} \in L$  for all  $n \in \mathbb{N}$ , we must have  $j = 2i$ , which implies  $i > 0$ . By way of contradiction, assume that  $G'$  has an SPP whose component for  $E$  is  $C_E^{\triangleleft}$  for some finite  $C_E \subseteq \Sigma^* \times \Sigma^*$ . We must have

$$L \supseteq a^{i_1} C_E^{\triangleleft} b^{j_1}, \quad C_E^{\triangleleft} \supseteq a^i C_E^{\triangleleft} b^{2i}, \quad C_E^{\triangleleft} \supseteq \{a^{i_2} c b^{j_2}\}.$$

By the last inclusion, every element of  $C_E$  must be of the form  $(a^{n-i_2}, b^{2n-j_2})$  for some  $n$  such that  $n \geq i_2$  and  $2n \geq j_2$ . Take an  $m$  large enough that  $m \geq 2n$  for all  $n$  such that  $(a^{n-i_2}, b^{2n-j_2}) \in C_E$ . Then for all such  $n$ ,

$$2m + 2n - j_2 \leq 3m \leq 3m + n - i_2 \leq 4m \leq 2(2m + 2n - j_2),$$

which implies  $a^{3m} d b^{2m} \in C_E^{\triangleleft}$ . But since  $L \supseteq a^{i_1+ni} C_E^{\triangleleft} b^{j_1+2ni}$  for all  $n \in \mathbb{N}$ , we get  $a^{i_1+ni+3m} d b^{j_1+2ni+2m} \in L$  and hence  $i_1 + ni + 3m \geq j_1 + 2ni + 2m$  for all  $n$ , which contradicts  $i > 0$ .

(ii). Let  $L = \{a^m b^n \mid m \neq n\}$ . Let  $G$  be the following CFG with a unique initial nonterminal  $S$ :

$$S \rightarrow A \mid B \mid aSb, \quad A \rightarrow a \mid aA, \quad B \rightarrow b \mid bB.$$

Then  $L(G, S) = L = \{(\varepsilon, \varepsilon)\}^{\triangleleft}$ ,  $L(G, A) = a^+ = \{(\varepsilon, ab)\}^{\triangleleft}$ ,  $L(G, B) = b^+ = \{(ab, \varepsilon)\}^{\triangleleft}$ , so  $G$  has the strong 1-FCP.

Now suppose that  $L$  has a CFG with the very weak FKP. Then  $L = K_1^{\triangleright\triangleleft} \cup \dots \cup K_n^{\triangleright\triangleleft}$  for some finite sets  $K_1, \dots, K_n \subseteq \Sigma^*$ . Clearly,  $K_i \subseteq L$  for  $i = 1, \dots, n$ . Let  $p = \max\{|m-n| \mid a^m b^n \in K_i \text{ for some } i\}$ . Then for each  $i$ ,  $K_i^{\triangleright} \supseteq \{a^m, b^n \mid |m-n| > p\}$ , which implies  $K_i^{\triangleright\triangleleft} \subseteq \{a^m b^n \mid |m-n| \leq p\}$ . This is a contradiction, since  $L \not\subseteq \{a^m b^n \mid |m-n| \leq p\}$ .  $\square$

## 4 Distributional Learning

The distributional learner has access to an infinite stream of positive examples and the membership oracle for the target language  $L_*$ . Let  $\triangleright$  and  $\triangleleft$  be understood relative to  $L_*$ .

The dual algorithm forms productions of the form  $C_0 \rightarrow w_0 C_1 w_1 \dots C_n w_n$ , where each nonterminal  $C_i$  is a finite set of contexts  $(u, v)$  contained in the input positive data  $D$  (i.e.,  $uvw \in D$  for some  $w$ ). Such a production is *valid* if

$$C_0^{\triangleleft} \supseteq w_0 C_1^{\triangleleft} w_1 \dots C_n^{\triangleleft} w_n.$$

The membership of a string in  $C_i^{\triangleleft}$  can be determined by a membership query, but since  $C_i^{\triangleleft}$  is in general infinite, the validity of a production cannot be determined in finite time. At each stage, the algorithm uses only those productions that are *valid on* the set  $E$  of all substrings of the positive examples, in the sense that

$$C_0^{\triangleleft} \supseteq w_0(E \cap C_1^{\triangleleft})w_1 \dots (E \cap C_n^{\triangleleft})w_n.$$

Testing this condition for all candidate productions can be done with a polynomial number of membership queries if there is a fixed bound on  $n$  and the cardinality of each  $C_i$ . Since  $E$  continues to grow, if the output of the algorithm stabilizes on a particular grammar, all its productions will be valid.

The primal algorithm in contrast constructs productions of the form  $K_0 \rightarrow w_0K_1w_1 \dots K_nw_n$  where each  $K_i$  is a finite set of strings  $w$  drawn from the input positive data  $D$  (i.e.,  $uvw \in D$  for some  $u, v$ ). Such a production is *valid* if  $K_0^{\triangleright\triangleleft} \supseteq w_0K_1^{\triangleright\triangleleft}w_1 \dots K_n^{\triangleright\triangleleft}w_n$ . It is not difficult to see that this condition is equivalent to

$$K_0^{\triangleright} \subseteq (w_0K_1w_1 \dots K_nw_n)^{\triangleright}.$$

In hypothesizing a production, the primal algorithm checks that it is *valid on* the set  $J$  of all contexts contained in the input positive data, in the following sense:

$$J \cap K_0^{\triangleright} \subseteq (w_0K_1w_1 \dots K_nw_n)^{\triangleright}.$$

Again, this ensures the validity of all productions in the limit.

---

**Algorithm 1: Dual learner for CFGs.**


---

**Parameters:** Positive integers  $r, k$ ;

**Data:** A positive presentation  $t_1, t_2, \dots$  of  $L_* \subseteq \Sigma^*$ ; membership oracle for  $L_*$ ;

**Result:** A sequence of grammars  $G_1, G_2, \dots$ ;

let  $D_0 := \emptyset; E_0 := \emptyset; J_0 := \emptyset; H_0 := \emptyset; G_0 := (\emptyset, \Sigma, \emptyset, \emptyset)$ ;

**for**  $i = 1, 2, \dots$  **do**

let  $D_i := D_{i-1} \cup \{t_i\}; E_i := \text{Sub}(D_i)$ ;

**if**  $D_i \not\subseteq L(G_{i-1})$  **then**

let  $J_i := \text{Con}(D_i); H_i := \text{Sub}^{\leq r+1}(D_i)$ ;

**else**

let  $J_i := J_{i-1}; H_i := H_{i-1}$ ;

output  $G_i := (N_i, \Sigma, P_i, I_i)$  where

$N_i = \{C \subseteq J_i \mid 1 \leq |C| \leq k\}$ ,

$P_i = \{C_0 \rightarrow w_0C_1w_1 \dots C_nw_n \mid (w_0, w_1, \dots, w_n) \in H_i,$

$C_0, C_1, \dots, C_n \in N_i, C_0 \rightarrow w_0C_1w_1 \dots C_nw_n \text{ is valid on } E_i\}$ ,

$I_i = \{C \in N_i \mid E_i \cap C^{\langle L_* \rangle} \subseteq L_*\}$ ;

---

Exact formulations of the two algorithms are given in Algorithms 1 and 2.<sup>2</sup> In the algorithms,  $\text{Sub}(D) = \{w \in \Sigma^* \mid uvw \in D \text{ for some } u, v\}$ ,  $\text{Sub}^n(D) =$

<sup>2</sup> The present formulations follow [4].

$\{(w_1, \dots, w_n) \in (\Sigma^*)^n \mid u_0 w_1 u_1 \dots w_n u_n \in D \text{ for some } u_0, \dots, u_n\}$ ,  $\text{Sub}^{\leq m}(D)$   
 $= \bigcup_{1 \leq n \leq m} \text{Sub}^n(D)$ , and  $\text{Con}(D) = \{(u, v) \in \Sigma^* \times \Sigma^* \mid uvv \in D \text{ for some } w\}$ .

---

**Algorithm 2:** Primal learner for CFGs.
 

---

**Parameters:** Positive integers  $r, k$ ;  
**Data:** A positive presentation  $t_1, t_2, \dots$  of  $L_* \subseteq \Sigma^*$ ; membership oracle for  $L_*$ ;  
**Result:** A sequence of grammars  $G_1, G_2, \dots$ ;  
 let  $D_0 := \emptyset; E_0 := \emptyset; J_0 := \emptyset; H_0 := \emptyset; G_0 := (\emptyset, \Sigma, \emptyset, \emptyset)$ ;  
**for**  $i = 1, 2, \dots$  **do**  
     let  $D_i := D_{i-1} \cup \{t_i\}; J_i := \text{Con}(D_i)$ ;  
     **if**  $D_i \not\subseteq L(G_{i-1})$  **then**  
         | let  $E_i := \text{Sub}(D_i); H_i := \text{Sub}^{\leq r+1}(D_i)$ ;  
     **else**  
         | let  $E_i := E_{i-1}; H_i := H_{i-1}$ ;  
     output  $G_i := (N_i, \Sigma, P_i, I_i)$  where  
          $N_i = \{K \subseteq E_i \mid 1 \leq |K| \leq k\}$ ,  
          $P_i = \{K_0 \rightarrow w_0 K_1 w_1 \dots K_n w_n \mid (w_0, w_1, \dots, w_n) \in H_i,$   
                  $K_0, K_1, \dots, K_n \in N_i, K_0 \rightarrow w_0 K_1 w_1 \dots K_n w_n \text{ is valid on } J_i\}$ ,  
          $I_i = \{K \in N_i \mid K \subseteq L_*\}$ ;

---

We say that the dual or primal learner *converges to  $G$  on  $L_*$*  if, given a positive presentation (i.e., enumeration) of  $L_*$  and the membership oracle for  $L_*$ , the output of the learner eventually stabilizes on  $G$ .

**Theorem 3.** *Suppose that the dual learner converges to a grammar  $G$  on  $L_*$ . Then  $L(G) = L_*$  and  $G$  has the very weak  $k$ -FCP.*

*Proof.* Let  $G = (N, \Sigma, P, I)$ . It is easy to see that  $D_i \subseteq L(G_i)$  holds at each stage, so  $L_* \subseteq L(G)$ . Since every element of  $\text{Sub}(L_*)$  eventually appears in  $E_i$ , each production of  $G$  must be valid. Hence the sequence  $(C^{\langle L_* \rangle})_{C \in N}$  is a pre-fixed point of  $G$ . By the same token, we must have  $C^{\langle L_* \rangle} \subseteq L_*$  for all  $C \in I$ . So  $\bigcup_{C \in I} C^{\langle L_* \rangle} \subseteq L_* \subseteq L(G)$ . Since  $(L(G, C))_{C \in N}$  is the least pre-fixed point of  $G$ ,  $L(G) = \bigcup_{C \in I} L(G, C) \subseteq \bigcup_{C \in I} C^{\langle L_* \rangle}$ . Hence  $L(G) = L_*$  and  $(C^{\langle L(G) \rangle})_{C \in N}$  is an SPP of  $G$ , which means that  $G$  has the very weak  $k$ -FCP.  $\square$

Similarly, we have

**Theorem 4.** *Suppose that the primal learner converges to a grammar  $G$  on  $L_*$ . Then  $L(G) = L_*$  and  $G$  has the very weak  $k$ -FKP.*

The following theorems are clear from the existing proof of correctness of the dual and primal learners (see, e.g., [4]).

**Theorem 5.** *Let  $G$  be a CFG each of whose productions has at most  $r$  non-terminals on the right-hand side. If  $G$  has the very weak  $k$ -FCP, then the dual learner converges to a grammar for  $L(G)$  on  $L(G)$ .*

**Theorem 6.** *Let  $G$  be a CFG each of whose productions has at most  $r$  nonterminals on the right-hand side. If  $G$  has the very weak  $k$ -FKP, then the primal learner converges to a grammar for  $L(G)$  on  $L(G)$ .*

## 5 The Strong vs. Weak Finite Context and Kernel Properties

We write  $x^R$  for the reversal of a string  $x$ , and  $|x|_a$  for the number of occurrences of a symbol  $a$  in  $x$ . Let

$$\begin{aligned}\Sigma &= \{a, b, c, d, e, \#, \$\}, \\ L &= L_1 \cup L_2 \cup L_3, \\ L_1 &= \{w_1 \# w_2 \# \dots \# w_n \$ w_n^R \dots w_2^R w_1^R \mid n \geq 1, w_1, \dots, w_n \in \{a, b\}^*\}, \\ L_2 &= \{w y c^i d^j e^k z \mid w, z \in \{a, b\}^*, y \in (\#\{a, b\}^*)^*, i, j \geq 0, |w|_a \geq |w|_b\}, \\ L_3 &= \{w y c^i d^j e^k z \mid w, z \in \{a, b\}^*, y \in (\#\{a, b\}^*)^*, i, j \geq 0, |w|_a \leq |w|_b\}.\end{aligned}$$

**Lemma 7.** *Every CFG  $G$  for  $L$  has a nonterminal  $E$  such that  $L(G, E)$  is not a closed set relative to  $L$ .*

*Proof.* Let  $G$  be a CFG for  $L$ . By applying Ogden's [7] lemma<sup>3</sup> to a derivation tree of a sufficiently long string in  $L_1$  of the form  $a^p b^p \# a^p \$ a^p b^p a^p$ , we obtain

$$\begin{aligned}S_1 &\Rightarrow_G^+ a^{m_1} A a^{l_1}, & A &\Rightarrow_G^+ a^{n_1} A a^{n_1}, & A &\Rightarrow_G^+ a^{m_2} b^{m_3} B b^{l_3} a^{l_2}, \\ B &\Rightarrow_G^+ b^{n_2} B b^{n_2}, & B &\Rightarrow_G^+ b^{m_4} \# a^{m_5} D a^{l_5} b^{l_4}, & D &\Rightarrow_G^+ a^{m_6} \$ a^{l_6},\end{aligned}$$

for some  $n_1, n_2, m_1, m_2, m_3, m_4, m_5, m_6 \geq 1, l_1, l_2, l_3, l_4, l_5, l_6 \geq 0$  such that  $m_1 + n_1 + m_2 = m_3 + n_2 + m_4 = m_5 + m_6 = l_1 + n_1 + l_2 = l_3 + n_2 + l_4 = l_5 + l_6 = p$ , where  $S_1$  is an initial nonterminal. We show that  $L(G, D)$  is not a closed set.

Let  $(u, v) \in L(G, D)^\triangleright$ . Then  $u a^{m_6} \$ a^{l_6} v \in L$ . Since  $y \$ z \in L$  implies  $y c^i d^j e^k z \in L$  for every  $y, z \in \Sigma^*$  and  $i \geq 0$ , we have  $u a^{m_6} c^i d^j e^k a^{l_6} v \in L$  for every  $i \geq 0$ . This shows

$$\{a^{m_6} c^i d^j e^k a^{l_6} \mid i \geq 0\} \subseteq L(G, D)^\triangleright. \quad (2)$$

On the other hand, since

$$S_1 \Rightarrow_G^* a^{m_1} (a^{n_1})^i a^{m_2} b^{m_3} (b^{n_2})^j b^{m_4} \# a^{m_5} D a^{l_5} b^{l_4} (b^{n_2})^j b^{l_3} a^{l_2} (a^{n_1})^i a^{l_1}$$

for all  $i, j \geq 0$ , there are  $w, w', z, z' \in \{a, b\}^*$  such that  $|w|_a > |w|_b, |w'|_a < |w'|_b$ , and

$$S_1 \Rightarrow_G^* w \# a^{m_5} D z, \quad S_1 \Rightarrow_G^* w' \# a^{m_5} D z', \quad (3)$$

<sup>3</sup> It is clear from Ogden's proof that the lemma is really about one particular derivation tree of a context-free grammar. If  $p$  is the constant of Ogden's lemma for  $G$ , we obtain the required decomposition of the derivation tree by first marking the initial  $a^p$ , then the  $b^p$  preceding  $\#$ , and then the  $a^p$  immediately following  $\#$ .

Now suppose  $a^{m_6}c^i d^j e^k a^{l_6} \in L(G, D)^{\triangleright\triangleleft}$ . Since (3) implies  $(w\#a^{m_5}, z) \in L(G, D)^{\triangleright}$  and  $(w'\#a^{m_5}, z') \in L(G, D)^{\triangleright}$ , we must have  $w\#a^{m_5}a^{m_6}c^i d^j e^k a^{l_6}z \in L_2$  and  $w'\#a^{m_5}a^{m_6}c^i d^j e^k a^{l_6}z \in L_3$ . It follows that

$$a^{m_6}c^i d^j e^k a^{l_6} \in L(G, D)^{\triangleright\triangleleft} \text{ only if } i = j = k. \quad (4)$$

By (2) and (4),  $L(G, D)^{\triangleright\triangleleft} \cap a^{m_6}c^*d^*e^*a^{l_6} = \{a^{m_6}c^i d^i e^i a^{l_6} \mid i \geq 0\}$ , which implies that  $L(G, D)^{\triangleright\triangleleft}$  is not context-free. Therefore,  $L(G, D) \neq L(G, D)^{\triangleright\triangleleft}$  and  $L(G, D)$  is not a closed set.  $\square$

The above lemma implies that  $L$  has no CFG that has either the strong FCP or the strong FKP.

**Lemma 8.** *There is a CFG for  $L$  that has both the weak 2-FCP and the weak 2-FKP.*

*Proof.* Let  $G$  be the following CFG, where  $S_1, S_2, S_3$  are the initial nonterminals.

$$\begin{aligned} S_1 &\rightarrow \$ \mid aS_1a \mid bS_1b \mid \#S_1, & C &\rightarrow \varepsilon \mid cC, \\ Q &\rightarrow \varepsilon \mid aQbQ \mid bQaQ, & J &\rightarrow \varepsilon \mid dJe, \\ F &\rightarrow Q\# \mid Fa \mid Fb \mid F\#, & S_2 &\rightarrow HE \mid FS_2 \mid QS_2 \mid aS_2 \mid S_2a \mid S_2b, \\ H &\rightarrow \varepsilon \mid cHd, & S_3 &\rightarrow CJ \mid FS_3 \mid QS_3 \mid bS_3 \mid S_3a \mid S_3b, \\ E &\rightarrow \varepsilon \mid Ee, \end{aligned}$$

We have

$$\begin{aligned} L(G, S_1) &= L_1, \\ L(G, S_1)^{\triangleright} &= \{(w_1\#w_2\#\dots\#w_n, w_n^R \dots w_2^R w_1^R) \mid n \geq 1, w_1, \dots, w_n \in \{a, b\}^*\}, \\ L(G, S_1)^{\triangleright\triangleleft} &= L_1 \cup \{yc^i d^i e^i z \mid y \in \{a, b, \#\}^*, z \in \{a, b\}^*, i \geq 0\} \\ &= \{(a\#, a), (b\#, b)\}^{\triangleleft} = \{\$\}^{\triangleright\triangleleft}, \\ L(G, Q) &= \{w \in \{a, b\}^* \mid |w|_a = |w|_b\} = \{(\varepsilon, \#cd), (a, b\#de)\}^{\triangleleft} = \{\varepsilon, ab\}^{\triangleright\triangleleft}, \\ L(G, F) &= \{w\#y \mid w \in \{a, b\}^*, |w|_a = |w|_b, y \in \{a, b, \#\}^*\} \\ &= \{(\varepsilon, cd), (\varepsilon, ade)\}^{\triangleleft} = \{\#, ab\# \}^{\triangleright\triangleleft}, \\ L(G, H) &= \{c^i d^i \mid i \geq 0\} = \{(a\#c, d)\}^{\triangleleft} = \{\varepsilon, cd\}^{\triangleright\triangleleft}, \\ L(G, E) &= e^* = \{(a\#cd, e)\}^{\triangleleft} = \{\varepsilon, e\}^{\triangleright\triangleleft}, \\ L(G, C) &= c^* = \{(b\#c, de)\}^{\triangleleft} = \{\varepsilon, c\}^{\triangleright\triangleleft}, \\ L(G, J) &= \{d^i e^i \mid i \geq 0\} = \{(b\#d, e)\}^{\triangleleft} = \{\varepsilon, de\}^{\triangleright\triangleleft}, \\ L(G, S_2) &= L_2, \\ L(G, S_2)^{\triangleright} &= \{(wy, z) \mid w, z \in \{a, b\}^*, y \in (\#\{a, b\}^*)^*, |w|_a \geq |w|_b\}, \\ L(G, S_2)^{\triangleright\triangleleft} &= L_2 \cup \{vc^i d^i e^i z \mid v \in \{a, b, \#\}^*, z \in \{a, b\}^*, i \geq 0\} \\ &= \{(\varepsilon, \varepsilon), (a\#, b)\}^{\triangleleft} = \{cda\}^{\triangleright\triangleleft}, \\ L(G, S_3) &= L_3, \end{aligned}$$

$$\begin{aligned}
L(G, S_3)^\triangleright &= \{(wy, z) \mid w, z \in \{a, b\}^*, y \in (\#\{a, b\}^*)^*, |w|_a \leq |w|_b\}, \\
L(G, S_3)^{\triangleright\triangleleft} &= L_3 \cup \{vc^i d^j e^k z \mid v \in \{a, b, \#\}^*, z \in \{a, b\}^*, i \geq 0\} \\
&= \{(\varepsilon, \varepsilon), (b\#, a)\}^{\triangleleft} = \{\#de\}^{\triangleright\triangleleft}.
\end{aligned}$$

This shows that  $G$  has both the weak 2-FCP and the weak 2-FKP.  $\square$

**Theorem 9.** *There is a language that has a CFG with both the weak 2-FCP and the weak 2-FKP but has no CFG with either the strong FCP or the strong FKP.*

## 6 The Weak vs. Very Weak Finite Context and Kernel Properties

**Proposition 10.** *If a language  $L$  has a CFG with the very weak  $k$ -FKP, then  $L$  has a CFG with the weak  $k$ -FKP.*

*Proof.* Let  $G = (N, \Sigma, P, I)$  be a CFG and let  $L = L(G)$ . Suppose that  $K_A \subseteq \Sigma^*$  is a finite set for each  $A \in N$  such that  $|K_A| \leq k$  and  $(K_A^{|L\rangle\langle L|})_{A \in N}$  is an SPP for  $G$ . Let  $G' = (N, \Sigma, P', I)$ , where  $P' = P \cup \{A \rightarrow w \mid A \in N, w \in K_A\}$ . Then  $(L(G', A))_{A \in N}$  is the least pre-fixed point  $(X_A)_{A \in N}$  of  $G$  such that  $K_A \subseteq X_A$ . Since  $K_A \subseteq K_A^{|L\rangle\langle L|}$ , this implies  $L(G', A) \subseteq K_A^{|L\rangle\langle L|}$  for each  $A \in N$ . As a consequence,  $L(G') \subseteq \bigcup_{A \in I} K_A^{|L\rangle\langle L|} = L$ . Clearly,  $L = L(G) \subseteq L(G')$ , so  $L(G') = L$ . Since  $K_A \subseteq L(G', A)$ , we get  $K_A^{|L\rangle\langle L|} \subseteq L(G', A)^{|L\rangle\langle L|}$  and hence  $K_A^{|L\rangle\langle L|} = L(G', A)^{|L\rangle\langle L|}$ . This shows that  $G'$  has the weak  $k$ -FKP.  $\square$

The proof of Proposition 10 shows that Theorem 4 can be strengthened to

**Corollary 11.** *If the primal learner converges to a grammar  $G$  on  $L_*$ , then  $L(G) = L_*$  and  $G$  has the weak FKP.*

Let

$$\begin{aligned}
\Sigma' &= \{a, b, c, d, e, f, \#, \$, \%, \}, \\
L' &= L_1 \cup L_2 \cup L_3 \cup L_4 \cup L_5, \\
L_4 &= \{v\$zc^k\%f^l \mid v \in \{a, b, \#\}^*, z \in \{a, b\}^*, k, l \geq 0\}, \\
L_5 &= \{vc^i d^j e^k zc^k\%f^l \mid v \in \{a, b, \#\}^*, z \in \{a, b\}^*, i, j, k, l \geq 0, j \neq k\},
\end{aligned}$$

where  $L_1, L_2, L_3$  are as defined in Section 5.

**Lemma 12.** *There is a CFG for  $L'$  that has the very weak 2-FCP.*

*Proof.* Let  $G'$  be the extension of the grammar  $G$  in the proof of Lemma 8 with the following additional productions:

$$\begin{aligned}
T &\rightarrow \varepsilon \mid Ta \mid Tb, & S_{4,5} &\rightarrow \$TC\% \mid HU\% \mid aS_{4,5} \mid bS_{4,5} \mid \#S_{4,5} \mid S_{4,5}f. \\
U &\rightarrow EeT \mid TcC \mid eUc,
\end{aligned}$$

The nonterminals  $T, U, S_{4,5}$  are new and  $S_{4,5}$  is an initial nonterminal. All the old nonterminals except  $S_1$  continue to be (weakly) characterized by the same sets of contexts as before, but we now have

$$\begin{aligned} L(G', S_1)^\triangleright &= \{ (w_1 \# w_2 \# \dots \# w_n, w_n^R \dots w_2^R w_1^R) \mid n \geq 1, w_1, \dots, w_n \in \{a, b\}^* \} \\ &\quad \cup \{ (v, zc^k\%f^l) \mid v \in \{a, b, \#\}^*, z \in \{a, b\}^*, k, l \geq 0 \}, \\ L(G', S_1)^\triangleleft &= L(G', S_1) = L_1, \end{aligned}$$

where of course  $\triangleright, \triangleleft$  are now relative to  $L'$ . For the new nonterminals, we have

$$\begin{aligned} L(G', T) &= \{a, b\}^* = \{(\$ , a\%)\}^\triangleleft, \\ L(G', U) &= \{e^j zc^k \mid z \in \{a, b\}^*, j, k \geq 0, j \neq k\} = \{(e, c\%)\}^\triangleleft, \\ L(G', S_{4,5}) &= L_4 \cup L_5 = \{(\varepsilon, f)\}^\triangleleft. \end{aligned}$$

Now if we let

$$\begin{aligned} X_{S_1} &= C(G', S_1)^\triangleleft \\ &= \{ (w_1 \# w_2 \# \dots \# w_n, w_n^R \dots w_2^R w_1^R) \mid n \geq 1, w_1, w_2, \dots, w_n \in \{a, b\}^* \}^\triangleleft \\ &= L_1 \cup \{vc^i d^i e^i z \mid v \in \{a, b, \#\}^*, z \in \{a, b\}^*, i \geq 0\}, \end{aligned}$$

then we have  $X_{S_1} \supseteq \{\$\} \cup aX_{S_1}a \cup bX_{S_1}b \cup \#X_{S_1}$ , so  $X_{S_1}$  together with (the closures of) the languages of the other nonterminals constitute an SPP for  $G'$ . Since  $X_{S_1} = \{(a\#, a), (b\#, b)\}^\triangleleft$ , this shows that  $G'$  has the very weak 2-FCP.  $\square$

**Lemma 13.** *There is no CFG for  $L'$  that has the weak FCP.*

*Proof.* Let  $G$  be any CFG for  $L'$ . As in the proof of Lemma 7, we obtain

$$\begin{aligned} S &\Rightarrow_G^* a^{m_1} (a^{n_1})^i a^{m_2} b^{m_3} (b^{n_2})^j b^{m_4} \# a^{m_5} D a^{l_5} b^{l_4} (b^{n_2})^j b^{l_3} a^{l_2} (a^{n_1})^i a^{l_1}, \\ D &\Rightarrow_G^+ a^{m_6} \$ a^{l_6}, \end{aligned}$$

for every  $i$  and  $j$ , where  $S$  is an initial nonterminal,  $D$  is a nonterminal, and  $n_1, n_2 \geq 1$ . It is easy to see that

$$\begin{aligned} L(G, D) &\subseteq \{a, b, \#\}^* (\{\$\} \cup \{c^i d^i e^i \mid i \geq 0\}) \{a, b\}^*, \\ L(G, D)^\triangleright &\subseteq \{a^{m_6} \$ a^{l_6}\}^\triangleright \\ &\subseteq \{a, b, \#\}^* \times \{a, b\}^* (\{\varepsilon\} \cup c^* \% f^*). \end{aligned}$$

Since  $L(G, D)$  is context-free, there must be a  $k_1$  such that

$$L(G, D) \cap \{a, b, \#\}^* c^i d^i e^i \{a, b\}^* = \emptyset \text{ for all } i > k_1.$$

It follows that  $\{(\varepsilon, c^i\%) \mid i > k_1\} \subseteq L(G, D)^\triangleright$  and so

$$L(G, D)^\triangleright \cap \{a, b, \#\}^* c^i d^i e^i \{a, b\}^* = \emptyset \text{ for all } i > k_1.$$

However, for any finite set  $W \subseteq L(G, D)^\triangleright$ , there is a  $k_2$  such that  $W \cap (\{a, b, \#\}^* \times \{a, b\}^* c^i \% f^*) = \emptyset$  for all  $i > k_2$ , which implies  $\{c^i d^i e^i \mid i > k_2\} \subseteq W^\triangleleft$ . It follows that  $L(G, D)^\triangleright \neq W^\triangleleft$  for any finite  $W$ .  $\square$

**Theorem 14.** *There is a language that has a CFG with the very weak 2-FCP but has no CFG with the weak FCP.*

## 7 Conclusion

The basic idea of distributional learning has been to let a finite set of contexts/strings determine (the distributions of) the strings derived from each nonterminal. We have shown that while this is indeed a necessary condition for the primal learner, the dual learner may succeed in the absence of such a characterizing finite set of contexts.

## References

1. Aho, A.V., Ullman, J.D.: *The Theory of Parsing, Translation, and Compiling*, vol. I. Prentice-Hall, Englewood Cliffs, N.J. (1972)
2. Clark, A.: Learning context free grammars with the syntactic concept lattice. In: Sempere, J.M., García, P. (eds.) *Grammatical Inference: Theoretical Results and Applications*, 10th International Colloquium, ICGI 2010. pp. 38–51. *Lecture Notes in Artificial Intelligence*, Springer, Berlin (2010)
3. Clark, A.: The syntactic concept lattice: Another algebraic theory of the context-free languages? *Journal of Logic and Computation* 25(5), 1203–1229 (2015), first published online: July 30, 2013
4. Clark, A., Kanazawa, M., Kobele, G.M., Yoshinaka, R.: Distributional learning of some nonlinear tree grammars. *Fundamenta Informaticae* 146(4), 339–377 (2016)
5. Clark, A., Yoshinaka, R.: Distributional learning of context-free and multiple context-free grammars. In: Heinz, J., Sempere, J.M. (eds.) *Topics in Grammatical Inference*, pp. 143–172. Springer, Berlin (2016)
6. Leiß, H.: Learning context free grammars with the finite context property: A correction of A. Clark’s algorithm. In: Morrill, G., Muskens, R., Osswald, R., Richter, F. (eds.) *Formal Grammar 2014*. pp. 121–137. *Lecture Notes in Computer Science*, Springer, Berlin (2014)
7. Ogden, W.: A helpful result for proving inherent ambiguity. *Mathematical Systems Theory* 2(3), 191–194 (1968)
8. Yoshinaka, R.: Towards dual approaches for learning context-free grammars based on syntactic concept lattices. In: Mauri, G., Leporati, A. (eds.) *Developments in Language Theory*, 15th International Conference, DLT 2011. pp. 429–440. *Lecture Notes in Computer Science*, Springer, Berlin (2011)
9. Yoshinaka, R.: General perspectives on distributionally learnable classes. In: Kuhlmann, M., Kanazawa, M., Kobele, G.M. (eds.) *Proceedings of the 14th Meeting on the Mathematics of Language*. pp. 87–98. Association for Computational Linguistics, Stroudsburg, PA (2015)
10. Yoshinaka, R.: Learning conjunctive grammars and contextual binary feature grammars. In: Dediu, A.H., Formenti, E., Martín-Vide, C., Truthe, B. (eds.) *Language and Automata Theory and Applications*, LATA 2015. pp. 623–635. *Lecture Notes in Computer Science*, Springer, Berlin (2015)
11. Yoshinaka, R., Kanazawa, M.: Distributional learning of abstract categorial grammars. In: Pogodalla, S., Prost, J.P. (eds.) *Logical Aspects of Computational Linguistics*, LACL 2011. pp. 251–266. *Lecture Notes in Computer Science*, Springer, Berlin (2011)