An Ambiguity Hierarchy of Weighted Context-Free Grammars

Yusuke Inoue(✉), Kenji Hashimoto, and Hiroyuki Seki

Graduate School of Informatics, Nagoya University, Nagoya, Japan
{y-inoue,seki}@sqlab.jp, k-hasimt@i.nagoya-u.ac.jp

Abstract. Weighted context-free grammar (WCFG) is a quantitative extension of context-free grammar (CFG). It is known that unambiguous weighted automata (WA), finitely-ambiguous WA, polynomially-ambiguous WA and general WA over the tropical semiring have different expressive powers. We prove that there exists a similar ambiguity hierarchy of WCFG over the tropical semiring, using an extended Ogden's lemma. Furthermore, we show that the hierarchy we proved is different from the known ambiguity hierarchy of unweighted CFG.

Keywords: Weighted context-free grammar · Ambiguity · Pumping lemma

1 Introduction

Weighted context-free grammar (WCFG) is a quantitative extension of context-free grammar (CFG). WCFG originates from the study of algebraic formal series by Chomsky and Schützenberger [2]. Since then, mathematical properties of WCFG and the formal series (or functions) defined by WCFG have been extensively studied. There are various applications of WCFG to real-world problems such as parsing natural language sentences and biological sequence analysis [4]. In some applications, weights correspond to probabilities, which are useful for selecting better estimations of the hidden structure from experimental or observable data. However, it is not yet very clear whether and how a hierarchy in terms of the expressive power is induced in the class of context-free languages by introducing weights to CFG.

In general, a weighted model (automaton, grammar, etc.) is defined with a semiring, and each model defines a function that maps a word to an element of the semiring, instead of a language. When the semiring is positive, the support of the function defined by a weighted model naturally corresponds to the language generated by the unweighted counterpart of the model, where the support is a homomorphism from the semiring to Boolean semiring $\{0, 1\}$.

The expressive power of weighted automata (WA) has been studied in the literature. In particular, it is known that unambiguous WA, finitely-ambiguous WA, polynomially-ambiguous WA and general WA over the tropical semiring have different expressive powers [1,7]. Unambiguous WA (resp. finitely-ambiguous WA, polynomially-ambiguous WA) are WA such that the number

P. Caron and L. Mignot (Eds.): CIAA 2022, LNCS 13266, pp. 238–250, 2022.
https://doi.org/10.1007/978-3-031-07469-1_19

of accepting runs is bounded by one (resp. by a constant, by a polynomial in the size of an input) for any input. Similar results are known for weighted tree automata over the tropical semiring [6] although the tree languages proved to be in the gaps between the adjacent two layers are essentially the same as those in [1,7]. For an (unweighted) finite automaton (FA), the ambiguity does not affect the expressive power since the determinization is possible for nondeterministic FA and a deterministic FA is apparently unambiguous. Therefore, the above mentioned results on WA indicate that the strict ambiguity hierarchy is caused by introducing weights. On the other hand, the ambiguity already increases the expressive power for unweighted CFG because there exist inherently ambiguous CFG [8]. In fact, it is shown that unambiguous CFG, finitely-ambiguous CFG, polynomially-ambiguous CFG and general CFG have different expressive powers [9].

In this paper, we study an ambiguity hierarchy of WCFG over the tropical semiring where the ambiguity of a word w in a WCFG G means the number of distinct parse trees of w in G. We show that there is a strict ambiguity hierarchy of WCFG over the tropical semiring caused by introducing weights. Specifically, we prove that there exist functions $f_{\mathrm{EX2}}, f_{\mathrm{EX3}}, f_{\mathrm{EX4}} \in$ U-CF such that $f_{\mathrm{EX2}} \in$ FA-WCF $\setminus$ U-WCF, $f_{\mathrm{EX3}} \in$ PA-WCF $\setminus$ FA-WCF and $f_{\mathrm{EX4}} \in$ WCF $\setminus$ PA-WCF. U-CF is the class of functions defined by WCFG over the tropical semiring whose supports coincide with the languages defined by unambiguous CFG, i.e., U-CF corresponds to the class of unambiguous context-free languages. U-WCF, FA-WCF, PA-WCF and WCF are the classes of functions defined by unambiguous WCFG, finitely-ambiguous WCFG, polynomially-ambiguous WCFG and general WCFG over the tropical semiring, respectively. That is, functions f_{EX2}, f_{EX3} and f_{EX4} exist in the gaps caused by introducing weights (see Fig. 1).

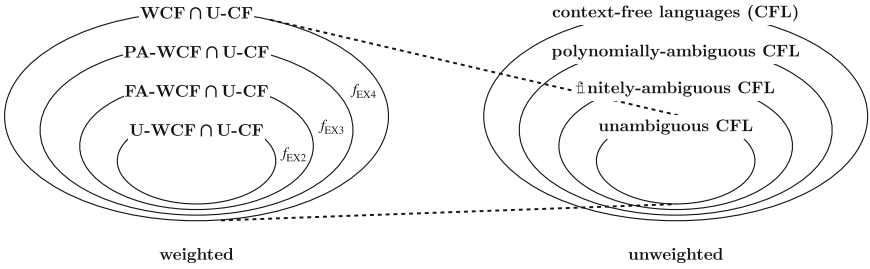

Fig. 1. The ambiguity hierarchy caused by introducing weights

Deciding the expressive power of weighted models is more difficult than that of unweighted ones. For unweighted automata (resp. grammars), we only need to check the existence of an accepting run (resp. a parse tree). For weighted automata (resp. grammars), we have to consider all accepting runs (resp. parse trees) to compute the weight of a given word because the weight of a word is defined by the semiring sum of the weights of all accepting runs (resp. parse

trees) of the word. For example, we have to find the minimum weight among all accepting runs when we compute the function value defined by WA over the tropical semiring. (Note that the sum in the tropical semiring means the minimum.) This difficulty is more remarkable for WCFG than for WA. This is because, the ambiguity for WA is caused by only the choice of a state, while the ambiguity for WCFG is also caused by the shape of a parse tree. Therefore, the expressive power of WCFG cannot be determined by a simple iteration property. For these reasons, we cannot show a strict ambiguity hierarchy of WCFG by a straightforward extension of the discussion on the ambiguity hierarchy for WA. To overcome this problem, we focus on the functions defined by WCFG that assign non-zero weights only to the words having specific form such as palindromes and well-nested parentheses (Dyck words).

In Sect. 2, we introduce semiring, weighted context-free grammar and weight function. Furthermore, we show some examples of functions defined by WCFG (Examples 1 to 5). These functions will be used to prove the strict hierarchy in Sect. 4. In Sect. 3, we show a pumping lemma for CFG, which is helpful for proving the hierarchy (Lemma 3). The lemma is an extension of the theorem for CFG known as Ogden's lemma. In Sect. 4, we prove that functions f_{EX2}, f_{EX3} and f_{EX4} defined in Sect. 2 lie in the gaps caused by introducing weights (Theorems 1, 2 and 3), and as a corollary of them, we show the strict ambiguity hierarchy of WCFG (Corollary 1).

2 Preliminaries

Let $\mathbb{N}$ be the set of all non-negative integers. The cardinality of a set X is denoted by $|X|$. Let Σ be a (finite) alphabet. For a word $w \in \Sigma^*$ and a letter $a \in \Sigma$, the length of w and the number of occurrences of a in w are denoted by $|w|$ and $|w|_a$, respectively. The empty word is denoted by ε, i.e., $|\varepsilon| = 0$. Let w^R be the reversal of w. For example, $(aab)^R = baa$. We say that $w' \in \Sigma^*$ is an (even) palindrome if there exists a word $w \in \Sigma^*$ such that $w' = ww^R$.

2.1 Semirings

A *semiring* $(\mathbb{S}, \oplus, \odot, \mathbb{0}, \mathbb{1})$ is an algebraic structure where

- $(\mathbb{S}, \oplus, \mathbb{0})$ is a commutative monoid,
- $(\mathbb{S}, \odot, \mathbb{1})$ is a monoid,
- $\odot$ distributes over $\oplus$,
- $\mathbb{0}$ is the zero element of $\odot$.

A semiring $(\mathbb{S}, \oplus, \odot, \mathbb{0}, \mathbb{1})$ is called a commutative semiring if $(\mathbb{S}, \odot, \mathbb{1})$ is also a commutative monoid. We abbreviate $(\mathbb{S}, \oplus, \odot, \mathbb{0}, \mathbb{1})$ as $\mathbb{S}$.

In this paper, we mainly consider the following two semirings : the *tropical semiring* $\mathbb{N}_{\min,+} = (\mathbb{N} \cup \{\infty\}, \min, +, \infty, 0)$ and *Boolean semiring* $\mathbb{B} = (\{0, 1\}, \vee, \wedge, 0, 1)$.

For a commutative semiring $\mathbb{S}$, we define the mapping $h_{\mathbb{S}} : \mathbb{S} \to \mathbb{B}$ as follows: $h_{\mathbb{S}}(x) = 0$ if $x = \mathbb{0}$, and $h_{\mathbb{S}}(x) = 1$ otherwise. A semiring $\mathbb{S}$ is said to be *positive* if $h_{\mathbb{S}} : \mathbb{S} \to \mathbb{B}$ is a semiring homomorphism, i.e., $h_{\mathbb{S}}(\mathbb{0}) = 0$, $h_{\mathbb{S}}(\mathbb{1}) = 1$, $h_{\mathbb{S}}(a \oplus b) = h_{\mathbb{S}}(a) \vee h_{\mathbb{S}}(b)$ and $h_{\mathbb{S}}(a \odot b) = h_{\mathbb{S}}(a) \wedge h_{\mathbb{S}}(b)$ for all $a, b \in \mathbb{S}$ [3]. Note that $\mathbb{N}_{\min,+}$ is a positive semiring.

2.2 Weighted Context-Free Grammars

Let $\mathbb{S}$ be a commutative semiring. A *weighted context-free grammar* (WCFG) over $\mathbb{S}$ is a tuple $G = (V, \Sigma, P, I, \mathrm{wt})$, where

- V is a finite set of *nonterminals*, and $I \in V$ is the *initial* symbol,
- Σ is a finite set of *terminals*, disjoint from V,
- P is a set of *productions* of the form: $A \to \gamma$ where $A \in V$ and $\gamma \in (V \cup \Sigma)^*$,
- $\mathrm{wt} : P \to \mathbb{S} \setminus \{\mathbb{0}\}$ is a *weight function.*

We say that $(\alpha A \beta, \alpha\gamma\beta) \in ((V \cup \Sigma)^*)^2$ is a *direct derivation* if there exists a production $p = A \to \gamma \in P$, and we write $\alpha A \beta \Rightarrow \alpha\gamma\beta$ or $\alpha A \beta \overset{c}{\Rightarrow} \alpha\gamma\beta$ where $c = \mathrm{wt}(p)$. For a sequence of direct derivations $\rho : \alpha_0 \overset{c_1}{\Rightarrow} \alpha_1 \overset{c_2}{\Rightarrow} \cdots \overset{c_n}{\Rightarrow} \alpha_n$ $(n \geq 0)$, the weight of ρ is defined by $\mathrm{wt}(\rho) = c_1 \odot c_2 \odot \cdots \odot c_n$. We say that ρ is a *derivation*, and we write $\alpha_0 A_0 \beta_0 \Rightarrow^* \alpha_n A_n \beta_n$ or $\alpha_0 A_0 \beta_0 \overset{c}{\Rightarrow}{}^* \alpha_n A_n \beta_n$ where $c = \mathrm{wt}(\rho)$. If a derivation ρ_1 can be written as $\alpha \Rightarrow^* \alpha_1\gamma\beta_1 \Rightarrow^* \alpha_1\delta\beta_1 \Rightarrow^* \eta$ where $\rho_2 : \gamma \Rightarrow^* \delta$ is also a derivation, we say that ρ_2 is a *subderivation* of ρ_1. A derivation $\rho : \alpha_0 A_0 \beta_0 \overset{c_1}{\Rightarrow} \cdots \overset{c_n}{\Rightarrow} \alpha_n A_n \beta_n$ $(n \geq 0)$ is said to be a *leftmost derivation* if $\alpha_0, \cdots, \alpha_n \in \Sigma^*$. A leftmost derivation $\rho : I \overset{c}{\Rightarrow}{}^* w$ is said to be a *complete leftmost derivation* of w if $c \neq \mathbb{0}$ and $w \in \Sigma^*$. Note that for each word $w \in \Sigma^*$, complete leftmost derivations of w have a one-to-one correspondence with parse trees of w in the usual sense [5]. Therefore, we will call a complete leftmost derivation $\rho : I \Rightarrow^* w$ a parse tree of w. For a word $w \in \Sigma^*$, the weight of w is defined by $[\![G]\!](w) = \bigoplus_{T \in \mathrm{parse}(w)} \mathrm{wt}(T)$ where $\mathrm{parse}(w)$ is the set of parse trees of w. We say that $[\![G]\!] : \Sigma^* \to \mathbb{S}$ is the function defined by WCFG G over $\mathbb{S}$.

For a WCFG $G = (V, \Sigma, P, I, \mathrm{wt})$, we say that CFG $G'' = (V, \Sigma, P, I)$ is the underlying CFG of G. If $[\![G]\!](w) \neq 0$, then $w \in L(G'')$ where $L(G'')$ is the language generated by G'' in the standard definition. However, the converse direction does not always hold. For example, if there are two derivations T_1 and T_2 of w in G where $\mathrm{wt}(T_1) = 1$ and $\mathrm{wt}(T_2) = -1$, then $[\![G]\!](w) = 0$ over $(\mathbb{Z}, +, \times, 0, 1)$ while $w \in L(G'')$.

Assume that $\mathbb{S}$ is positive (see Sect. 2.1). For the function $f = [\![G]\!]$ defined by a WCFG $G = (V, \Sigma, P, I, \mathrm{wt})$ over $\mathbb{S}$, the *support* of f is defined by $\mathrm{supp}(f) = h_{\mathbb{S}} \circ f$. Then, $\mathrm{supp}(f)$ coincides with the function defined by WCFG $G' = (V, \Sigma, P, I, \mathrm{wt}')$ over $\mathbb{B}$ where $\mathrm{wt}'(p) = h_{\mathbb{S}}(\mathrm{wt}(p))$. Let $G'' = (V, \Sigma, P, I)$ be the underlying CFG of G. Since S is positive,

$$[\![G]\!](w) \neq \mathbb{0} \iff \mathrm{supp}([\![G]\!])(w) = 1 \iff w \in L(G'').$$

A WCFG G over $\mathbb{S}$ is *unambiguous* (U-WCFG) if $|\mathrm{parse}(w)| \leq 1$ for all $w \in \Sigma^*$. G is *finitely-ambiguous* (FA-WCFG) if there exists $m \in \mathbb{N}$ such that $|\mathrm{parse}(w)| \leq m$ for all $w \in \Sigma^*$. G is *polynomially-ambiguous* (PA-WCFG) if there exists a polynomial $p(\cdot)$ such that $|\mathrm{parse}(w)| \leq p(|w|)$ for all $w \in \Sigma^*$.

Fix a semiring $\mathbb{S}$ and assume that $\mathbb{S}$ is positive. We define U-WCF, FA-WCF, PA-WCF and WCF as the classes of functions defined by U-WCFG, FA-WCFG, PA-WCFG and WCFG over $\mathbb{S}$, respectively. Clearly, U-WCF $\subseteq$ FA-WCF $\subseteq$ PA-WCF $\subseteq$ WCF. Furthermore, we define U-CF $= \{f \mid \exists$ U-WCFG G over $\mathbb{B}.\, \mathrm{supp}(f) = [\![G]\!]\}$. That is, U-CF is the class of functions whose supports are defined by some U-WCFG over $\mathbb{B}$. In this paper, we fix the semiring $\mathbb{S}$ to $\mathbb{N}_{\min,+}$ when we refer to these classes of functions.

Example 1. Let $G_1 = (\{I\}, \{a, b\}, P, I, \mathrm{wt})$ where $P = \{$

$I \to aIa \mid bIb \quad$ (weight : 1),
$I \to \varepsilon \quad$ (weight : 0) $\}$.

G_1 is a WCFG over $\mathbb{N}_{\min,+}$ and the function f_{EX1} defined by G_1 is

$$f_{\mathrm{EX1}}(w') = \begin{cases} |w| & w' = ww^R \\ \infty & \text{otherwise} \end{cases}.$$

Clearly G_1 is unambiguous, and hence $f_{\mathrm{EX1}} \in$ U-WCF.

Example 2. Let $G_2 = (\{I, A, B\}, \{a, b\}, P, I, \mathrm{wt})$ where $P = \{$

$I \to A \mid B \quad$ (weight : 0),
$A \to aAa \quad$ (weight : 1), $\quad A \to bAb \mid \varepsilon \quad$ (weight : 0),
$B \to bBb \quad$ (weight : 1), $\quad B \to aBa \mid \varepsilon \quad$ (weight : 0) $\}$.

G_2 is a WCFG over $\mathbb{N}_{\min,+}$, and the function f_{EX2} defined by G_2 is

$$f_{\mathrm{EX2}}(w') = \begin{cases} \min\{|w|_a, |w|_b\} & w' = ww^R \\ \infty & \text{otherwise} \end{cases}.$$

G_2 is finitely-ambiguous because there are two parse trees of w' if w' is a palindrome. One of them counts the number of letter a using nonterminal A, and the other counts the number of letter b using nonterminal B. Hence, $f_{\mathrm{EX2}} \in$ FA-WCF.

Example 3. Let $G_3 = (\{A, B\}, \{a, b, \$\}, P, B, \mathrm{wt})$ where $P = \{$

$B \to aBa \mid A \mid \$\$ \quad$ (weight : 0), $\quad B \to bBb \quad$ (weight : 1),
$A \to bAb \mid \$\$ \quad$ (weight : 0), $\quad A \to aAa \quad$ (weight : 1) $\}$.

G_3 is a WCFG over $\mathbb{N}_{\min,+}$, and the function f_{EX3} defined by G_3 is

$$f_{\mathrm{EX3}}(w') = \begin{cases} \min_{0 \leq i \leq n}\{|a_1 \cdots a_i|_b + |a_{i+1} \cdots a_n|_a\} & w' = ww^R,\ w = a_1 a_2 \cdots a_n\$ \\ \infty & \text{otherwise} \end{cases}.$$

For a palindrome $w' = ww^R$, G_3 counts the number of letter b using nonterminal B, and counts the number of letter a using nonterminal A. G_3 has a choice when to start counting a. Hence, G_3 is polynomially-ambiguous and $f_{\mathrm{EX3}} \in$ PA-WCF.

Example 4. Let $G_4 = (\{I, A, B\}, \{a, b, \#, \$\}, P, I, \mathrm{wt})$ where $P = \{$

$$I \to A \mid B \mid \$\$ \quad (\text{weight} : 0),$$
$$A \to aAa \quad (\text{weight} : 1), \quad A \to bAb \mid \#I\# \quad (\text{weight} : 0),$$
$$B \to bBb \quad (\text{weight} : 1), \quad B \to aBa \mid \#I\# \quad (\text{weight} : 0) \quad \}.$$

G_4 is a WCFG over $\mathbb{N}_{\min,+}$ and the function f_{EX4} defined by G_4 is

$$f_{\mathrm{EX4}}(w') = \begin{cases} \sum_{1 \le i \le n} \min\{|w_i|_a, |w_i|_b\} & w' = ww^R,\ w = w_1\#w_2\#\cdots w_n\#\$\ , \\ \infty & \text{otherwise}\ . \end{cases}$$

For a palindrome $w' = ww^R$, G_4 counts the number of letter a or letter b in w_i. For each i $(1 \le i \le n)$, G_4 has a choice whether to count the number of a in w_i using nonterminal A or to count the number of b in w_i using nonterminal B. Hence, G_4 is not polynomially-ambiguous.

Example 5. Let $G_5 = (\{I\}, \{a, b\}, P, I, \mathrm{wt})$ where $P = \{I \to aIa \mid bIb \mid \varepsilon\}$ and $\mathrm{wt}(p) = 1$ for all $p \in P$. G_5 is a WCFG over $\mathbb{B}$ and the underlying CFG of G_5 is $G'_5 = (\{I\}, \{a, b\}, P, I)$. The function f_{EX5} defined by G_5 satisfies $f_{\mathrm{EX5}}(w') = 1$ iff $w' = ww^R$. Furthermore, $\mathrm{supp}(f_{\mathrm{EX1}}) = \mathrm{supp}(f_{\mathrm{EX2}}) = f_{\mathrm{EX5}}$. Clearly G_5 is unambiguous, and hence $f_{\mathrm{EX1}}, f_{\mathrm{EX2}}, f_{\mathrm{EX5}} \in$ U-CF. We can also show that $f_{\mathrm{EX3}}, f_{\mathrm{EX4}} \in$ U-CF by considering variants of G_5.

3 An Extended Ogden's Lemma

In this section, we give an extension of Ogden's lemma, which is useful for proving the main results of this paper. We first review Ogden's lemma. The original Ogden's lemma in [8] is a statement for a word w, but we slightly extend it to a statement for a word w and a parse tree T of w. It is clear from the proof of Ogden's lemma in [8] that this extension also holds.

Lemma 1 (Ogden's Lemma [8]). *For each CFG $G = (V, \Sigma, P, I)$, there exists a constant $N \in \mathbb{N}$ that satisfies the following condition :*

> *Let w be any word in $L(G)$ and T be any parse tree of w in G. For any way to mark at least N positions in w as distinguished, there exist $A \in V$ and $u, v, x, y, z \in \Sigma^*$ such that*
> - *T can be represented as $I \Rightarrow^* uAz \Rightarrow^* uvAyz \Rightarrow^* uvxyz = w$,*
> - *x has at least one of the distinguished positions,*
> - *Either u and v both have distinguished positions, or y and z both have distinguished positions, and*
> - *vxy has at most N distinguished positions.*

□

We define the relation $\sqsubseteq_w \subseteq (\Sigma^*)^3 \times (\Sigma^*)^3$ as follows: for a word $w = uvx = \lambda\mu\nu \in \Sigma^*$, $(u, v, x) \sqsubseteq_w (\lambda, \mu, \nu)$ if there exist words $\lambda', \nu' \in \Sigma^*$ such that $\mu = \lambda' v \nu'$, $\lambda\lambda' = u$, $\nu'\nu = x$. If $(u, v, x) \sqsubseteq_w (\lambda, \mu, \nu)$ where the word w and the partitions $w = uvx = \lambda\mu\nu$ are clear or not relevant, we say that μ *contains* v.

Lemma 2. *Let $G = (V, \Sigma, P, I)$ be a CFG and L be the language defined by G. There exists a constant $N \in \mathbb{N}$ that satisfies the following condition :*

Let $w = \lambda\mu\nu \in \Sigma^*$ be any word such that $w \in L$ and $|\mu| \geq N$. For every parse tree T of w, there exist $A \in V$ and $u, v, x, y, z \in \Sigma^*$ such that T can be represented as

$$I \Rightarrow^* uAz \Rightarrow^* uvAyz \Rightarrow^* uvxyz = w \ ,$$

and the following (i) or (ii) holds.
(i) $1 \leq |v| < N$ and μ contains v, i.e., $(u, v, xyz) \sqsubseteq_w (\lambda, \mu, \nu)$.
(ii) $1 \leq |y| < N$ and μ contains y, i.e., $(uvx, y, z) \sqsubseteq_w (\lambda, \mu, \nu)$.

Proof. The above property can be obtained by applying Lemma 1, by letting all letters in μ be distinguished positions. □

Lemma 2 states that every word $w \in L$ having a sufficiently long subword μ can be divided as $w = uvxyz$ such that μ contains one of v and y. We call such a pair (v, y) a pump in w.

As stated in the next theorem, Lemma 2 can be generalized in such a way that if a word $w \in L$ has $2n$ long subwords $\mu_1, \ldots, \mu_{2n}$, then w has n pumps (v_i, y_i) $(1 \leq i \leq n)$ such that some n subwords out of $\mu_1, \ldots, \mu_{2n}$ either contains the left subwords v_i $(1 \leq i \leq n)$ or the right subwords y_i $(1 \leq i \leq n)$. This generalization is essential for proving the existence of a function not in FA-WCF (Theorem 2) and a function not in PA-WCF (Theorem 3).

Lemma 3. *Let $G = (V, \Sigma, P, I)$ be a CFG and L be the language generated by G. There exists a constant $N \in \mathbb{N}$ that satisfies the following condition :*

Let $w = \lambda_1 \cdot \mu_1 \cdot \lambda_2 \cdot \mu_2 \cdot \cdots \cdot \lambda_{2n} \cdot \mu_{2n} \cdot \lambda_{2n+1} \in \Sigma^*$ be any word such that $w \in L$ and $|\mu_1|, \ldots, |\mu_{2n}| \geq N$. For every parse tree T of w, there are subderivations $A_i \Rightarrow^* v_i A_i y_i$ of T where $A_i \in V$, $v_i, y_i \in \Sigma^*$ for each i $(1 \leq i \leq n)$ such that there exists a monotone injection $g : \{1, \ldots, n\} \to \{1, \ldots, 2n\}$ and the following (i) or (ii) holds.
(i) For each i $(1 \leq i \leq n)$, $1 \leq |v_i| < N$ and $\mu_{g(i)}$ contains v_i.
(ii) For each i $(1 \leq i \leq n)$, $1 \leq |y_i| < N$ and $\mu_{g(i)}$ contains y_i.

Proof. Let N be a constant in Lemma 2 and λ'_j, ν'_j be $\lambda'_j = \lambda_1\mu_1 \cdots \lambda_j$, $\nu'_j = \lambda_{j+1}\mu_{j+1} \cdots \lambda_{2n+1}$ for each j $(1 \leq j \leq 2n)$ (see Fig. 2). By applying Lemma 2 to $w = \lambda'_j \mu_j \nu'_j$ (note that $|\mu_j| \geq N$) and a parse tree of w, we obtain that there is a subderivation $A_j \Rightarrow^* v_j A_j y_j$ of T, and (i') μ_j contains v_j such that $1 \leq |v_j| < N$ or (ii') μ_j contains y_j such that $1 \leq |y_j| < N$. Since we have $2n$ subwords $\mu_1, \ldots, \mu_{2n}$ that do not pairwise overlap in w, there exist $j_1, j_2, \cdots, j_n$ $(1 \leq j_1 < j_2 < \cdots < j_n \leq 2n)$ such that the following (i) or (ii) holds.

(i) For each j_i $(1 \leq i \leq n)$, μ_{j_i} contains v_{j_i}.
(ii) For each j_i $(1 \leq i \leq n)$, μ_{j_i} contains y_{j_i}.

Let $A_i = A_{j_i}$, $v_i = v_{j_i}$, $y_i = y_{j_i}$ and define the injection g as $g(i) = j_i$, then the claim of the theorem holds. □

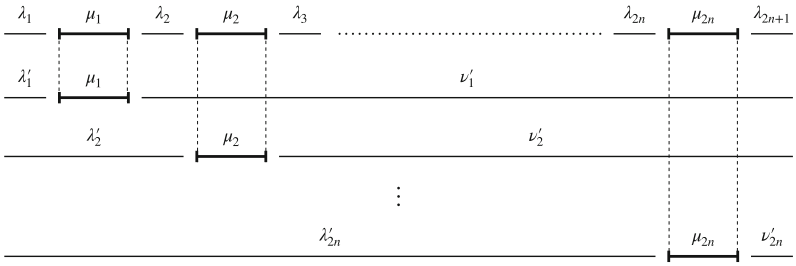

Fig. 2. Illustration for the proof of Lemma 3

4 An Ambiguity Hierarchy of WCFG over $\mathbb{N}_{\min,+}$

The purpose of this paper is to prove a strict ambiguity hierarchy caused by introducing weights. Namely, we would like to prove that there exists a function in (U-CF∩FA-WCF)\U-WCF (resp. a function in (U-CF∩PA-WCF)\FA-WCF, a function in (U-CF∩WCF)\PA-WCF). We use f_{EX2} (resp. f_{EX3}, f_{EX4}) as such a function that exists in the gap. We already know that $f_{\mathrm{EX2}} \in$ U-CF ∩ FA-WCF (resp. $f_{\mathrm{EX3}} \in$ U-CF ∩ PA-WCF, $f_{\mathrm{EX4}} \in$ U-CF ∩ WCF) by Example 2 (resp. Example 3, Example 4) and Example 5. Therefore, we just need to prove $f_{\mathrm{EX2}} \notin$ U-WCF (resp. $f_{\mathrm{EX3}} \notin$ FA-WCF, $f_{\mathrm{EX4}} \notin$ PA-WCF).

To prove them, we use Lemma 3. Note that $\mathbb{N}_{\min,+}$ is a positive semiring (see Sect. 2.1), and hence $[\![G]\!](w) \neq \infty$ iff $w \in L(G'')$ where G is a WCFG over $\mathbb{N}_{\min,+}$ and G'' is the underlying CFG of G. Therefore, Lemma 3 can be applied to WCFG over $\mathbb{N}_{\min,+}$, by regarding "Let $G = (V, \Sigma, P, I)$ be a CFG and L be the language generated by G" as "Let $G = (V, \Sigma, P, I, \mathrm{wt})$ be a WCFG over $\mathbb{N}_{\min,+}$ and f be the function defined by G" and "$w \in L$" as "$f(w) \neq \infty$".

Theorem 1. *$f_{\mathrm{EX2}} \notin$ U-WCF.*

Proof. We suppose that f_{EX2} can be defined by an unambiguous WCFG $G = (V, \Sigma, P, I, \mathrm{wt})$ and let N be a constant in Lemma 3. Consider the word $w = b^N a^{N+1} a^{N+1} b^N$. Clearly, w is a palindrome and $f_{\mathrm{EX2}}(w) = N$. Let T be a parse tree of w such that $\mathrm{wt}(T) = N$.

Let us apply Lemma 3 to w and T by letting $n = 1$ and $w = \lambda_1\mu_1\lambda_2\mu_2\lambda_3$ where $\lambda_1 = \lambda_3 = \varepsilon$, $\mu_1 = \mu_2 = b^N$ and $\lambda_2 = a^{N+1}a^{N+1}$. Then, T can be written as $I \Rightarrow^* uAz \underset{c}{\Rightarrow}^* uvAyz \Rightarrow^* uvxyz = w$ for some $A \in V$ and $u, v, x, y, z \in \Sigma^*$, and one of the following four conditions holds: (i-1) μ_1 contains v, or (i-2) μ_2 contains v, or (ii-1) μ_1 contains y, or (ii-2) μ_2 contains y (see Fig. 3). We examine these four cases.

The case (i-2) contradicts the definition of f_{EX2}. This is because $w_2 = uvvxyyz = b^N a^{N+1} a^{N+1} b^{N'}$ $(N' > N)$ has a parse tree whose weight is $N + c$ but $f_{\mathrm{EX2}}(uvvxyyz) = \infty$ since w_2 is not a palindrome. The case (ii-1) is not possible by a similar reason to (i-2).

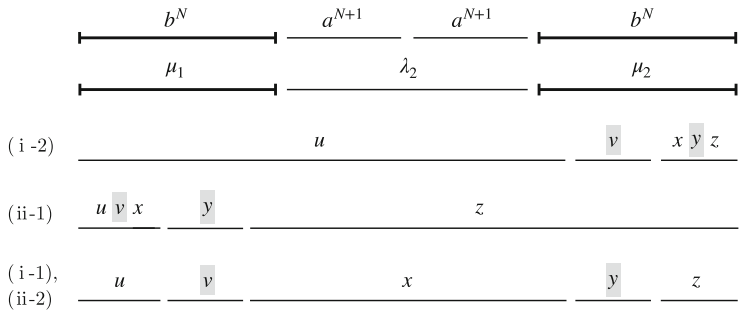

Fig. 3. Case analysis for the proof of Theorem 1

If (i-1) holds, it follows that $v = y = b^k$ $(1 \leq k < N)$ and μ_2 contains y because, by the definition of f_{EX2}, $f_{\mathrm{EX2}}(w) \neq \infty$ iff w is a palindrome. If (ii-2) holds, $v = y = b^k$ $(1 \leq k < N)$ and μ_1 contains v by the same reason as in the case (i-1). For these subcases (i-1) and (ii-2), consider the parse tree T' of $w' = uv^3xy^3z = b^{N+2k}a^{N+1}a^{N+1}b^{N+2k}$, which is constructed by pumping the subderivation $A \overset{c}{\Rightarrow}{}^* vAy$ in T twice. Apparently, $\mathrm{wt}(T') = N + 2c$. Because $k \geq 1$, $N + 1 = f_{\mathrm{EX2}}(w') \neq \mathrm{wt}(T')$ for any $c \in \mathbb{N}$. Hence, there exists a parse tree of w' whose weight is $N + 1$. Therefore, $|\mathrm{parse}(w')| \geq 2$, but it contradicts the assumption that G is unambiguous. □

Remark 1. We used a WCFG that generates palindromes to prove Theorem 1, but the technique can be applied to other WCFG. For example, we consider the following function f'_{EX2} defined by FA-WCFG :

$$f'_{\mathrm{EX2}}(w) = \begin{cases} \min\{|w|_{[}, |w|_{\langle}\} & w \in \mathrm{Dyck}([\,], \langle\,\rangle) \\ \infty & \text{otherwise} \end{cases}$$

where $\mathrm{Dyck}([\,], \langle\,\rangle)$ is Dyck language consisting of two types of brackets $[\,]$ and $\langle\,\rangle$. We can prove that f'_{EX2} is not in U-WCF using the word $\langle^N[^{N+1}]^{N+1}\rangle^N$ as well. For Theorems 2 and 3 below, we also use palindromes for simplicity.

Every non-empty (even) palindrome can be written as ww^R where $w = a_1^{n_1}a_2^{n_2}\cdots a_k^{n_k}$, $n_j \geq 1$ for each j $(1 \leq j \leq k)$ and $a_j \neq a_{j+1}$ for each j $(1 \leq j < k)$. We call each $a_j^{n_j}$ a *block* in w. We say that $a_j^{n_j}$ in w and $a_j^{n_j}$ in w^R forms a *symmetrical block pair* of ww^R.

To prove Theorems 2 and 3 below, we show a pumping lemma for CFL that contain only palindromes. Lemma 4 states that if a parse tree T of a palindrome with distinct central positions such as $w\$\$w^R$ where $w \in (\Sigma - \{\$\})^*$ has pumps (v_i, y_i), T must consist of only linear recursions of nonterminals and v_i, y_i are contained in a symmetrical block pair, respectively. This is a generalization of the case analysis in the proof of Theorem 1.

Lemma 4. *Let $G = (V, \Sigma, P, I)$ be a CFG that generates only palindromes, and Σ is divided as $\Sigma = \Gamma \cup \Delta \cup \Omega$ with Γ, Δ, Ω pairwise disjoint. There exists a constant $N \in \mathbb{N}$ that satisfies the following condition :*

Let $ww^R = \lambda_1 \cdot \mu_1 \cdot \lambda_2 \cdot \mu_2 \cdot \cdots \cdot \lambda_{2n} \cdot \mu_{2n} \cdot \lambda_{2n+1} \in L(G)$ where $|\mu_i| \geq N$, $\mu_i = \mu_{2n+1-i} \in a^*$ with some $a \in \Gamma$, $\lambda_{2n-i+2} = (\lambda_i)^R \in \Delta^*$ for every i $(1 \leq i \leq n)$ and $\lambda_{n+1} \in \Omega^+$ is a palindrome. For every parse tree T of ww^R, there are subderivations $A_i \Rightarrow^* v_i A_i y_i$ of T where $A_i \in V$, $v_i, y_i \in \Sigma^*$ for each i $(1 \leq i \leq n)$ such that
(1) $1 \leq |v_i| < N$ and μ_i contains v_i.
(2) $v_i = y_i$, and
(3) v_i and y_i are contained in a symmetrical block pair of ww^R.

Proof. Let N be a constant in Lemma 2. By applying Lemma 2 to the assumed CFG G in the same way as the proof in Lemma 3, there are subderivations $A_i \Rightarrow^* v_i A_i y_i$ of T where (i') μ_i contains v_i such that $1 \leq |v_i| < N$ or (ii') μ_i contains y_i such that $1 \leq |y_i| < N$, for each i $(1 \leq i \leq 2n)$.

If (ii') holds for some $i \leq n$, then T can be represented as $I \Rightarrow^* uA_i z\lambda_{n+1}z' \Rightarrow^* uv_i A_i y_i z\lambda_{n+1}z' \Rightarrow^* uv_i x y_i z\lambda_{n+1}z' = ww^R$ for some $u, x, z, z' \in \Sigma^*$. Note that $\lambda_{n+1} \in \Omega^*$, $uv_i x y_i z, z' \in (\Gamma \cup \Delta)^*$ and $y_i \neq \varepsilon$, contradicting the assumption that G generates only palindromes. Therefore, (i') holds for every i $(1 \leq i \leq n)$. That is, (1) $1 \leq |v_i| < N$ and μ_i contains v_i for every i $(1 \leq i \leq n)$. Furthermore, we can show the following in the same way as the proof of Theorem 1. Subderivations $A_i \Rightarrow^* v_i A_i y_i$ satisfy the conditions (2) and (3) for each i $(1 \leq i \leq n)$, otherwise, G can generate non palindromes by pumping (v_i, y_i).

Theorem 2. *$f_{\mathrm{EX3}} \notin$ FA-WCF.*

Proof. We suppose that f_{EX3} can be defined by a WCFG $G = (V, \Sigma, P, I, \mathrm{wt})$ such that there exists $m \in \mathbb{N}$ and $|\mathrm{parse}(w)| \leq m - 1$ for all $w \in \Sigma^*$. Let N be a constant in Lemma 4. For each ℓ $(1 \leq \ell \leq m)$, consider the word

$$w_\ell = \alpha_1\beta_1\alpha_2\beta_2 \cdots \alpha_m\beta_m\$\$\beta_{m+1}\alpha_{m+1} \cdots \beta_{2m-1}\alpha_{2m-1}\beta_{2m}\alpha_{2m}$$

where

$$(\alpha_j, \beta_j) = \begin{cases} (a^{N(m\cdot N!+1)},\, b^{N(m\cdot N!+1)}) & j = \ell, 2m-\ell+1\ , \\ (a^N, b^N) & \text{otherwise}\ , \end{cases}$$

for each j $(1 \leq j \leq 2m)$. Note that w_ℓ is a palindrome. We include long subwords $a^{N(m\cdot N!+1)}$ in w_ℓ by the following reason. Below we will show that there are pumps (a^{k_i}, a^{k_i}) and (b^{k_i}, b^{k_i}) where $k_i < N$ $(1 \leq i \leq 2m(1 + N!))$. We would like to obtain an identical word of the form (*3) below from multiple w_ℓ for different ℓ by repeating some of the above pumps depending on ℓ.

By the definition of f_{EX3}, the value $f_{\mathrm{EX3}}(w_\ell)$ is obtained when we divide w_ℓ into $\alpha_1\beta_1\alpha_2\beta_2 \cdots \beta_{\ell-1}\alpha_\ell$ and $\beta_\ell\alpha_{\ell+1}\beta_{\ell+1}\alpha_{\ell+2} \cdots \alpha_m\beta_m\$$. Hence, $f_{\mathrm{EX3}}(w_\ell) = |\alpha_1\beta_1\alpha_2\beta_2 \cdots \beta_{\ell-1}\alpha_\ell|_b + |\beta_\ell\alpha_{\ell+1}\beta_{\ell+1}\alpha_{\ell+2} \cdots \alpha_m\beta_m|_a = (\ell-1)N + (m-\ell)N = (m-1)N$. Let T_ℓ be a parse tree of w_ℓ such that $\mathrm{wt}(T_\ell) = (m-1)N$.

Let us apply Lemma 4 to w_ℓ and T_ℓ by letting $n = 2m(1+N!)$, $\Gamma = \{a, b\}, \Delta = \emptyset, \Omega = \{\$\}$ and $\mu_1, \cdots, \mu_{2n} \in \{a^N, b^N\}$, $\lambda_1 = \cdots = \lambda_n = \varepsilon$, $\lambda_{n+2} = \cdots = \lambda_{2n+1} = \varepsilon$, $\lambda_{n+1} = \$\$$ (*1). Then, there are subderivations $A_i \Rightarrow^* v_i A_i y_i$ of T_ℓ where $A_i \in V$, $v_i, y_i \in \Sigma^*$ for each i $(1 \leq i \leq n)$ (*2) such that $1 \leq |v_i| \leq N$, μ_i contains v_i, $v_i = y_i = a^{k_i}$ (or $= b^{k_i}$), and α_{2m-j+1} (or β_{2m-j+1}) contains y_i if α_j (or β_j) contains v_i.

Consider $A_i \overset{c_i}{\Rightarrow}{}^* v_i A_i y_i$ such that v_i is contained in α_ℓ among the subderivations mentioned in (*2). There are exactly $m \cdot N! + 1$ of such subderivations by the following reason. We have $\alpha_\ell = a^{N(m \cdot N!+1)}$ and by the assumption (*1), α_ℓ is the concatenation of some μ_i of length N and hence the number of such μ_i is exactly $m \cdot N! + 1$. Since $\mathrm{wt}(T_\ell) = (m-1)N < m \cdot N! + 1$ and $c_i \in \mathbb{N}$, there is at least one i such that $c_i = 0$. (Otherwise, $\mathrm{wt}(T_\ell)$ would be greater than or equal to $m \cdot N! + 1$.) For any i such that $c_i = 0$, $v_i = y_i = a^{k_i}$ in α_ℓ can be pumped with weight 0. The same property holds for $v_i = b^{k_i}$ contained in β_ℓ.

Next, we consider $v_i = y_i = a^{k_i}$ (resp. $v_i = y_i = b^{k_i}$) that is not contained in $\alpha_\ell, \alpha_{2n-\ell+1}$ (resp. $\beta_\ell, \beta_{2n-\ell+1}$). Note that k_i $(< N)$ must be a devisor of $m \cdot N!$ and all pumps (v_i, y_i) are nested each other on T_ℓ. Hence we can construct a parse tree T'_ℓ of $w' =$

$$\underbrace{a^{N(m\cdot N!+1)} b^{N(m\cdot N!+1)} \cdots b^{N(m\cdot N!+1)}}_{m} \$\$ \underbrace{b^{N(m\cdot N!+1)} \cdots b^{N(m\cdot N!+1)} a^{N(m\cdot N!+1)}}_{m} \tag{*3}$$

by pumping subderivations in T_l.

We now consider two parse trees T'_{ℓ_1}, T'_{ℓ_2} of w' $(1 \leq \ell_1 < \ell_2 \leq m)$. Note that T'_{ℓ_1} can pump subwords a^{k_1} contained in ℓ_1-th $a^{N(m\cdot N!+1)}$ and $b^{k'_1}$ contained in ℓ_1-th $b^{N(m\cdot N!+1)}$ with weight 0, while T'_{ℓ_2} can pump subwords a^{k_2} contained in ℓ_2-th $a^{N(m\cdot N!+1)}$ and $b^{k'_2}$ contained in ℓ_2-th $b^{N(m\cdot N!+1)}$ with weight 0. By the definition of f_{EX3}, the value of f_{EX3} increases if subwords $a^{k_1}, b^{k'_1}, a^{k_2}, b^{k'_2}$ can be all pumped simultaneously. If $T'_{\ell_1} = T'_{\ell_2}$, then this simultaneous pump does not increase the weight, which is a contradiction. Hence, T'_{ℓ_1} and T'_{ℓ_2} are different trees. Thus, $T_1, T_2, \cdots, T_m$ are pairwise different and $|\mathrm{parse}(w')| \geq m$. However, this contradicts the assumption that the ambiguity of G is at most $m-1$. □

Remark 2. In the proof of Theorem 2, we said that some $v_i = a^{k_i}$ contained in α_ℓ can be pumped with weight 0, but we can also say that every v_i contained in α_ℓ can be pumped with weight 0. That is because, if a subword of α_ℓ is generated by a derivation $A_i \overset{c_i}{\Rightarrow}{}^* a^{k_i} A_i a^{k_i} \Rightarrow^* a^{k_i} a^k A_j a^k a^{k_i} \overset{c_j}{\Rightarrow}{}^* a^{k_i} a^k a^{k_j} A_j a^{k_j} a^k a^{k_i}$ (with pairwise different subderivations $A_i \overset{c_i}{\Rightarrow}{}^* a^{k_i} A_i a^{k_i}$ and $A_j \overset{c_j}{\Rightarrow}{}^* a^{k_j} A_j a^{k_j}$), there are 2^n ways to derive $a^{(k_i+k_j)n+k}$. This contradicts the assumption that G is finitely-ambiguous. Therefore, all $v_i = a^{k_i}$ contained in α_ℓ are generated by the same subderivation. This remark also holds for the proof in Theorem 3.

We can prove that $f_{\mathrm{EX4}} \notin$ PA-WCF in a similar way to the proof of Theorem 2.

Theorem 3. $f_{\mathrm{EX4}} \notin$ *PA-WCF.*

Corollary 1. *U-WCF* $\subsetneq$ *FA-WCF* $\subsetneq$ *PA-WCF* $\subsetneq$ *WCF. Furthermore, (U-WCF* $\cap$ *U-CF)* $\subsetneq$ *(FA-WCF* $\cap$ *U-CF)* $\subsetneq$ *(PA-WCF* $\cap$ *U-CF)* $\subsetneq$ *(WCF* $\cap$ *U-CF).*

5 Conclusion

We proved a pumping lemma for CFG, which is helpful for demonstrating an iteration without increasing weights, and showed the strict ambiguity hierarchy of WCFG. Since the functions proved to exist in the gaps are all in U-CF, this hierarchy is different from the ambiguity hierarchy of CFG known as inherent ambiguity. In other words, the hierarchy shown to exist in this paper is caused by introducing weights.

We defined U-CF as the class of functions whose supports are defined by some U-WCFG over $\mathbb{B}$. Similarly, we can define FA-CF and PA-CF as the classes of functions whose supports are defined by some FA-WCFG over $\mathbb{B}$ and some PA-WCFG over $\mathbb{B}$, respectively. For these classes, we expect to prove the inclusion (FA-WCF $\cap$ FA-CF) $\subsetneq$ (PA-WCF $\cap$ FA-CF) $\subsetneq$ (WCF $\cap$ FA-CF) and (PA-WCF $\cap$ PA-CF) $\subsetneq$ (WCF $\cap$ PA-CF) in the same way.

The discussion on the ambiguity hierarchy of WA in [1,7] is generalized by using pumping lemmas that correspond to each hierarchy level. Showing similar pumping lemmas for U-WCFG, FA-WCFG and PA-WCFG is left as future work. However, showing them seems difficult because the expressive power of WCFG cannot be determined by a simple iteration property, as explained in Sect. 1.

The techniques in Theorems 1, 2 and 3 could be applied to other weighted models and other semirings. In particular, Remark 2 is useful. For example, if there are n of the same subderivations $A \overset{c}{\Rightarrow}{}^{*} vAy$ and $f(w) = W$, then c must be smaller than or equal to $W^{1/n}$ for WCFG over the semiring $(\mathbb{N} \cup \{\infty\}, +, \times, 0, 1)$ of natural numbers.

References

1. Chattopadhyay, A., Mazowiecki, F., Muscholl, A., Riveros, C.: Pumping Lemmas for Weighted Automata, CoRR abs/2001.06272 (2020)
2. Chomsky, N., Schützenberger, M.P.: The algebraic theory of context-free languages. Stud. Logic Found. Math. **26**, 118–161 (1959)
3. M. Droste, W. Kuich and H. Vogler, Handbook of Weighted Automata, Springer Science & Business Media, Berlin (2009). https://doi.org/10.1007/978-3-642-01492-5
4. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge (1998)
5. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages, and Computation, Addison-Wesley, Boston (1979)
6. Maletti, A., Nasz, T., Stier, K., Ulbricht, M.: Ambiguity hierarchies for weighted tree automata. In: Maneth, S. (ed.) CIAA 2021. LNCS, vol. 12803, pp. 140–151. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-79121-6_12

7. Mazowiecki, F., Riveros, C.: Pumping lemmas for weighted automata. STACS **50**(1–50), 14 (2018)
8. Ogden, W.: A helpful result for proving inherent ambiguity. Math. Syst. Theor. **2**(3), 191–194 (1968)
9. Wich, K.: Exponential Ambiguity of Context-free Grammars, DLT 1999, pp. 125–138 (1999)