The Parikh Property for Weighted Context-Free Grammars

Pierre Ganty
IMDEA Software Institute, Madrid, Spain
pierre.ganty@imdea.org

Elena Gutiérrez
IMDEA Software Institute, Madrid, Spain
Universidad Politécnica de Madrid, Spain
elena.gutierrez@imdea.org

───── **Abstract** ─────

Parikh's Theorem states that every context-free grammar (CFG) is equivalent to some regular CFG when the ordering of symbols in the words is ignored. The same is not true for the so-called weighted CFGs, which additionally assign a weight to each grammar rule. If the result holds for a given weighted CFG G, we say that G satisfies the Parikh property. We prove constructively that the Parikh property holds for every weighted nonexpansive CFG. We also give a decision procedure for the property when the weights are over the rationals.

1 Introduction

The celebrated Parikh's Theorem [17] establishes that every context-free language is *Parikh-equivalent* to some regular language. Two words w, w' over an alphabet of symbols are Parikh-equivalent if the number of occurrences of each symbol in w coincides with that of w'. For instance, the words $aabb$ and $abab$ over the alphabet $\{a, b\}$ are Parikh-equivalent as both have 2 a's and 2 b's. Two languages L and L' are Parikh-equivalent if for each word in L there is a Parikh-equivalent word in L', and viceversa, e.g., the language $\{ab, aabb\}$ is Parikh-equivalent to the language $\{ba, abab, abba\}$. Consider, for instance, the context-free language $L = \{a^n b^n \mid n \geq 0\}$. Then, a regular language that satisfies Parikh's Theorem is $(ab)^*$. In fact, given a context-free grammar, one can construct a finite-state automaton that recognizes a Parikh-equivalent language [5]. Parikh's Theorem has been applied in automata theory for decision problems concerning Parikh-equivalence such as membership, universality, equivalence and disjointness [4, 11, 12, 13], to establish complexity bounds on verification problems for counter machines [8], equational Horn clauses [22], among many others. It has also found application in the analysis of asynchronous programs with procedures [7, 21] where the Parikh-equivalent finite-state automaton is used to compute another asynchronous program without procedures that preserves safety bugs.

Weighted finite-state automata are a generalization of the classical nondeterministic finite-state automata in which each transition carries a weight. This weight can be defined, for instance, as a nonnegative number representing the cost of its execution. Then, the weight of a path in the weighted automaton can be computed by adding the weights of its transitions. If we are interested in the minimal cost of execution of a given word, we can compute its weight as the minimum of the weights of the paths accepting that word. In

general, the algebraic structure underlying the computation of the weights is that of a *semiring*, an algebraic structure with two operations · (product) and + (sum) used to compute the weight of a path and the weight of a word, respectively. In the same way, it is possible to add weights to the transitions of a pushdown automaton. The later model, so-called weighted pushdown automata, has been used to perform data-flow analysis of programs with procedures [19].

In this paper we study the question of whether Parikh's Theorem can be extended to the weighted case. Roughly speaking, for a given weighted pushdown automaton $\mathcal{P}$, we ask whether there is a weighted finite-state automaton $\mathcal{F}$ that accepts a Parikh-equivalent language and such that for every word w, the sum of the weights of all words Parikh-equivalent to w in $\mathcal{P}$ coincides with that of all Parikh-equivalent words to w in $\mathcal{F}$. Extending Parikh's Theorem to the weighted case has the potential of reaching new applications, for instance, the analysis of event-driven asynchronous programs with procedures where each transition is augmented with the probability of the event associated to it. Finding a weighted finite-state automaton that is Parikh-equivalent to the original program and preserves the probabilities can be used to perform probabilistic analysis of programs following this paradigm.

We will present our results using the grammar model (as opposed to the automata model). It is well-known that both models are equivalent, in the sense that both representations generate the same family of languages of weighted words. Using weighted context-free grammars (WCFGs for short) allows us to exploit their connection with algebraic systems of equations to give more simple and convincing proofs of our results. In a WCFG, a weight is assigned to each rule of the grammar. The notion of weight is extended from rules to parse trees by multiplying the weights of the rules used along a tree, and from parse trees to words by adding the weights of all the possible parse trees that yield to a word. We say that two WCFGs G_1 and G_2 are Parikh-equivalent if for each Parikh-equivalence class $\mathcal{E}$, the sum of the weights of every word in $\mathcal{E}$ under G_1 and G_2 coincide.

We consider the following problem: given a WCFG G, does there exist a Parikh-equivalent WCFG G' that is regular? If the answer is positive we say that G satisfies the Parikh property. It follows from a known counterexample by Petre [18] that the property is not true in general. Recently, Bhattiprolu et al. [2] further investigated this question. They show a class of WCFGs over the unary alphabet that always satisfy the Parikh property. Now, we show that every *nonexpansive* WCFG (over an arbitrary alphabet and arbitrary semiring) satisfies the Parikh property. A WCFG is nonexpansive if no grammar derivation is of the form $X \Rightarrow^* w_0\, X\, w_1\, X\, w_2$. Note that nonexpansiveness is decidable as it reduces to computing predecessors of a regular set [6]. We can show that in the unary case the class of nonexpansive grammars strictly contains the class defined by Bhattiprolu et al. [2] (see Appendix D). However, nonexpansiveness is a sufficient condition for the Parikh property, but not necessary. In particular, we give an example of an expansive WCFG for which there exists a Parikh-equivalent regular WCFG. This shows that a conjecture formulated by Baron and Kuich [1] in 1981 is false[1]. Furthermore, we can show that nonexpansiveness is not necessary for the property even when the alphabet is unary by means of a similar example.

In the second part of our work, we study the question of whether the Parikh property is decidable. As far as we can tell, this question is open. However, it implicitly follows from a result by Kuich et al. [15] that, when we equivalently formulate the property in terms

---

[1]   Essentially, they conjectured that every unambiguous WCFG G is nonexpansive iff G has the Parikh property [1, Conjecture C].

of formal power series, it is decidable over the semiring of rational numbers. Their proof relies on an ad-hoc elimination procedure which is hard to perform even on small examples. Now we give a decision procedure that sidesteps this problem by using a different technique that allows to illustrate the algorithm on small examples with the support of mainstream open-source computer algebra systems.

The document is organized as follows. After preliminaries in Section 2, we show in Section 3 that every nonexpansive WCFG is Parikh-equivalent to a regular WCFG. In Section 4, we give a decision procedure for the property when the weight domain is over the rational numbers and we illustrate its use with several examples. Finally, we give further details of the related work in Section 5, and conclusions and further work in Section 6. Missing proofs can be found in the Appendix.

2 Preliminaries

We denote by Σ^* ($\Sigma^\oplus$) the free (commutative) monoid generated by Σ. The elements of Σ^* are written as words over the alphabet Σ, typically denoted by w, w' and w_i ($i \in \mathbb{N}$), while the elements of $\Sigma^\oplus$ are written as monomials in the variables Σ and they are typically denoted by v, v' and v_i. For instance, if $\Sigma = \{a, b\}$ then all the elements in Σ^* of length two containing 1 a and 1 b are the words ab and ba while the only element with that property in $\Sigma^\oplus$ is the monomial ab.

We denote a *context-free grammar* (CFG for short) as a tuple (V, Σ, S, R) where V is a finite set of *variables* including S, the *start* variable, Σ is the set of *terminals* and $R \subseteq V \times (\Sigma \cup V)^*$ is a finite set of *rules*. Rules are conveniently denoted $X \to \gamma$. We will always assume that CFGs are *cycle-free*, i.e., there is no derivation of the form $X \Rightarrow^+ X$ with $X \in V$. This guarantees that the number of parse trees for one given word is finite and thus the weight of a word is a well-defined function. W.l.o.g., we assume that every *regular* CFG is *right-regular* , i.e., $\gamma \in \Sigma^+(\varepsilon \cup V)$ for each γ. A CFG is *nonexpansive* if no derivation is of the form $X \Rightarrow^* w_0 X w_1 X w_2$ with $X \in V$ and $w_i \in (\Sigma \cup V)^*$. Otherwise, it is *expansive.*

A *semiring* is a structure $(A, +, \cdot, 0_A, 1_A)$ where $(A, +, 0_A)$ is a commutative monoid with identity 0_A, $(A, \cdot, 1_A)$ is a monoid with identity 1_A, $\cdot$ distributes over $+$ and 0_A satisfies that $a \cdot 0_A = 0_A \cdot a = 0_A$, for all $a \in A$. A semiring is called *commutative* iff $a \cdot b = b \cdot a$ for every $a, b \in A$. In the sequel, we assume that A is always a commutative semiring. An *idempotent* semiring is one that satisfies $a + a = a$, for all $a \in A$, . A (commutative) *ring* is a (commutative) semiring where $(A, +, 0_A)$ is a commutative *group* (i.e., every element in A has an additive inverse). Finally, a *field* is a ring where $(A \setminus \{0_A\}, \cdot, 1_A)$ is a *commutative group* (i.e., every element in A except 0_A has a multiplicative inverse). We sometimes use A for both the structure and the underlying set when the meaning is clear from the context. We abuse notation and use $+$ and $\cdot$ to denote the ordinary sum and product in $\mathbb{N}$ and $\mathbb{Q}$. Classical examples of a commutative semirings are $(\mathbb{N}, +, \cdot, 0, 1)$ and $(\mathbb{Q}, +, \cdot, 0, 1)$. The later is also a *field* and we will refer to it as the *rational semiring.* Another classical example of a commutative semiring is the *tropical semiring*, defined as $(\mathbb{N} \cup \{\infty\}, min, +, \infty, 0)$. Note that this semiring is also idempotent as $min(a, a) = a$, for all $a \in \mathbb{N} \cup \{\infty\}$.

A *weighted context-free grammar* (WCFG for short) is a pair (G, W) where G is a CFG as defined above and W is a mapping with the signature $W : R \to A$ that assigns a weight from A to each production in R, for some (commutative) semiring A. Note that W may assign 0_A to some rules in R. The mapping W is usually referred to as the *weight function* of

the WCFG. We extend the definition of W from rules to *derivation sequences*[2] by assigning to each derivation sequence ψ a weight value which is the product of the weights of the rules applied in ψ. We assume that, the derivation policy for G, i.e., the derivation strategy that determines the next variable to rewrite along a derivation, defines one unique derivation sequence for each parse tree. We also assume that the $\cdot$ operation is commutative, i.e., we will always consider *commutative semirings*. Then, the weight of a derivation sequence does not depend on the choice of the derivation policy. Under these assumptions we can extend the definition from rules to *parse trees* (instead of derivation sequences). Before, we recall some definitions. We define a *labeled tree* $c(\tau_1, \ldots, \tau_n)$ (with $n \geq 0$) as a finite tree whose nodes are labeled, where c is the label of the root and $\tau_1, \ldots, \tau_n$ are labeled trees, the children of the roots. When $n = 0$ we prefer to write c instead of $c()$. We simply write $\tau = c(\ldots)$ when the children nodes $\tau_1, \ldots, \tau_n$ are not important. We will write parse trees as labeled trees of the form $\tau = \pi(\tau_1, \ldots, \tau_n)$ to denote that the topmost level of τ is induced by the grammar rule π and has exactly n children nodes which root (from left to right) the parse trees $\tau_1, \ldots, \tau_n$, i.e., the right-hand side of π contains n grammar variables where the i-th (from the left) is derived according to τ_i. We thus define the *yield* of a parse tree $\tau = \pi(\tau_1, \ldots, \tau_n)$, denoted as $\mathcal{Y}(\tau)$ inductively as follows. If $n = 0$, then $\mathcal{Y}(\tau) = \gamma$ where π is of the form $X \to \gamma$ and $\gamma \in \Sigma^* \cup \{\varepsilon\}$. Otherwise, $\mathcal{Y}(\tau) = \alpha_1 \mathcal{Y}(\tau_1) \ldots \alpha_n \mathcal{Y}(\tau_n) \alpha_{n+1}$ where π is of the form $X \to \alpha_1 X_1 \ldots \alpha_n X_n \alpha_{n+1}$ with $\alpha_i \in \Sigma^* \cup \{\varepsilon\}$, and each X_i corresponds to the left-hand side of the rule in the root of τ_i. Define the *weight of a parse tree* $\tau = \pi(\tau_1, \ldots, \tau_n)$ inductively as:

$$W(\tau) \stackrel{\text{def}}{=} W(\pi) \prod_{i=1}^{n} W(\tau_i) \ .$$

Note that $W(\tau)$ does not depend on the order in which we consider the rules in τ as we assume that $\cdot$ is commutative. Denote by $\mathcal{T}_G$ the set of all parse trees of a CFG G. Then, define the *weight of a word* $w \in \Sigma^*$ as follows:

$$W(w) \stackrel{\text{def}}{=} \sum_{\substack{\mathcal{Y}(\tau)=w \\ \tau \in \mathcal{T}_G}} W(\tau) \ .$$

If for some $w \in \Sigma^*$, the set $\{\tau \mid \mathcal{Y}(\tau) = w, \tau \in \mathcal{T}_G\} = \emptyset$ then $W(w) \stackrel{\text{def}}{=} 0_A$. Define the *semantics* of a WCFG (G, W), denoted by $[\![G]\!]_W$, as the mapping $[\![G]\!]_W : \Sigma^* \to A$ such that $[\![G]\!]_W(w) \stackrel{\text{def}}{=} W(w)$. Define the *Parikh image* of a word $w \in \Sigma^*$ with $\Sigma = \{a_1, \ldots, a_n\}$, denoted by $\wr w \wr$ as the monomial $a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_n^{\alpha_n} \in \Sigma^\oplus$ such that α_i is the number of occurrences of a_i in w. Define the *Parikh image* of a weighted context-free grammar (G, W), denoted by $Pk[\![G]\!]_W$, as the mapping $Pk[\![G]\!]_W : \Sigma^\oplus \to A$ such that:

$$Pk[\![G]\!]_W(v) \stackrel{\text{def}}{=} \sum_{\substack{v=\wr w \wr \\ w \in \Sigma^*}} [\![G]\!]_W(w) \ .$$

We write $[\![G]\!]_W$ and $Pk[\![G]\!]_W$ as the formal sums $\sum_{w \in \Sigma^*} [\![G]\!]_W(w) \, w$ and $\sum_{v \in \Sigma^\oplus} Pk[\![G]\!]_W(v) \, v$, respectively. Two WCFGs (G, W) and (G', W') are *language-equivalent* iff $[\![G]\!]_W = [\![G']\!]_{W'}$, while (G, W) and (G', W') are *Parikh-equivalent* iff $Pk[\![G]\!]_W = Pk[\![G']\!]_{W'}$. Finally, a WCFG (G, W) is *regular/nonexpansive/cycle-free* iff G is regular/nonexpansive/cycle-free, respectively.

---

[2] For a definition of *derivation sequence* go to the beginning of Appendix A.

▶ **Definition 1** (Parikh property). A WCFG (G, W) satisfies the *Parikh property* iff there exists a WCFG (G_ℓ, W_ℓ) such that:

1. (G_ℓ, W_ℓ) is regular, and
2. $Pk[\![G]\!]_W = Pk[\![G_\ell]\!]_{W_\ell}$.

3 Sufficient condition for the Parikh property

Petre [18] shows that the Parikh property is not true in general. In the following example we show a well-known WCFG (for instance, see [2, 18]) for which no regular Parikh-equivalent WCFG exists.

▶ **Example 2.** Consider the WCFG (G, W) with $G = (\{X\}, \{a\}, X, \{X \to aXX, X \to a\})$ and the weight function W over $(\mathbb{N}, +, \cdot, 0, 1)$ that assigns 1 to each production in the grammar. Note that, because the alphabet is unary, we have that $Pk[\![G]\!]_W = [\![G]\!]_W$. As W assigns 1 to each grammar rule, the weight of each word can be interpreted as its ambiguity according to G. Then, the reader can check that:

$$[\![G]\!]_W = \sum_{n \geq 0} C_n\, a^{2n+1} = 1a + 1a^3 + 2a^5 + 5a^7 + 14a^9 + 42a^{11} + 132a^{13} + 429a^{15} + \ldots$$

with $C_n = \frac{1}{n+1}\binom{2n}{n}$ the n-th Catalan number. We will see in Example 10 that this formal power series cannot be generated by a regular WCFG.

Now we show that every nonexpansive WCFG over an arbitrary commutative weight domain satisfies the Parikh property.

▶ **Theorem 3.** *Let (G, W) be an arbitrary WCFG. If G is nonexpansive then (G, W) satisfies the Parikh property.*

Proof. The proof is constructive. Here we give the main intuition of the construction. For a complete proof go to Appendix A. For every nonexpansive WCFG (G, W), we give a 2-step construction that results in a Parikh-equivalent regular WCFG (G_ℓ, W_ℓ). The steps are:

1. construct a new WCFG $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$, with $k \in \mathbb{N}$, language-equivalent to (G, W); and
2. construct a regular WCFG (G_ℓ, W_ℓ) Parikh-equivalent to $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$.

The general idea behind the first step is to build a WCFG $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$ that contains all the information needed to define a "strategic" derivation policy. This derivation policy is strategic in the sense that the total number of grammar variables in all *derivation sentences*[3] produced along a derivation sequence is bounded.

In the second step of the construction, we use $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$ to build a regular WCFG (G_ℓ, W_ℓ) that is Parikh-equivalent. Each grammar variable of (G_ℓ, W_ℓ) represents each possible sentence (without the terminals) along a derivation sequence of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$, and each rule simulates a derivation step of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$. Because the number of variables in the sentences is bounded, the number of variables and rules of (G_ℓ, W_ℓ) is necessarily finite.
◀

The converse of Theorem 3 is not true. The next counterexample illustrates this fact by defining an expansive WCFG G_2 for which a Parikh-equivalent regular WCFG G_1 exists. Thus, nonexpansiveness does not provide an exact characterization of the Parikh property.

---

[3] For a definition of *derivation sentence* go to the beginning of Appendix A.

▶ **Example 4.** Consider the WCFG (G_1, W_1) where $G_1 = (\{X_1\}, \{a, \overline{a}\}, X_1, R_1 = \{X_1 \to aX_1, X_1 \to \overline{a}X_1, X_1 \to \varepsilon\})$ and W_1 is defined over $(\mathbb{N}, +, \cdot, 0, 1)$ and assigns 1 to each rule in R_1. First, note that (G_1, W_1) is regular and the weight of each word can be interpreted as its ambiguity according to G_1. Because G_1 is unambiguous, the weight of each word in the language of G_1 is 1. It is easy to see that $[\![G_1]\!]_{W_1}$ is:

$$[\![G_1]\!]_{W_1} = (a+\overline{a})^* = \sum_{n \geq 0}(a+\overline{a})^n = 1\varepsilon + 1a + 1\overline{a} + 1a\overline{a} + 1\overline{a}a + 1aaa + 1aa\overline{a} + 1a\overline{a}a + 1a\overline{a}\overline{a} + \ldots$$

Now consider the expansive WCFG (G_D, W_D) where $G_D = (\{D\}, \{a, \overline{a}\}, D, R_D = \{D \to aD\overline{a}D, D \to \varepsilon\})$ and W_D is defined over $\mathbb{N}$ and assigns 1 to each rule in R_D. The grammar G_D generates the Dyck language L_D over the alphabet $\{a, \overline{a}\}$ and it is also unambiguous (i.e., the weight of each $w \in L_D$ is 1). It is well-known that L_D is a deterministic context-free language (DCFL). Then the complement of L_D, namely $\{a, \overline{a}\}^* \setminus L_D$, is also a DCFL and thus admits an unambiguous CFG. Let $G_{\overline{D}} = (V_{\overline{D}}, \{a, \overline{a}\}, \overline{D}, R_{\overline{D}})$ be the unambiguous CFG that generates $\{a, \overline{a}\}^* \setminus L_D$, and define $(G_{\overline{D}}, W_{\overline{D}})$ where $W_{\overline{D}}$ is defined over $\mathbb{N}$ and assigns 1 to each rule in $R_{\overline{D}}$.

W.l.o.g., assume $V_D \cap V_{\overline{D}} = \emptyset$ and consider a new variable $X_2 \notin V_D \cup V_{\overline{D}}$. Define the WCFG (G_2, W_2) where $G_2 = (\{X_2\} \cup V_D \cup V_{\overline{D}}, \{a, \overline{a}\}, X_2, R_2)$, R_2 is defined as $R_2 = \{X_2 \to D, X_2 \to \overline{D}\} \cup R_D \cup R_{\overline{D}}$ where W_2 is defined over $\mathbb{N}$ and assigns 1 to each rule in R_2. First, G_2 is expansive because G_D is expansive. Furthermore, D and $\overline{D}$ generate unambiguously languages that are complementary over $\{a, \overline{a}\}$. As the weight of each word in (G_2, W_2) corresponds to its ambiguity, we have that $[\![G_2]\!]_{W_2} = (a + \overline{a})^*$. Hence $[\![G_1]\!]_{W_1} = [\![G_2]\!]_{W_2}$ and thus $Pk[\![G_1]\!]_{W_1} = Pk[\![G_2]\!]_{W_2}$. Recall that (G_1, W_1) is regular. We conclude that (G_2, W_2) is expansive and satisfies the Parikh property.   ◀

We can give a similar counterexample over a unary alphabet (see Appendix B). This shows that nonexpansiveness is not necessary for the Parikh property even in the unary case.

4   A decision procedure for the Parikh property over the rationals

In this section we give a decision procedure that tells whether or not a given WCFG with weights over the rational semiring satisfies the Parikh property. Our procedure relies on a decidability result by Kuich and Salomaa [15, Theorem 16.13]. It implicitly follows from this result that the Parikh property is decidable over the rational semiring. However, their decision procedure is hard to follow as it relies on algebraic methods beyond the scope of this field. This makes its implementation rather involved even for small instances. We propose an alternative method to sidestep this problem using Groebner basis theory.

First, we give some preliminaries. In what follows, A will denote a partially ordered commutative semiring. Given A and an alphabet Σ, a *formal power series in commuting variables* is a mapping of $\Sigma^\oplus$ into A. $A\langle\langle\Sigma^\oplus\rangle\rangle$ denotes the set of all formal power series in commuting variables Σ and coefficients in A. The values of a formal power series r are denoted by (r, v) where $v \in \Sigma^\oplus$. As r is a mapping of $\Sigma^\oplus$ into A, it can be written as a formal sum as $r = \sum_{v \in \Sigma^\oplus}(r, v)\, v$. When $v = \varepsilon$ we will write the term $(r, \varepsilon)\varepsilon$ of r simply as (r, ε). We define the *support* of a formal power series as $supp(r) \stackrel{\text{def}}{=} \{v \mid (r, v) \neq 0_A\}$. The subset of $A\langle\langle\Sigma^\oplus\rangle\rangle$ consisting of all series with a finite support is denoted by $A\langle\Sigma^\oplus\rangle$ and its elements are called *polynomials*. Finally, define, for $k \geq 0$, the operator R_k by $R_k(r) \stackrel{\text{def}}{=} \sum_{|v| \leq k}(r, v)v$ where $r \in A\langle\langle\Sigma^\oplus\rangle\rangle$.

Now we establish the connection between WCFGs and algebraic systems in commuting variables. Let (G, W) be a WCFG with $G = (V, \Sigma, X_1, R)$, $V = \{X_1, \ldots, X_n\}$, and W

defined over the semiring A. We associate to (G, W) the algebraic system in commuting variables defined as follows. For each $X_i \in V$:

$$X_i = \sum_{\substack{\pi \in R \\ \pi = (X_i \to \gamma)}} W(\pi) \, \rbrack \, \gamma \, \lbrack \ . \tag{1}$$

We refer to this system as the *algebraic system (in commuting variables) corresponding to* (G, W). Sometimes, we write $A\langle \Sigma^{\oplus} \rangle$-algebraic system to indicate that the coefficients of the system lie in $A\langle \Sigma^{\oplus} \rangle$. Note that (1) can be written as follows. For each $X_i \in V$:

$$X_i = p_i \ , \ \text{with } p_i \in A\langle (\Sigma \cup V)^{\oplus} \rangle \ . \tag{2}$$

A *solution* to (2) is defined as an n-tuple $r = (r_1, \ldots, r_n)$ of elements of $A\langle\langle \Sigma^{\oplus} \rangle\rangle$ such that $r_i = r(p_i)$, for $i = 1, \ldots, n$, where $r(p_i)$ denotes the series obtained from p_i by replacing, for $j = 1, \ldots, n$, simultaneously each occurrence of X_j by r_j. Note that, r_1, the first component of r, always corresponds to the solution for X_1, the initial variable of G. The *approximation sequence* $\sigma^0, \sigma^1, \ldots, \sigma^j, \ldots$ where each σ^j is an n-tuple of elements of $A\langle \Sigma^{\oplus} \rangle$ associated to an algebraic system as (2) is defined as $\sigma^0 = (0_A, \ldots, 0_A)$ and $\sigma^{j+1} = (\sigma^j(p_1), \ldots, \sigma^j(p_n))$ for all $j \geq 0$. We have that $\lim_{j \to \infty} = \sigma^j$ iff for all $k \geq 0$ there exists an $m(k)$ such that $R_k(\sigma^{m(k)+j}) = R_k(\sigma^{m(k)}) = R_k(\sigma)$ for all $j \geq 0$. If $\lim_{j \to \infty} \sigma^j = \sigma$, then σ is a solution of (2) (from Theorem 14.1 in [15]) and is referred to as the *strong solution*. Note that, by definition, the strong solution is unique whenever it exits. Finally, if (G, W) is a regular WCFG then each p_i in its corresponding algebraic system written as in (2) is a polynomial in $A\langle \mathcal{M} \rangle$, where $\mathcal{M}$ denotes the set of monomials of the form $a_1^{\alpha_1} \ldots a_m^{\alpha_m} X_1^{\beta_1} \ldots X_n^{\beta_n}$ with $a_i \in \Sigma$, $\alpha_i, \beta_j \in \mathbb{N}$ for all i and j, and $\sum_{i=1}^{n} \beta_i \leq 1$. We call a system of this form a *regular algebraic system*. Conversely, we associate to each $A\langle \Sigma^{\oplus} \rangle$-algebraic system S in commuting variables of the form (2) a WCFG (G, W) over the semiring A as follows. Define $G = (\{X_1, \ldots, X_n\}, \Sigma, X_1, R)$ and such that $\pi = (X_i \to \gamma) \in R$ iff $(p_i, \gamma) \neq 0_A$. If $\pi \in R$ then $W(\pi) = (p_i, \gamma)$. We will refer to (G, W) as the *WCFG corresponding to the algebraic system* S. Note that if we begin with an algebraic system in commuting variables, then go to the corresponding WCFG and back again to an algebraic system, then the latter coincides with the original. However, if we begin with the WCFG, form the corresponding algebraic system and then again the corresponding WCFG, then the latter grammar may differ from the original.

Next theorem shows that the Parikh image of a cycle-free WCFG corresponds to the solution for the initial variable in the corresponding algebraic system.

▶ **Theorem 5.** *Let (G, W) be a cycle-free WCFG and let S be the algebraic system in commuting variables corresponding to (G, W). Then, the strong solution r of S exists and the first component of r corresponds to $Pk[\![G]\!]_W$.*

Now we introduce the class of *rational* power series in commuting variables Σ with coefficients in the semiring A, denoted by $A^{rat}\langle\langle \Sigma^{\oplus} \rangle\rangle$.

▶ **Definition 6.** $r \in A^{rat}\langle\langle \Sigma^{\oplus} \rangle\rangle$ iff r is the first component of the solution of a regular algebraic system in commuting variables.

From the previous definition and Theorem 5 we can characterize the WCFGs that satisfy the Parikh property as follows.

▶ **Lemma 7.** *Let (G, W) be a cycle-free WCFG. Then (G, W) satisfies the Parikh property iff $Pk[\![G]\!]_W \in A^{rat}\langle\langle \Sigma^{\oplus} \rangle\rangle$.*

Next we observe that every WCFG (G, W) defined over a commutative *ring* with the Parikh property satisfies a linear equation of a special kind. This result directly follows from Theorem 16.4 in [15].

▶ **Theorem 8.** *Let (G, W) be a cycle-free WCFG with W defined over a commutative ring A. Then (G, W) satisfies the Parikh property iff $Pk[\![G]\!]_W$ satisfies a linear equation of the form: $X = sX + t$, for some $s, t \in A\langle\Sigma^\oplus\rangle$ with $(s, \varepsilon) = 0$.*

Proof. The result is a consequence of Theorem 16.4 in [15] and Lemma 7.   ◀

We conclude from the previous theorem that, given a WCFG (G, W) with W defined over a commutative ring, if such a linear equation exists then (G, W) satisfies the Parikh property; otherwise it does not. Now we will use a result by Kuich et al. [15] to conclude that, if (G, W) is defined over $\mathbb{Q}$ then there exists an irreducible polynomial $q(X)$ such that q evaluates to 0 when $X = Pk[\![G]\!]_W$, denoted by $q(Pk[\![G]\!]_W) \equiv 0$. Intuitively, this polynomial contains all the information needed to decide whether or not (G, W) has the Parikh property.

▶ **Theorem 9** (from Theorem 16.9 in [15])**.** *Let S be the $\mathbb{Q}\langle\Sigma^\oplus\rangle$-algebraic system in commuting variables corresponding to a cycle-free WCFG. Let r_1 be the first component of its strong solution. Then there exists an irreducible polynomial $q(X_1)$ with coefficients in $\mathbb{Q}\langle\Sigma^\oplus\rangle$, and unique up to a factor in $\mathbb{Q}\langle\Sigma^\oplus\rangle$, such that $q(r_1) \equiv 0$.*

Kuich et al. [15] show that the polynomial q is effectively computable by means of a procedure based on the classical elimination theory. Now we develop an alternative method using Groebner bases. Before introducing this technique, we give some intuition on the ideas presented above by revisiting the examples of the previous section.

▶ **Example 10.** Consider the cycle-free WCFG (G, W) defined in Example 2 where the weight function W is now defined over $(\mathbb{Q}, +, \cdot, 0, 1)$ and assigns 1 to each production in the grammar. The algebraic system S corresponding to (G, W) is given by the equation $X = a X^2 + a$. Let r_1 be its strong solution. Assume for now that the irreducible polynomial $q(X) \in \mathbb{Q}\langle\{a\}^\oplus\rangle\langle X\rangle$ from Theorem 9 is $q(X) = aX^2 - X + a$ (later we will give its construction using Groebner bases). We will see later that the fact that $q(X)$ is not linear is enough to conclude that (G, W) does not satisfy the Parikh property (as we expected). Note that the solution of S is $r_1 = \frac{1 - \sqrt{1 - 4a^2}}{2a}$, which written as a series corresponds to $r_1 = \sum_{n \geq 0} C_n a^{2n+1}$, with $C_n = \frac{1}{n+1}\binom{2n}{n}$ the n-th Catalan number. It is known that this formal power series cannot be written as the solution of a linear equation with coefficients in $\mathbb{Q}\langle\{a\}^\oplus\rangle$ [2].

▶ **Example 11.** Now we will consider the WCFG given in Example 4. This time we will give a complete definition of its grammar rules and, as in the previous example, we will extend its weight domain from $\mathbb{N}$ to $\mathbb{Q}$. Define the WCFG (G_2, W_2) where $G_2 = (\{X_2, \overline{D}, D, Y, Z\}, \{a, \overline{a}\}, X_2, R_2)$, R_2 is given by:

$$X_2 \to D \mid \overline{D} \qquad\qquad \overline{D} \to D\,\overline{a}\,Y \mid D\,a\,Z \qquad\qquad Z \to D\,a\,Z \mid D \ .$$
$$D \to a\,D\,\overline{a}\,D \mid \varepsilon \qquad\qquad Y \to a\,Y \mid \overline{a}\,Y \mid \varepsilon$$

and the weight function W_2 is defined over $(\mathbb{Q}, +, \cdot, 0, 1)$ and assigns 1 to each production in the grammar. Note that (G_2, W_2) is cycle-free. The grammar variable D generates all

the words in the Dyck language L_D over the alphabet $\{a, \overline{a}\}$, while the variable $\overline{D}$ generates $\{a, \overline{a}\}^* \setminus L_D$. The system S corresponding to (G_2, W_2) consists of the following equations:

$$X_2 = D + \overline{D} \qquad\qquad \overline{D} = D\,\overline{a}\,Y + D\,a\,Z \qquad\qquad Z = D\,a\,Z + D \ .$$
$$D = a\,D\,\overline{a}\,D + 1 \qquad\qquad Y = a\,Y + \overline{a}\,Y + 1$$

Let $\sigma = (r_1, r_2, r_3, r_4, r_5)$ be its strong solution where r_1 corresponds to the solution for the initial variable X_2. Assume for now that the irreducible polynomial $q(X_2) \in \mathbb{Q}\langle\{a, \overline{a}\}^\oplus\rangle\langle X_2\rangle$ described by Theorem 9 is:

$$q(X_2) = (1 - (a + \overline{a}))X_2 - 1 \ .$$

We observe that q is linear in X_2 and can be written as:

$$q(X_2) = (1 - s)X_2 - t = (1 - (a + \overline{a}))X_2 - 1 \ ,$$

with $(s, \varepsilon) = 0$. Thus, by Theorem 8, we conclude that (G_2, W_2) satisfies the Parikh property as we expected.

Now we develop the technique we will use to construct the irreducible polynomial of Theorem 9: Groebner bases. A Groebner basis is a set of polynomials in one or more variables enjoying certain properties. Given a set of polynomials F with coefficients in a field, one can compute a Groebner basis G of F with the property that G has the same solutions as F when interpreted as a polynomial system of equations. Then, problems such as finding the solutions for the system induced by F, or looking for alternative representations of polynomials in terms of other polynomials become easier using G instead of F. One of the main insights for using Groebner bases is that they are *effectively constructable* using standard computer algebra systems, for any set of polynomials with coefficients in a field.

We are interested in computing Groebner bases of algebraic systems in commuting variables corresponding to weighted CFGs. Given a WCFG and its corresponding algebraic system, our goal is to obtain a system with the same solution as the original, and such that one equation in the new system depends only on the initial grammar variable X_1. This equation will contain all the information needed to decide whether or not the given WCFG satisfies the Parikh property. We will not enter into the technical details of how Groebner bases are constructed and their properties as these lie beyond the scope of this document (however, an explicit reference will be given in connection with each result applied). Instead, we will give a result that encapsulates all the preconditions and postconditions we need for our purpose (Theorem 13). We first introduce the definitions that will appear in the theorem.

In what follows, K will always denote a field. First we need to introduce the notion of *ideal*. Let $K\langle V^\oplus\rangle$ denote the ring of polynomials in variables V and with coefficients in K. A subset $I \subset K\langle V^\oplus\rangle$ is an *ideal* iff (i) $0_K \in I$, (ii) if $f, g \in I$ then $f + g \in I$, and (iii) if $f \in I$ and $h \in K\langle V^\oplus\rangle$ then $h \cdot f \in I$. Given a set of polynomials $F = \{f_1, \ldots, f_n\}$, we define $\langle F\rangle$ as $\langle F\rangle \stackrel{\text{def}}{=} \{\sum_{i=1}^n h_i \cdot f_i \mid h_i \in K\langle V^\oplus\rangle, f_i \in F\}$. It can be shown that $\langle F\rangle$ is an ideal [3] and we call it the *ideal generated by F*. When an ideal is generated by a finite number of polynomials $g_1, \ldots, g_n \in K\langle V^\oplus\rangle$, we say that $g_1, \ldots, g_n$ is a *basis* of the ideal. It is known that every ideal in $K\langle V^\oplus\rangle$ has a basis (actually many, but the ones we are particularly interested in are the so-called Groebner bases) [3]. If one considers the set of polynomial equations $\{f = 0 \mid f \in F\}$, denoted by $F = \mathbf{0}$, then the set of all solutions of $F = \mathbf{0}$ is defined as $\{(r_1, r_2, \ldots, r_n) \in K^n \mid f(r_1, \ldots, r_n) \equiv 0, \text{ for all } f \in F\}$. Then, given two sets of polynomials F and G, if $\langle F\rangle = \langle G\rangle$ then the set of solutions of $F = \mathbf{0}$ coincides

with the set of solutions of $G = \mathbf{0}$ [3]. To construct a Groebner basis of an ideal I, one needs to impose first a total ordering on the monomials of variables occuring in I. This choice is significant as different orderings lead to different Groebner bases with different properties. We are interested in computing Groebner bases with the *elimination property* for the initial variable X_1, i.e., bases where at least one polynomial depends only on X_1. Hence, we will always impose the *reverse lexicographic ordering* to construct Groebner bases.

▶ **Definition 12.** Let $V = \{X_1, \ldots, X_n\}$ be a set of variables. Let α and β be two monomials in $V^\oplus$ and let $\overline{\alpha}$ (resp. $\overline{\beta}$) be the vector in $\mathbb{N}^n$ such that its i-th component corresponds to the number of occurrences of the variable X_i in α (resp. β). Then we say that α *is greater than β w.r.t. the reverse lexicographic ordering*, denoted by $\alpha \succ_{revlex} \beta$, iff the first non-zero component of the vector $\overline{\alpha} - \overline{\beta}$ is negative.

Notice that Definition 12 implies an ordering of the variables: $X_n \succ_{revlex} X_{n-1} \succ_{revlex} \ldots \succ_{revlex} X_1$. The reason for choosing the *reverse* lexicographic ordering is that, in order to compute a Groebner basis with the elimination property for the initial variable X_1, we need X_1 to be the least monomial (with one or more variable). In what follows, the phrase "w.r.t. the reverse lexicographic ordering" (for some given $V = \{X_1, \ldots, X_n\}$) will refer to the one described in Definition 12 with variables V, unless stated otherwise. Fixed a total monomial ordering, we define the *leading monomial* of a polynomial p as the greatest monomial in p, and we denote it by $LM(p)$. We define the *leading term* of p as the leading monomial of p together with its coefficient, and we denote it by $LT(p)$. Finally, we introduce the notion of a *reduced* Groebner basis which allows to define uniquely a Groebner basis of an ideal of polynomials. Let F be a set of polynomials and G a Groebner basis of $\langle F \rangle$. We say that G is a *reduced* Groebner basis of $\langle F \rangle$ iff for each $g_i \in G$ (i) the coefficient of $LT(g_i) = 1$; and (ii) $LM(g_i)$ does not divide any term of any g_j with $i \neq j$. For a given set of polynomials F and monomial ordering $\succ$, there exists exactly one reduced Groebner basis of $\langle F \rangle$ w.r.t. $\succ$ [3]. We abuse notation and write $K\langle X \rangle$ instead of $K\langle \{X\}^\oplus \rangle$ to refer to the ring of polynomials in the variable X with coefficients in K. Now we are ready to give the theorem.

▶ **Theorem 13.** *Let K be a field and $V = \{X_1, \ldots, X_n\}$ a set of variables. Let $F \subseteq K\langle V^\oplus \rangle$ be a set of polynomials such that the strong solution of the system $F = \mathbf{0}$ is $(r_1, \ldots, r_n)$ where r_i corresponds to the solution for X_i. Let G be the reduced Groebner basis of $\langle F \rangle$ w.r.t. the reverse lexicographic ordering. Then the following properties are satisfied:*

1. *$(r_1, \ldots, r_n)$ is also the strong solution of the system $G = \mathbf{0}$ and,*
2. *there is exactly one polynomial $g \in G$ s.t. $g \in K\langle X_1 \rangle$, and for that g we have $g(r_1) \equiv 0$.*

Proof. Property 1. follows from the fact that G is a basis of $\langle F \rangle$. Now we prove property 2. G is a Groebner basis of $\langle F \rangle$ w.r.t. the reverse lexicographic ordering. Then, as a result of the Elimination Theorem [3, Theorem 3.1.2], $G \cap K\langle X_1 \rangle$ is a Groebner basis of $\langle F \rangle \cap K\langle X_1 \rangle$. Assume first that $G \cap K\langle X_1 \rangle$ contains only the zero polynomial (the constant polynomial whose coefficients are equal to 0). Then the ideal $\langle F \rangle \cap K\langle X_1 \rangle$ also contains only the zero polynomial. But this contradicts Theorem 9. Then $G \cap K\langle X_1 \rangle$ contains at least one nonzero polynomial g. Assume now that $G \cap K\langle X_1 \rangle$ contains two different elements g_1 and g_2 in $K\langle X_1 \rangle$. W.l.o.g., let g_1 be such that $LM(g_1) \preceq_{lex} LM(g_2)$. Thus, $LM(g_1)$ divides (at least) the leading term of g_2. Then G is not in reduced form (contradiction). We conclude that there is exactly one (nonzero) polynomial $g \in G$ such that $g \in K\langle X_1 \rangle$. Finally, $g(r_1) \equiv 0$ follows from 1. and the fact that $g \in (G \cap K\langle X_1 \rangle)$. ◀

Now we show in Theorem 14 how to construct q using Groebner bases. Finally, we give in Theorem 15 the main result of this section.

▶ **Theorem 14.** *Let S be a $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$-algebraic system in commuting variables corresponding to a cycle-free WCFG and r_1 be the first component of its strong solution. Then an irreducible polynomial $q(X_1)$ with coefficients in $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$ such that $q(r_1) \equiv 0$ can be effectively constructed.*

Proof. We begin with the first part of the algorithm. Let K be the fraction field of $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$, i.e., the smallest field (w.r.t. inclusion) containing $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$. Consider S as defined in (2) (page 7) where now each polynomial p_i has its coefficients in K and its variables in V, and let $F \subseteq K\langle V^{\oplus}\rangle$ be the set of polynomials $\{p_i \mid 1 \leq i \leq n\}$. Construct the reduced Groebner basis G of F w.r.t. the reverse lexicographic ordering. Let $G = \{g_1, \ldots, g_s\}$ with $s \geq 1$. By Theorem 13, there is exactly one $g \in G$ such that $g \in K\langle X_1\rangle$, and g satisfies $g(r_1) \equiv 0$.

We cannot conclude yet that $g(X_1)$ is the polynomial $q(X_1)$ we are looking for since $g(X_1)$ might not be irreducible in the fraction field of $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$. This constitutes the second part of the algorithm which follows the method given in [15] to obtain from $g(X_1)$ an irreducible polynomial $q(X_1)$ such that $q(r_1) \equiv 0$. Compute the factorization[4] of g in the fraction field of $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$ and let $\{q_1(X_1), \ldots, q_m(X_1)\}$ with $m \geq 1$ be the set of all irreducible polynomials obtained thus as factors. Because $g(r_1) \equiv 0$, there exists an index j_0 with $1 \leq j_0 \leq m$ such that $q_{j_0}(r_1) \equiv 0$ and $q_j(r_1) \not\equiv 0$ for $j \neq j_0$ and $1 \leq j \leq m$. Now we show how to find j_0. Using the operator R_k introduced in the beginning of Section 4, we have that $R_k(q_{j_0}(R_k(r_1))) \equiv 0$ for all $k \geq 0$, while for each $j \neq j_0$ there is always an index k_j such that $R_{k_j}(q_j(R_{k_j}(r_1))) \not\equiv 0$. Then, eventually an index j_0 is always found. Let $q_{j_0}(X_1) = \frac{n_k}{d_k}X_1^k + \frac{n_{k-1}}{d_{k-1}}X_1^{k-1} + \ldots + \frac{n_0}{d_0}$ with $k \geq 0, n_i, d_i \in \mathbb{Q}\langle\Sigma^{\oplus}\rangle$ and $d_i \neq 0$ for all i. Let $lcm(d_0, \ldots, d_k)$ denote the least common multiple of $d_0, \ldots, d_k$ and define $q(X_1) = lcm(d_0, \ldots, d_k) \cdot q_{j_0}(X_1)$. Now $q(X_1) \in \mathbb{Q}\langle\Sigma^{\oplus}\rangle\langle X_1\rangle$ and this completes the algorithm.   ◀

▶ **Remark.** It is worth noting that, even though $q(X_1)$ is an irreducible polynomial over K, the fraction field of $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$, it might not be irreducible over $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$ since it might have a factorization consisting of a polynomial $\tilde{q}(X_1) \in \mathbb{Q}\langle\Sigma^{\oplus}\rangle\langle X_1\rangle$ of the same degree and one or more constant polynomials over $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$, i.e., polynomials of degree zero, that are not units in $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$. Ignoring such constant factors, it remains the irreducible polynomial mentioned in Theorem 14.

▶ **Theorem 15.** *Let (G, W) be a cycle-free WCFG with W defined over $\mathbb{Q}$. Then, it is decidable whether or not (G, W) verifies the Parikh property.*

Proof. Let S be the $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$-algebraic system corresponding to G and let r_1 be the first component of its strong solution. Construct the irreducible polynomial $q(X_1)$ with coefficients in $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$ as in Theorem 14. By Theorem 8, we only need to check whether or not the equation $q(X_1) = 0$ can be written as a linear equation of the form: $(1 - s)X_1 - t = 0$, with $s, t \in \mathbb{Q}\langle\Sigma^{\oplus}\rangle$ and $(s, \varepsilon) = 0$. Observe that the procedure given in Theorem 14 is complete, i.e., if the polynomial q obtained is not linear in X_1 then there cannot exist a polynomial $q_\ell(X_1)$ with coefficients in $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$ and linear in X_1 such that $q_\ell(r_1) \equiv 0$. If it were the case, then q_ℓ would be necessarily a factor of q, and this contradicts the fact that q is irreducible over $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$. Then, if q is not linear in X_1, we conclude that (G, W) does not satisfy the Parikh property. Otherwise, $q(X_1)$ can be rewritten as $q(X_1) = (1 - s)X_1 - t$ with $s, t \in \mathbb{Q}\langle\Sigma^{\oplus}\rangle$ and $(s, \varepsilon) = 0$, and we conclude that (G, W) satisfies the Parikh property.   ◀

---

[4] Polynomial factorizations are performed w.r.t. polynomials with coefficients in the fraction field of $\mathbb{Q}\langle\Sigma^{\oplus}\rangle$ which is a computable field.

Consider a WCFG (G, W) with r_1 the first component of the solution of its corresponding algebraic system. Observe that, if the decision procedure returns a positive answer for (G, W) then the polynomial $q(X_1)$ constructed as in Theorem 14 is of the form:

$$q(X_1) = (s_0 - s_1)X_1 - t = 0 \ ,$$

with $s_0 \in \mathbb{Q}, s_0 \neq 0$ and $s_1, t \in \mathbb{Q}\langle \Sigma^{\oplus} \rangle$ with $(s_1, \varepsilon) = (t, \varepsilon) = 0$. It follows that the algebraic system consisting of the equation:

$$X_1 = \frac{1}{s_0} s_1 X_1 + \frac{1}{s_0} t \ , \tag{3}$$

has also r_1 as solution. Then a regular WCFG Parikh-equivalent to (G, W) is the one corresponding to the regular algebraic system (3).

Now we complete Examples 10 and 11 by following the decision procedure given in Theorem 15 and giving the construction of a Parikh-equivalent regular WCFG (if exists). Additionally, we give a third example.

▶ **Example 16.** Consider the WCFG (G, W) given in Example 10. Recall that its corresponding algebraic system S is given by the equation $X = a\,X^2 + a$. Let r be its strong solution. Now we construct the irreducible polynomial $q(X) \in \mathbb{Q}\langle \{a\}^{\oplus} \rangle \langle X \rangle$ following the procedure given in Theorem 14. Let $F = \{a\,X^2 - X + a\}$. The reduced Groebner basis G of F w.r.t. reverse lexicographic ordering is (trivially) $G = \{X^2 - \frac{1}{a}X + 1\}$. Then the polynomial $g \in G$ such that $g \in K\langle X \rangle$ where K is the fraction field of $\mathbb{Q}\langle \{a\}^{\oplus} \rangle$, and $g(r_1) \equiv 0$ is:

$$g(X) = X^2 - \frac{1}{a}X + 1 \ .$$

Note that this polynomial cannot be reduced into factors in the fraction field of $\mathbb{Q}\langle \{a\}^{\oplus} \rangle$. Multiplying g by a, we get $q(X) = aX^2 - X + a \in \mathbb{Q}\langle \{a\}^{\oplus} \rangle \langle X \rangle$ and we conclude that $q(X)$ is the irreducible polynomial described by Theorem 9 (up to a factor in $\mathbb{Q}\langle \{a\}^{\oplus} \rangle$). As $q(X)$ is not linear we conclude that (G, W) does not satisfy the Parikh property.

▶ **Example 17.** Now consider the WCFG given in Example 4 and its corresponding algebraic system S. We construct the irreducible polynomial $q(X_2) \in \mathbb{Q}\langle \{a, \overline{a}\}^{\oplus} \rangle \langle X_2 \rangle$ following the procedure given in Theorem 14. Given F, the set of polynomials in the left-hand sides of the equations of S after moving all monomials from right to left, we construct the reduced Groebner basis G of F w.r.t. reverse lexicographic ordering. For clarity, we just show the polynomial $g \in G$ such that $g \in K\langle X_2 \rangle$ where K is the fraction field of $\mathbb{Q}\langle \{a, \overline{a}\}^{\oplus} \rangle$, and verifies $g(r_1) \equiv 0$:

$$g(X_2) = X_2 - \frac{1}{1 - (a + \overline{a})} \ .$$

This polynomial is linear so it is irreducible over the fraction field of $\mathbb{Q}\langle \{a, \overline{a}\}^{\oplus} \rangle$. Now we multiply g by $(1 - (a + \overline{a}))$ and thus obtain $q(X_2) = (1 - (a + \overline{a}))X_2 - 1 \in \mathbb{Q}\langle \{a, \overline{a}\}^{\oplus} \rangle \langle X_2 \rangle$ which is the irreducible polynomial described by Theorem 9 (up to a factor in $\mathbb{Q}\langle \{a, \overline{a}\}^{\oplus} \rangle$). Now we apply the decision procedure described in Theorem 15. We observe that q can be written as follows:

$$q(X_2) = (1 - s)X_2 - t = (1 - (a + \overline{a}))X_2 - 1 \ ,$$

with $(s, \varepsilon) = 0$. Thus, we conclude that (G, W) satisfies the Parikh property. Finally, we give a regular Parikh-equivalent WCFG (G_ℓ, W_ℓ). The regular algebraic system:

$$(1 - (a + \overline{a}))X_2 - 1 = 0 \iff X_2 = (a + \overline{a})X_2 + 1 \tag{4}$$

has r_1 as solution. Then, the WCFG (G_ℓ, W_ℓ) corresponding to (4) is given by $G_\ell = (\{X_2\}, \{a, \overline{a}\}, R_\ell, X_2)$ with R_ℓ defined as:

$$\pi_1 = X_2 \to aX_2$$
$$\pi_2 = X_2 \to \overline{a}X_2$$
$$\pi_3 = X_2 \to \varepsilon$$

and W_ℓ defined over $(\mathbb{Q}, +, \cdot, 0, 1)$ as $W_\ell(\pi_i) = 1$, for all i. Notice that (G_ℓ, W_ℓ) coincides with (G_1, W_1) in Example 4.

▶ **Example 18.** Consider the cycle-free WCFG (G, W) given by $G = (\{X_1, X_2\}, \{a, b\}, R, X_1)$ with R defined as follows:

$$X_1 \to aX_2X_2$$
$$X_2 \to bX_2 \mid a \ ,$$

and the weight function W over $(\mathbb{Q}, +, \cdot, 0, 1)$ that assigns 1 to each production in the grammar. The algebraic system S corresponding to (G, W) is defined as follows:

$$\begin{cases} X_1 = a \, X_2^2 \\ X_2 = b \, X_2 + a \ . \end{cases}$$

Let $\sigma = (r_1, r_2)$ be its strong solution. Now we construct the irreducible polynomial $q(X_1) \in \mathbb{Q}\langle\{a, b\}^\oplus\rangle\langle X_1\rangle$ following the procedure given in Theorem 14. Let $F = \{X_1 - a \, X_2^2, X_2 - b \, X_2 - a\}$. The reduced Groebner basis[5] G of F w.r.t. lexicographic ordering is:

$$G = \left\{ X_1 - \frac{a^3}{b^2 - 2b + 1}, X_2 + \frac{a}{b - 1} \right\} \ .$$

Clearly, the polynomial $g \in G$ such that $g \in K\langle X_1\rangle$ where K is the fraction field of $\mathbb{Q}\langle\Sigma^\oplus\rangle$, and $g(r_1) \equiv 0$ is:

$$g(X_1) = X_1 - \frac{a^3}{b^2 - 2b + 1} \ .$$

This polynomial cannot be reduced into factors in the fraction field of $\mathbb{Q}\langle\Sigma^\oplus\rangle$. Now we multiply g by $(b^2 - 2b + 1)$ and thus obtain $q(X_1) = (b^2 - 2b + 1)X_1 - a^3$ in $\mathbb{Q}\langle\Sigma^\oplus\rangle\langle X_1\rangle$ which is the irreducible polynomial described by Theorem 9 (up to a factor in $\mathbb{Q}\langle\Sigma^\oplus\rangle$). Now we apply the decision procedure described in Theorem 15. We observe that q is linear in X_1 and can be written as:

$$q(X_1) = (1 - s)X_1 - t = (1 - (2b - b^2))X_1 - a^3 \ ,$$

---

[5] The Groebner basis G was computed using the `groebner_basis` method of the open-source mathematics software system SageMath.

with $(s, \varepsilon) = 0$. Then we conclude that (G, W) satisfies the Parikh property. Note that this is the result expected as (G, W) is nonexpansive. Finally, we give a regular Parikh-equivalent WCFG (G_ℓ, W_ℓ). We know that the algebraic system:

$$X_1 = (2b - b^2)X_1 + a^3 \tag{5}$$

has r_1 as solution. Then the WCFG (G_ℓ, W_ℓ) corresponding to the regular system (6) is given by $G_\ell = (\{X_1\}, \{a, b\}, R_\ell, X_1)$ with R_ℓ defined as:

$$\pi_1 = X_1 \to bX_1$$
$$\pi_2 = X_1 \to b^2 X_1$$
$$\pi_3 = X_1 \to a^3$$

and W_ℓ defined over $(\mathbb{Q}, +, \cdot, 0, 1)$ as:

$$W_\ell(\pi) = \begin{cases} 2 & \text{if } \pi = \pi_1 \\ -1 & \text{if } \pi = \pi_2 \\ 1 & \text{if } \pi = \pi_3 \end{cases} .$$

5   Related Work

The problem of extending Parikh's Theorem to the weighted case has been significantly considered in the literature [2, 14, 16, 18]. Petre [18] establishes that the family of power series in commuting variables that can be generated by regular WCFGs is *strictly* contained in that of the series generated by arbitrary WCFGs. In this way, he shows that Parikh's Theorem does not hold in the weighted case. It is well-known that the Parikh property holds in a commutative and idempotent semiring [2, 14, 16]. Luttenberger et al. [16] deal with WCFGs where the weight of a word corresponds to its ambiguity (or commutative ambiguity when considering monomials instead of words) and they show that if a CFG is nonexpansive then its commutative ambiguity can be expressed by a weighted rational expression relying on the fact that all the parse trees of a nonexpansive CFG are of bounded dimension. We used this fact to give a Parikh-equivalent regular WCFG construction, for a given nonexpansive WCFG defined over *any* commutative semiring. Baron and Kuich [1] gave a similar characterization of nonexpansive grammars using rational power series to that of Luttenberger et al. They also conjectured that an unambiguous WCFG is nonexpansive iff it has the Parikh property. This conjecture appears to be false as evidenced by Example 4. Bhattiprolu et al. [2] also show that the class of *polynomially ambiguous* WCFGs over the unary alphabet satisfies the property. In the unary case, this class is strictly contained in the class of nonexpansive grammars (a proof is given in Appendix D). Finally, our decision procedure relies on a result by Kuich and Saloma [15] that decides if an *algebraic* series in commuting variables with coefficients in $\mathbb{Q}$ is rational. To the best of our knowledge, the connection of this result to a decidability result for the Parikh property was only implicit.

6   Conclusions and Further Work

Note that from the theoretical point of view, our decision procedure can be applied to WCFGs over any arbitrary *field*. For arbitrary semirings, the decidability of the Parikh property remains open. It would be interesting to tackle the question first in the unary case. Finally, Theorem 3 shows an equivalent characterization of the Parikh property. Namely, the

Parikh property holds for a WCFG (G, W) iff there exists a Parikh-equivalent nonexpansive WCFG, i.e., iff (G, W) is not *inherently* expansive. It is known that inherent expansiveness is undecidable in the noncommutative and unweighted case [9], but the question remains unsolved in the commutative case when weights are considered.

References

1 Gerd Baron and Werner Kuich. The characterization of nonexpansive grammars by rational power series. *Information and Control*, 48(2):109–118, 1981. `doi:10.1016/S0019-9958(81)90634-3`.

2 Vijay Bhattiprolu, Spencer Gordon, and Mahesh Viswanathan. Extending Parikh's theorem to weighted and probabilistic context-free grammars. In *QEST 2017*, pages 3–19, 2017. `doi:10.1007/978-3-319-66335-7_1`.

3 David A. Cox, John Little, and Donal O'Shea. *Ideals, varieties, and algorithms - an introduction to computational algebraic geometry and commutative algebra (2. ed.)*. Undergraduate texts in mathematics. Springer, 1997.

4 Javier Esparza. Petri nets, commutative context-free grammars, and basic parallel processes. *Fundam. Inform.*, 31(1):13–25, 1997. `doi:10.3233/FI-1997-3112`.

5 Javier Esparza, Pierre Ganty, Stefan Kiefer, and Michael Luttenberger. Parikh's theorem: A simple and direct automaton construction. *Inf. Process. Lett.*, 111(12):614–619, 2011. `doi:10.1016/j.ipl.2011.03.019`.

6 Javier Esparza, Peter Rossmanith, and Stefan Schwoon. A uniform framework for problems on context-free grammars. *Bulletin of the EATCS*, 72:169–177, 2000.

7 Pierre Ganty and Rupak Majumdar. Algorithmic verification of asynchronous programs. *CoRR*, abs/1011.0551, 2010. `arXiv:1011.0551`.

8 Stefan Göller, Richard Mayr, and Anthony Widjaja To. On the computational complexity of verifying one-counter processes. In *LICS 2009*, pages 235–244, 2009. `doi:10.1109/LICS.2009.37`.

9 Jozef Gruska. A few remarks on the index of context-free grammars and languages. *Information and Control*, 19(3):216–223, 1971. `doi:10.1016/S0019-9958(71)90095-7`.

10 John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to automata theory, languages, and computation - international edition (2. ed)*. Addison-Wesley, 2003.

11 Dung T. Huynh. The complexity of equivalence problems for commutative grammars. *Information and Control*, 66(1/2):103–121, 1985. `doi:10.1016/S0019-9958(85)80015-2`.

12 Thiet-Dung Huynh. The complexity of semilinear sets. In *ICALP 1980*, pages 324–337, 1980. `doi:10.1007/3-540-10003-2_81`.

13 Thiet-Dung Huynh. Deciding the inequivalence of context-free grammars with 1-letter terminal alphabet is Σ_2^P-complete. In *FOCS 1982*, pages 21–31, 1982. `doi:10.1109/SFCS.1982.65`.

14 Werner Kuich. The Kleene and the Parikh theorem in complete semirings. In *ICALP 1987*, pages 212–225, 1987. `doi:10.1007/3-540-18088-5_17`.

15 Werner Kuich and Arto Salomaa. *Semirings, Automata, Languages*, volume 5 of *EATCS Monographs on Theoretical Computer Science*. Springer, 1986. `doi:10.1007/978-3-642-69959-7`.

16 Michael Luttenberger and Maximilian Schlund. Convergence of Newton's method over commutative semirings. *Inf. Comput.*, 246:43–61, 2016. `doi:10.1016/j.ic.2015.11.008`.

17 Rohit Parikh. On context-free languages. *J. ACM*, 13(4):570–581, 1966. `doi:10.1145/321356.321364`.

18 Ion Petre. Parikh's theorem does not hold for multiplicities. *Journal of Automata, Languages and Combinatorics*, 4(1):17–30, 1999.

19 Thomas W. Reps, Stefan Schwoon, Somesh Jha, and David Melski. Weighted pushdown systems and their application to interprocedural dataflow analysis. *Sci. Comput. Program.*, 58(1-2):206–263, 2005. `doi:10.1016/j.scico.2005.02.009`.

20 Arto Salomaa and Matti Soittola. *Automata-Theoretic Aspects of Formal Power Series.* Texts and Monographs in Computer Science. Springer, 1978. `doi:10.1007/978-1-4612-6264-0`.

21 Koushik Sen and Mahesh Viswanathan. Model checking multithreaded programs with asynchronous atomic methods. In *CAV 2006*, pages 300–314, 2006. `doi:10.1007/11817963_29`.

22 Kumar Neeraj Verma, Helmut Seidl, and Thomas Schwentick. On the complexity of equational Horn clauses. In *CADE 2005*, pages 337–352, 2005. `doi:10.1007/11532231_25`.

23 Klaus Wich. Exponential ambiguity of context-free grammars. In *DLT 1999*, pages 125–138, 1999.

A   Proof of Theorem 3

First, we give the definitions we will use in this section. Given a CFG $G = (V, \Sigma, S, R)$, define the *degree* of G as $max\{|\gamma\!\downarrow_V| : (X \to \gamma) \in R\} - 1$, where $\gamma\!\downarrow_V$ denotes the projection of γ onto the variables V. Given a production $\pi = (X \longrightarrow \gamma) \in R$ and a position $1 \leq i \leq |\alpha|$, we define a *derivation step* $\alpha \overset{\pi/i}{\Longrightarrow} \beta$ with $\alpha, \beta \in (\Sigma \cup V)^*$ iff $(\alpha)_i = X$ and $\beta = (\alpha)_1 \ldots (\alpha)_{i-1}\, \gamma\, (\alpha)_{i+1} \ldots (\alpha)_{|\alpha|}$. We omit the position i when it is not important. We say that α and β in $(\Sigma \cup V)^*$ are *derivation sentences* of G. We define a *derivation sequence* $\alpha_0 \overset{\pi_1}{\Longrightarrow} \alpha_1 \overset{\pi_2}{\Longrightarrow} \ldots \overset{\pi_n}{\Longrightarrow} \alpha_n$ iff for every $i \in \{1, \ldots, n\}$, $\alpha_{i-1} \overset{\pi_i}{\Longrightarrow} \alpha_i$ is a derivation step. We call the derivation step $\alpha_{i-1} \overset{\pi_i}{\Longrightarrow} \alpha_i$ the *i-step* of the derivation sequence. A derivation sequence $\psi = \alpha_0 \Rightarrow \cdots \Rightarrow \alpha_n$ of G has *index j*, denoted by $idx(\psi)$, if for every $i \in \{0, \ldots, n\}$, no word $(\alpha_i)\!\downarrow_V$ is longer than j. Now we define the dimension of a labeled tree as follows.

▶ **Definition 19.** Given a labeled tree $\tau = c(\tau_1, \ldots, \tau_n)$ $(n \geq 0)$, the *dimension* of τ represented as $dim(\tau)$ is defined as follows:

$$dim(c(\tau_1, \ldots, \tau_n)) \overset{\text{def}}{=} \begin{cases} 0 & \text{if } n = 0 \\ dim(\tau_i) & \text{if } n > 0 \land |\{i \mid \forall j\colon dim(\tau_j) \leq dim(\tau_i)\}| = 1 \\ dim(\tau_i) + 1 & \text{if } n > 0 \land |\{i \mid \forall j\colon dim(\tau_j) \leq dim(\tau_i)\}| > 1 \end{cases}$$

Now we present the proof of Theorem 3. All the definitions, lemmas and theorems referred there can be found below the proof.

Theorem 3 Let (G, W) be an arbitrary WCFG. If G is nonexpansive then (G, W) satisfies the Parikh property.

Proof. The proof is constructive. For every nonexpansive WCFG (G, W), we give a 2-step construction that results in a Parikh-equivalent regular WCFG (G_ℓ, W_ℓ). The steps are:

1. construct a new WCFG $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil}\right)$, with $k \in \mathbb{N}$, language-equivalent to (G, W); and
2. construct a regular WCFG (G_ℓ, W_ℓ) Parikh-equivalent to $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil}\right)$.

The first part of the construction consists of building a new WCFG $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil}\right)$ (Definition 20 below), so-called *at-most-k-dimension* WCFG of (G, W), which is language-equivalent to the original and where grammar variables are annotated with information about the dimension of the parse trees that can be obtained from these variables. Let us give an intuition on its construction.

For a given CFG G and $k \in \mathbb{N}$ (the choice of $k \in \mathbb{N}$ will be described later on), we define $G^{\lceil k \rceil}$ using the same construction as Luttenberger et al. [16]. They show how to construct, for a given CFG G, a new grammar $G^{\lceil k \rceil}$ with the property that $\mathcal{T}_{G^{\lceil k \rceil}}$ corresponds to the subset of $\mathcal{T}_G$ of trees of dimension at most k. They annotate each grammar variable with the superscript $[d]$ (resp. $\lceil d \rceil$) to denote that only parse trees of dimension exactly d (resp. at most d), where $d \leq k$, can be obtained from these variables. When constructing the grammar, they also consider those rules containing two or more variables in its right-hand side and distinguish which cases yield an increase of dimension. We recall the construction of $G^{\lceil k \rceil}$ in Definition 20.

To define the weight function $W^{\lceil k \rceil}$, we assign to each rule in $G^{\lceil k \rceil}$ the same weight as its corresponding version in G (note that for those rules in $G^{\lceil k \rceil}$ with no corresponding version in G, i.e. the so-called e-rules, we assign the identity 1_A with respect to $\cdot$, where A denotes the weight domain). Let us discuss the choice of k in $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$. Luttenberger et al. [16] also show that if G is a nonexpansive CFG then the dimension of every parse tree in $\mathcal{T}_G$ is bounded (Theorem 21). Moreover, the bound is at most the number of grammar variables of G. Then, for a given nonexpansive WCFG (G, W), define k as this bound. Because k is at most equal to the number of variables of G, such a value is always found and consequently, the first part of the construction always terminates. Finally, we show that the WCFG $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$ is language-equivalent to (G, W) (Lemma 22).

In the second part of the construction, we build a regular WCFG (G_ℓ, W_ℓ) that is Parikh-equivalent to $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$. Esparza et al. [5] show that if the dimension of a parse tree is bounded by k then there exists a derivation sequence for the yield of the tree whose index is bounded by some affine function of k (Lemma 23). We rely on this result to define a special derivation policy over at-most-k-dimension WCFGs, for which we know the dimension of every parse tree is bounded by k. They are called *lowest-dimension-first (LDF) derivations*. We prove that, for every WCFG $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$, the index of an LDF derivation sequence is always bounded by an affine function of k (Lemma 25). Then, each grammar variable of (G_ℓ, W_ℓ) represents each possible sentence (without the terminals) along an LDF derivation sequence of $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$, and each grammar rule is intended to simulate an LDF derivation step of $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$. Because the number of variables in these sentences is bounded, the sets of variables and rules of (G_ℓ, W_ℓ) are necessarily finite. A formal definition of the weighted regular (G_ℓ, W_ℓ) is given in Definition 26 . Finally we show that (G_ℓ, W_ℓ) is Parikh-equivalent to $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$ (Lemma 27) and this concludes the proof.   ◄

Now we give the construction of the at-most-k-dimension WCFG $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$ for a given WCFG (G, W) and $k \in \mathbb{N}$. For the construction of $G^{\lceil k \rceil}$, we rely on the one given by Luttenberger et al. [16].

▶ **Definition 20** (The at-most-k-dimension WCFG)**.** Let (G, W) be a WCFG with $G = (V, \Sigma, S, R)$ and W defined over the commutative semiring A, and let $k \in \mathbb{N}$. Define the *at-most-k-dimension WCFG* $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$ with $G^{\lceil k \rceil} = (V^{\lceil k \rceil}, \Sigma, S^{\lceil k \rceil}, R^{\lceil k \rceil})$ of (G, W) (with $u_0, \ldots, u_n \in \Sigma^*$) as follows:

- The set $V^{\lceil k \rceil}$ of variables is given by

    $$\{X^{[d]}, X^{\lceil d \rceil} \mid X \in V, 0 \leq d \leq k\} \ .$$

- The set $R^{\lceil k \rceil}$ of production rules is given by
    1. Linear rules:
        - $r_0(\pi) = \{X^{[0]} \to u_0\}$ for each rule $\pi = (X \to u_0) \in R$.

- $r_1(\pi) = \{X^{[d]} \to u_0 X_1^{[d]} u_1 \mid 0 \le d \le k\}$ for each rule $\pi = (X \to u_0 X_1 u_1) \in R$.

2. Non-linear rules:

   For each rule $\pi = (X \to u_0 X_1 u_1 \ldots u_{n-1} X_n u_n) \in R$

   - $r_2(\pi) = \{X^{[d]} \to u_0 Z_1 u_1 \ldots u_{n-1} Z_n u_n \mid 1 \le d \le k, J \subseteq \{1, \ldots, n\}$ with $|J| = 1 :$ $Z_i = X_i^{[d]}$ if $i \in J$, and $Z_i = X_i^{[d-1]}$ for all $i \in \{1, \ldots, n\} \setminus J\}$ and
   - $r_3(\pi) = \{X^{[d]} \to u_0 Z_1 u_1 \ldots u_{n-1} Z_n u_n \mid 1 \le d \le k, J \subseteq \{1, \ldots, n\}$ with $|J| \ge 2 :$ $Z_i = X_i^{[d-1]}$ for all $i \in J$ and $Z_i = X_i^{[d-1]}$ for all $i \in \{1, \ldots, n\} \setminus J\}$.

3. e-rules:

   - $r_4 = \{X^{\lceil d \rceil} \to X^{[e]} \mid 0 \le e \le d \le k\}$.

- The weight function $W^{\lceil k \rceil}$ is given by

$$
W^{\lceil k \rceil}(\varphi) = \begin{cases} W(\pi) & \text{if } \varphi \in r_0(\pi) \text{ for some } \pi = (X \to u_0) \in R \\ W(\pi) & \text{if } \varphi \in r_1(\pi) \text{ for some } \pi = (X \to u_0 X_1 u_1) \in R \\ W(\pi) & \text{if } \varphi \in r_2(\pi) \cup r_3(\pi) \text{ for some } \pi = (X \to u_0 Z_1 u_1, \ldots, u_{n-1} Z_n u_n) \in R \\ 1_A & \text{if } \varphi \in r_4 \end{cases}
$$

We say that a variable $Z \in V^{\lceil k \rceil}$ is *of dimension d* iff either $Z = X^{\lceil d \rceil}$, or $Z = X^{[d]}$, with $X \in V$, and we denote it by $dim(Z) = d$. Define $V^{(d)} \stackrel{\text{def}}{=} \{Z \in V^{\lceil k \rceil} \mid dim(Z) = d\}$, for each $0 \le d \le k$.

▶ **Theorem 21** (from Theorem 3.3 in [16]). *Let G be a nonexpansive CFG with n variables. Then there exists $k \in \mathbb{N}$ with $k \le n$ such that every parse tree in $\mathcal{T}_G$ has dimension at most k.*

▶ **Lemma 22.** $[\![G]\!]_W = [\![G^{\lceil k \rceil}]\!]_{W^{\lceil k \rceil}}$.

Proof. First recall that k corresponds to the nonnegative value such that every parse tree in G has dimension at most k. We want to show that there is a bijection μ from $\mathcal{T}_{G^{\lceil k \rceil}}$ to $\mathcal{T}_G$ that preserves the yield and the weight of each parse tree.

First, define $\mathcal{T}_G^{\le k} \stackrel{\text{def}}{=} \{\tau \mid \tau \in \mathcal{T}_G, dim(\tau) \le k\}$. Luttenberger et al. [16] prove that there is a bijection μ from $\mathcal{T}_{G^{\lceil k \rceil}}$ to $\mathcal{T}_G^{\le k}$ that preserves the yield of parse trees. Roughly speaking, μ contracts the edges corresponding to the e-rules and removes the superscripts from the labels of the trees. Note that $X^{\lceil d \rceil}$ can only be rewritten to $X^{[e]}$ for some $e \le d$. Then, contracting the corresponding edges cannot change the yield of the corresponding tree. Furthermore, the rules of $G^{\lceil k \rceil}$ that rewrite the variable $X^{[d]}$ are obtained from the rules of G that rewrite X by only adding a superscript. Hence, by removing these annotations again, every tree $\tau \in \mathcal{T}_{G^{\lceil k \rceil}}$ is mapped by μ to a tree in $\mathcal{T}_G^{\le k}$ with the same yield. The complete proof is in Lemma 3.2 in [16]. Furthermore, because G is nonexpansive we have that $\mathcal{T}_G^{\le k} = \mathcal{T}_G$. Thus, if G is nonexpansive, then μ is a bijection from $\mathcal{T}_{G^{\lceil k \rceil}}$ to $\mathcal{T}_G$ that preserves the yield of parse trees.

Now we show that μ also preserves the weights of parse trees, i.e., for each $\tau \in \mathcal{T}_{G^{\lceil k \rceil}}$: $W^{\lceil k \rceil}(\tau) = W(\mu(\tau))$. We proceed by induction on the number of nodes of τ. In the base case, τ has one node, i.e., it has no children. Then $\tau = \varphi$ with $\varphi = X^{[0]} \to u_0$ and $u_0 \in \Sigma^*$, and $\mu(\tau) = \varphi'$ with $\varphi' = X \to u_0$. Then we have:

$$
\begin{aligned}
W^{\lceil k \rceil}(\tau) &= W^{\lceil k \rceil}(\varphi) & \tau = \varphi \\
&= W(\varphi') & \text{by definition of } W^{\lceil k \rceil} \\
&= W(\mu(\tau)) & \mu(\tau) = \varphi'
\end{aligned}
$$

For the induction step, assume $\tau = \varphi(\tau_1, \ldots, \tau_n)$ with $n \geq 1$ and φ is a rule from the set r_i with $i \in \{1, \ldots, 5\}$ (see Definition 20). We distinguish three cases:

- Assume $\varphi \in r_1$ Then $\tau = \varphi(\tau_1)$ and $\mu(\tau) = \varphi'(\mu(\tau_1))$, with $\varphi' = X \to u_0 \, X_1 \, u_1$ and $u_0, u_1 \in \Sigma^*$.

$$
\begin{aligned}
W^{\lceil k \rceil}(\tau) &= W^{\lceil k \rceil}(\varphi) \cdot W^{\lceil k \rceil}(\tau_1) && \text{by definition of weight of } \tau \\
&= W(\varphi') \cdot W(\mu(\tau_1)) && \text{by definition of } W^{\lceil k \rceil} \text{ and induction hyp.} \\
&= W(\mu(\tau)) && \mu(\tau) = \varphi'(\mu(\tau_1))
\end{aligned}
$$

- Assume $\varphi \in r_4$ Then $\tau = \varphi(\tau_1)$ and $\mu(\tau) = \mu(\tau_1)$.

$$
\begin{aligned}
W^{\lceil k \rceil}(\tau) &= W^{\lceil k \rceil}(\varphi) \cdot W^{\lceil k \rceil}(\tau_1) && \tau = \varphi(\tau_1) \\
&= 1_A \cdot W(\mu(\tau_1)) && \text{by induction hyp and definition of } W^{\lceil k \rceil} \\
&= W(\mu(\tau)) && \mu(\tau) = \mu(\tau_1)
\end{aligned}
$$

- Assume $\varphi \in r_2 \cup r_3$ Then $\tau = \varphi(\tau_1, \ldots, \tau_n)$ and $\mu(\tau) = \varphi'(\mu(\tau_1), \ldots, \mu(\tau_n))$ with $\varphi' = X \to u_0 \, X_1 \, u_1 \ldots u_{n-1} \, X_n \, u_n$ and $u_0, \ldots, u_n \in \Sigma^*$.

$$
\begin{aligned}
W^{\lceil k \rceil}(\tau) &= W^{\lceil k \rceil}(\varphi) \prod_{i=1}^{n} W^{\lceil k \rceil}(\tau_i) && \tau = \varphi(\tau_1, \ldots, \tau_n) \\[2ex]
&= W(\varphi') \prod_{i=1}^{n} W(\mu(\tau_i)) && \text{by definition of } W^{\lceil k \rceil} \text{ and induction hyp.} \\[2ex]
&= W(\mu(\tau)) && \mu(\tau) = \varphi'(\mu(\tau_1), \ldots, \mu(\tau_n))
\end{aligned}
$$

Finally, for each $w \in \Sigma^*$:

$$
[\![G]\!]_W(w) = W(w) = \sum_{\substack{w = \mathcal{Y}(\tau) \\ \tau \in \mathcal{T}_G}} W(\tau) = \sum_{\substack{w = \mathcal{Y}(\tau') \\ \tau' \in \mathcal{T}_{G^{\lceil k \rceil}}}} W^{\lceil k \rceil}(\tau') = W^{\lceil k \rceil}(w) = [\![G^{\lceil k \rceil}]\!]_{W^{\lceil k \rceil}}(w) \ .
$$

$\blacktriangleleft$

▶ **Lemma 23** (from Lemma 2.2 in [5]). *Let G be a CFG of degree m and let $\tau \in \mathcal{T}_G$ with $dim(\tau) \leq k$ and $k \in \mathbb{N}$. Then there is a derivation sequence for $\mathcal{Y}(\tau)$ of index at most $km + 1$.*

Now we define a derivation policy over at-most-k-dimension WCFGs. We will prove that this derivation policy satisfies Lemma 23 and thus the index of every derivation is bounded. We call these derivations *lowest-dimension-first (LDF) derivations.*

Intuitively, given a parse tree τ of an at-most-k-dimension WCFG, we define the LDF derivation sequence of τ by performing a depth-first traversal of τ where nodes in the same level of the tree are visited from lower to greater dimension and, if more than one node has the same dimension, then from left to right. Recall that the dimension of a node corresponds to the dimension of the parse tree that it roots.

Before giving a formal definition, we introduce the following notation. Given a derivation sequence $\psi = \alpha_0 \Rightarrow \ldots \Rightarrow \alpha_n$ and β_0, β_1 (possibly empty) sequences of symbols and/or variables, we will denote by $\beta_0 \, \psi \, \beta_1$ the derivation sequence $\beta_0 \, \alpha_0 \, \beta_1 \Rightarrow \ldots \Rightarrow \beta_0 \, \alpha_n \, \beta_1$.

▶ **Definition 24.** Let $G^{\lceil k \rceil}$ be an at-most-k-dimension CFG as in Definition 20. Let $\tau = \pi(\tau_1, \ldots, \tau_n)$ be a parse tree of $G^{\lceil k \rceil}$. Define the *lowest-dimension-first (LDF) derivation sequence ψ of τ* inductively as follows:

■ If $n = 0$, then π is of the form $\pi = X^{[0]} \to u_0$, and $\tau = \pi$. Then, the LDF derivation sequence of τ is:

$$\psi = X^{[0]} \Rightarrow^{\pi}_{ldf} u_0 \ .$$

■ If $n \geq 1$, we distinguish the following cases:

1. If $\pi \in r_1$, i.e., π is of the form $\pi = X^{[d]} \to u_0 X^{[d]}_1 u_1$ with $0 \leq d \leq k$, and $\tau = \pi(\tau_1)$. Then, the LDF derivation sequence of τ is:

$$\psi = X^{[d]} \Rightarrow^{\pi}_{ldf} u_0 X^{[d]}_1 u_1 \Rightarrow_{ldf} u_0 \psi_1 u_1 \ ,$$

where ψ_1 is the LDF derivation sequence of τ_1.

2. If $\pi \in r_4$, i.e., π is of the form $\pi = X^{\lceil d \rceil} \to X^{[e]}$ with $0 \leq e \leq d \leq k$, and $\tau = \pi(\tau_1)$. Then, the LDF derivation sequence of τ is:

$$\psi = X^{\lceil d \rceil} \Rightarrow^{\pi}_{ldf} X^{[e]} \Rightarrow_{ldf} \psi_1 \ ,$$

where ψ_1 is the LDF derivation sequence of τ_1.

3. If $\pi \in r_2$, w.l.o.g., we assume that π is of the form:

$$\pi = X^{[d]} \to u_0 X^{[d]}_1 u_1 X^{\lceil d-1 \rceil}_2 u_2 \ldots u_{n-2} X^{\lceil d-1 \rceil}_{n-1} u_{n-1} X^{\lceil d-1 \rceil}_n u_n \ ,$$

with $1 \leq d \leq k$, and $\tau = \pi(\tau_1, \ldots, \tau_n)$. Define, for each $i \in \{2, \ldots, n\}$, the derivation sequence $\tilde{\psi}_i$ as follows:

$$\tilde{\psi}_i \overset{\text{def}}{=} u_0 \, X^{[d]}_1 \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, X^{\lceil d-1 \rceil}_i \, u_i \, \ldots \, u_{n-1} \, X^{\lceil d-1 \rceil}_n \, u_n$$
$$\Rightarrow^{*}_{ldf} u_0 \, X^{[d]}_1 \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, \psi_i \, u_i \, X^{\lceil d-1 \rceil}_{i+1} \, u_{i+1} \, \ldots \, u_{n-1} \, X^{\lceil d-1 \rceil}_n \, u_n \ ,$$

where ψ_i is the LDF derivation sequence of τ_i. And define:

$$\tilde{\psi}_1 \overset{\text{def}}{=} u_0 \, X^{[d]}_1 \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, u_{n-1} \, \mathcal{Y}(\tau_n) \, u_n$$
$$\Rightarrow^{*}_{ldf} u_0 \, \psi_1 \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, u_{n-1} \, \mathcal{Y}(\tau_n) \, u_n \ ,$$

where ψ_1 is the LDF derivation sequence of τ_1. Then the LDF derivation ψ of τ is:

$$\psi = X^{[d]} \Rightarrow^{\pi}_{ldf} \tilde{\psi}_2 \Rightarrow_{ldf} \ldots \Rightarrow_{ldf} \tilde{\psi}_n \Rightarrow_{ldf} \tilde{\psi}_1 \ .$$

4. If $\pi \in r_3$, w.l.o.g., we assume that π is of the form:

$$\pi = X^{[d]} \to u_0 X^{\lceil d-1 \rceil}_1 u_1 X^{\lceil d-1 \rceil}_2 u_2 \ldots u_{n-2} X^{[d-1]}_{n-1} u_{n-1} X^{[d-1]}_n u_n \ ,$$

with $1 \leq d \leq k$, and $\tau = \pi(\tau_1, \ldots, \tau_n)$. Define, for each $i \in \{1, \ldots, n\}$, the derivation sequence $\tilde{\psi}_i$ as follows:

$$\tilde{\psi}_i \overset{\text{def}}{=} u_0 \, \mathcal{Y}(\tau_1) \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, X^{\lceil d-1 \rceil}_i \, u_i \, \ldots \, u_{n-1} \, X^{[d-1]}_n \, u_n$$
$$\Rightarrow^{*}_{ldf} u_0 \, \mathcal{Y}(\tau_1) \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, \psi_i \, u_i \, X^{\lceil d-1 \rceil}_{i+1} \, u_{i+1} \, \ldots \, u_{n-1} \, X^{[d-1]}_n \, u_n \ ,$$

where ψ_i is the LDF derivation sequence of τ_i. The the LDF derivation ψ of τ is:

$$\psi = X^{[d]} \Rightarrow_{ldf} \tilde{\psi}_1 \Rightarrow_{ldf} \ldots \Rightarrow_{ldf} \tilde{\psi}_n \ .$$

Note that, given a parse tree τ of $G^{\lceil k \rceil}$, the LDF derivation sequence of τ is uniquely defined.

▶ **Lemma 25.** *Let $G^{\lceil k \rceil}$ be an at-most-k-dimension CFG of degree m and $\tau \in \mathcal{T}_{G^{\lceil k \rceil}}$ such that $dim(\tau) \leq k$. Then, the LDF derivation sequence of τ verifies $idx(\psi) \leq km + 1$.*

Proof. Let $G^{\lceil k \rceil} = (V^{\lceil k \rceil}, \Sigma, S^{\lceil k \rceil}, R^{\lceil k \rceil})$. We prove the more general statement: let m be the degree of $G^{\lceil k \rceil}$ and let $\tau \in \mathcal{T}_{G^{\lceil k \rceil}}$ such that $dim(\tau) \leq d$. Then, the LDF derivation sequence ψ of τ satisfies $idx(\psi) \leq dm + 1$. The proof goes by induction on the number of nodes of τ. In the base case, τ has one node, i.e., it has no children. Then, $d = 0$ and the LDF derivation of τ is $\psi = X^{[0]} \Rightarrow^{\pi}_{ldf} u_0$ with $\pi = (X^{[0]} \to u_0) \in R^{\lceil k \rceil}$. Clearly, the index of ψ is 1.

For the induction step, assume that $\tau = \pi(\tau_1, \ldots, \tau_n)$ with $n \geq 1$. We split the proof into the following four cases:

- If $\pi \in r_1$, then π is of the form $\pi = X^{[d]} \to u_0 X_1^{[d]} u_1$ with $0 \leq d \leq k$, and $\tau = \pi(\tau_1)$ with $dim(\tau) \leq d$. By induction hypothesis, the LDF derivation sequence ψ_1 of τ_1 verifies $idx(\psi_1) \leq (dm + 1)$. Then, the LDF derivation of τ is:

$$\psi = X^{[d]} \Rightarrow_{ldf} u_0 X_1^{[d]} u_1 \Rightarrow^*_{ldf} u_0 \psi_1 u_1,$$

  and verifies $idx(\psi) \leq dm + 1$.

- If $\pi \in r_4$, then π is of the form $\pi = X^{\lceil d \rceil} \to X^{[e]}$ with $0 \leq e \leq d \leq k$, and $\tau = \pi(\tau_1)$ with $dim(\tau) \leq d$. By induction hypothesis, the LDF derivation sequence ψ_1 of τ_1 s.t. $idx(\psi_1) \leq (em + 1)$. Then, the LDF derivation of τ is:

$$\psi = X^{\lceil d \rceil} \Rightarrow_{ldf} X^{[e]} \Rightarrow^*_{ldf} \psi_1,$$

  and verifies $idx(\psi) \leq em + 1 \leq dm + 1$.

- If $\pi \in r_2$, then, w.l.o.g., π is of the form:

$$\pi = X^{[d]} \to u_0 X_1^{[d]} u_1 X_2^{\lceil d-1 \rceil} u_2 \ldots u_{n-2} X_{n-1}^{\lceil d-1 \rceil} u_{n-1} X_n^{\lceil d-1 \rceil} u_n,$$

  with $1 \leq d \leq k$, and $\tau = \pi(\tau_1, \ldots, \tau_n)$ with $dim(\tau) \leq d$. By induction hypothesis, for each $i \in \{2, \ldots, n\}$, there is a derivation ψ_i of τ_i s.t. $idx(\psi_i) \leq ((d-1)m + 1)$, and there is a derivation ψ_1 for τ_1 s.t. $idx(\psi_1) \leq dm + 1$. Now, define, for each $i \in \{2, \ldots, n\}$, the derivation sequence $\tilde{\psi}_i$ as follows:

$$\tilde{\psi}_i \stackrel{\text{def}}{=} u_0 \, X_1^{[d]} \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, X_i^{\lceil d-1 \rceil} \, u_i \, \ldots \, u_{n-1} \, X_n^{\lceil d-1 \rceil} u_n$$
$$\Rightarrow^*_{ldf} u_0 \, X_1^{[d]} \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, \psi_i \, u_i \, X_{i+1}^{\lceil d-1 \rceil} \, u_{i+1} \, \ldots \, u_{n-1} \, X_n^{\lceil d-1 \rceil} u_n \ .$$

  And define:

$$\tilde{\psi}_1 \stackrel{\text{def}}{=} u_0 \, X_1^{[d]} \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, u_{n-1} \, \mathcal{Y}(\tau_n) \, u_n$$
$$\Rightarrow^*_{ldf} u_0 \, \psi_1 \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, u_{n-1} \, \mathcal{Y}(\tau_n) \, u_n \ .$$

  Then, the LDF derivation ψ of τ is:

$$\psi = X^{[d]} \Rightarrow^{\pi}_{ldf} \tilde{\psi}_2 \Rightarrow_{ldf} \ldots \Rightarrow_{ldf} \tilde{\psi}_n \Rightarrow_{ldf} \tilde{\psi}_1 \ .$$

  Observe that $n - 1 \leq m$ where n is the number of variables occurring in the right-hand side of π and m is the degree of $G^{\lceil k \rceil}$. For each $i \in \{2, \ldots, n\}$, the index of $\tilde{\psi}_i$ is at most $(d-1)m + 1 + (n - i) \leq dm + 1$. On the other hand, the index of $\tilde{\psi}_1$ is at most $dm + 1$. Then, performing the derivation steps of ψ in the order shown above we have that $idx(\psi) \leq dm + 1$.

- If $\pi \in r_3$, then, w.l.o.g., π is of the form:

$$\pi = X^{[d]} \to u_0 X_1^{\lceil d-1 \rceil} u_1 X_2^{\lceil d-1 \rceil} u_2 \ldots u_{n-2} X_{n-1}^{\lceil d-1 \rceil} u_{n-1} X_n^{\lceil d-1 \rceil} u_n,$$

with $1 \le d \le k$, and $\tau = \pi(\tau_1, \ldots, \tau_n)$ with $dim(\tau) \le d$. By induction hypothesis, for each $i \in \{1, \ldots, n\}$, there is a derivation ψ_i of τ_i s.t. $idx(\psi_i) \le ((d-1)m+1)$. Now, define, for each $i \in \{1, \ldots, n\}$, the derivation sequence $\tilde{\psi}_i$ as follows:

$$\tilde{\psi}_i \overset{\text{def}}{=} u_0 \, \mathcal{Y}(\tau_1) \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, X_i^{\lceil d-1 \rceil} \, u_i \, \ldots \, u_{n-1} \, X_n^{\lceil d-1 \rceil} \, u_n$$

$$\Rightarrow_{ldf}^* \; u_0 \, \mathcal{Y}(\tau_1) \, u_1 \, \mathcal{Y}(\tau_2) \, u_2 \, \ldots \, \mathcal{Y}(\tau_{i-1}) \, u_{i-1} \, \psi_i \, u_i \, X_{i+1}^{\lceil d-1 \rceil} \, u_{i+1} \, \ldots \, u_{n-1} \, X_n^{\lceil d-1 \rceil} \, u_n \; .$$

Then, the LDF derivation ψ of τ is:

$$\psi = X^{[d]} \Rightarrow_{ldf}^{\pi} \tilde{\psi}_1 \Rightarrow_{ldf} \ldots \Rightarrow_{ldf} \tilde{\psi}_n \; .$$

For each $i \in \{1, \ldots, n\}$, the index of $\tilde{\psi}_i$ is at most $(d-1)m + 1 + (n-i) \le dm + 1$. It follows that $idx(\psi) \le dm + 1$.

◀

Given a derivation sentence $\alpha \in \left(\Sigma \cup V^{\lceil k \rceil} \right)^*$ of an at-most-k-dimension CFG, define $\mathcal{LDF}(\alpha) \overset{\text{def}}{=} \alpha|_\Sigma \, \alpha|_{V^{(0)}} \, \alpha|_{V^{(1)}} \ldots \alpha|_{V^{(k)}}$ and $\mathcal{LDF}_{V^{\lceil k \rceil}}(\alpha) \overset{\text{def}}{=} (\mathcal{LDF}(\alpha))|_{V^{\lceil k \rceil}}$. Now we show how to construct a regular (G_ℓ, W_ℓ) that is Parikh-equivalent to $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$.

▶ **Definition 26** (Regular WCFG for $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$). Let $\left(G^{\lceil k \rceil}, W^{\lceil k \rceil} \right)$ be an at-most-k-dimension WCFG with $G^{\lceil k \rceil} = (V^{\lceil k \rceil}, \Sigma, S^{\lceil k \rceil}, R^{\lceil k \rceil})$ and degree m, and $W^{\lceil k \rceil}$ defined over the commutative semiring A. Define the WCFG (G_ℓ, W_ℓ) with $G_\ell = (V_\ell, \Sigma, S_\ell, R_\ell)$ as follows:

- Each variable in V_ℓ corresponds to a sequence $\alpha \in \left(V^{\lceil k \rceil} \right)^{km+1}$ where $\left(V^{\lceil k \rceil} \right)^{km+1}$ denotes the set $\{w \mid w \in (V^{\lceil k \rceil})^*, |w| \le km + 1\}$, and we denote it by $\langle \alpha \rangle$. Formally,

$$V_\ell \overset{\text{def}}{=} \{\langle \alpha \rangle \mid \alpha \in \left(V^{\lceil k \rceil} \right)^{km+1}\} \; .$$

- The initial variable is defined as $S_\ell \overset{\text{def}}{=} S^{\lceil k \rceil}$.
- For each rule $\pi = (X \longrightarrow \gamma) \in R^{\lceil k \rceil}$ define $\pi^\alpha \overset{\text{def}}{=} (\langle X \, \alpha \rangle \longrightarrow \gamma|_\Sigma \, \langle \mathcal{LDF}_{V^{\lceil k \rceil}}(\gamma) \, \alpha \rangle)$. The set R_ℓ of rules is given by

$$\{\pi^\alpha \mid \pi = (X \longrightarrow \gamma) \in R^{\lceil k \rceil} \text{ and } \langle X\alpha \rangle, \langle \mathcal{LDF}_{V^{\lceil k \rceil}}(\gamma) \, \alpha \rangle \in V_\ell\} \; .$$

- The weight function W_ℓ is given by

$$W_\ell(\pi^\alpha) \overset{\text{def}}{=} W^{\lceil k \rceil}(\pi) \text{ for all } \pi^\alpha \in R_\ell \; .$$

▶ **Lemma 27.** $Pk[\![G^{\lceil k \rceil}]\!]_{W^{\lceil k \rceil}} = Pk[\![G_\ell]\!]_{W_\ell}$.

Proof. First, we give the definitions and notation we will use in this proof. For convenience, we will give an alternative definition of the weight of a word using derivation sequences. Recall that we assume that the derivation policy of a grammar defines for each parse tree one unique derivation sequence. Given a CFG G and $w \in \Sigma^*$, define by $parse_G(w)$ as the subset of all derivations of G that yield to $w \in \Sigma^*$. Then, define for each derivation sequence $\psi = \alpha_0 \Rightarrow^{\pi_1} \alpha_1 \Rightarrow^{\pi_2} \ldots \Rightarrow^{\pi_n} \alpha_n$ of G the *weight of ψ* as follows:

$$W(\psi) \overset{\text{def}}{=} \prod_{i=1}^{n} W(\pi_i) \; .$$

Finally, define for each $w \in \Sigma^*$,

$$W(w) \stackrel{\text{def}}{=} \sum_{\psi \in parse_G(w)} W(\psi) \ .$$

If $parse_G(w) = \emptyset$ then $W(w) \stackrel{\text{def}}{=} 0_A$. Given a derivation sequence $\psi = \alpha_0 \Rightarrow \ldots \Rightarrow \alpha_n$ and β_0, β_1 (possibly empty) sequences of symbols and/or variables, recall that the notation $\beta_0 \psi \beta_1$ denotes the derivation sequence $\beta_0 \alpha_0 \beta_1 \Rightarrow \ldots \Rightarrow \beta_0 \alpha_n \beta_1$. When $\beta_1 \in V_\ell$ and thus $\beta_1 = \langle \beta_1' \rangle$ for some $\beta_1' \in \left(V^{\lceil k \rceil} \right)^{km+1}$, and α_i is of the form $w \langle \alpha_i' \rangle$ with $\alpha_i' \in V_\ell$ and $w \in \Sigma^*$, for some $i \in \{1, \ldots, n\}$, then $\beta_0 \alpha_i \beta_1$ denotes $\beta_0 w \langle \alpha_i' \beta_1' \rangle$.

We claim that there exists a one-to-one correspondence f that maps each LDF derivation sequence of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$ into a derivation sequence of (G_ℓ, W_ℓ) that preserves the Parikh images and the weights between derivations. Formally, there exists a one-to-one correspondence f such that for each LDF derivation sequence $\psi = X^{[d]} \Rightarrow^*_{ldf} w$ with $w \in \Sigma^*$ of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$, $f(\psi) = \langle X^{[d]} \rangle \Rightarrow^* w'$ with $w' \in \Sigma^*$ is a derivation sequence of (G_ℓ, W_ℓ) with the following properties:

1. $\wr w \wr = \wr w' \wr$ and,
2. $W^{\lceil k \rceil}(\psi) = W_\ell(f(\psi))$.

We now give an inductive definition of f. Along this definition we will prove inductively that: (i) f is an injective function from LDF derivation sequences of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$ to derivation sequences in (G_ℓ, W_ℓ); and (ii) properties 1. and 2. above hold.

Let ψ be a LDF derivation of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$.

1. If ψ is a 1-step derivation sequence then $\psi = X^{[0]} \Rightarrow^\pi_{ldf} u_0$ with $\pi \in r_0$. Then, define $f(\psi) \stackrel{\text{def}}{=} \langle X^{[0]} \rangle \Rightarrow^{\pi^\varepsilon} u_0$.
   Note that $f(\psi)$ is a one-step derivation sequence that uses the rule $(\langle X^{[0]} \rangle \to u_0) \in R_\ell$. It follows that f defines uniquely a derivation sequence of (G_ℓ, W_ℓ) for ψ. Note that property 1. holds trivially. By definition of (G_ℓ, W_ℓ), we have that:

   $$W_\ell(f(\psi)) = W_\ell(\langle X^{[0]} \rangle \to u_0) = W^{\lceil k \rceil}(X^{[0]} \to u_0) = W^{\lceil k \rceil}(\psi) \ .$$

2. If ψ is a n-step derivation sequence (with $n > 1$), then we have the following cases:
   - If $\psi = X^{[d]} \Rightarrow^\pi_{ldf} u_0 X_1^{[d]} u_1 \Rightarrow^*_{ldf} u_0 \psi' u_1$ where $\pi \in r_1$ and $\psi' = X_1^{[d]} \Rightarrow^*_{ldf} w$ with $w \in \Sigma^*$. Then, define $f(\psi) \stackrel{\text{def}}{=} \langle X^{[d]} \rangle \Rightarrow^{\pi^\varepsilon} u_0 u_1 \langle X_1^{[d]} \rangle \Rightarrow^* u_0 u_1 f(\psi')$.
     Note that the first step in the derivation $f(\psi)$ uses the rule $(\langle X^{[d]} \rangle \to u_0 u_1 \langle X_1^{[d]} \rangle) \in R_\ell$. Relying on this and the hypothesis of induction, f defines uniquely a derivation sequence of (G_ℓ, W_ℓ) for ψ. By hypothesis of induction, it is easy to check that property 1. holds. Finally, using the hypothesis of induction and the definition of (G_ℓ, W_ℓ) we have:

   $$W_\ell(f(\psi)) = W_\ell(\langle X^{[d]} \rangle \to u_0 u_1 \langle X_1^{[d]} \rangle) \cdot W_\ell(f(\psi')) = W^{\lceil k \rceil}(X^{[d]} \to u_0 X_1^{[d]} u_1) \cdot W^{\lceil k \rceil}(\psi') = W^{\lceil k \rceil}(\psi) \ .$$

   - If $\psi = X^{\lceil d \rceil} \Rightarrow^\pi_{ldf} X^{[e]} \Rightarrow^*_{ldf} \psi'$ where $\pi \in r_2$ and $\psi' = X_1^{[e]} \Rightarrow^*_{ldf} w$ with $w \in \Sigma^*$. Then, define $f(\psi) \stackrel{\text{def}}{=} \langle X^{\lceil d \rceil} \rangle \Rightarrow^{\pi^\varepsilon} \langle X^{[e]} \rangle \Rightarrow^* f(\psi')$.
     Note that the first step in the derivation $f(\psi)$ uses the rule $(\langle X^{\lceil d \rceil} \rangle \to \langle X^{[e]} \rangle) \in R_\ell$. Relying on this and the hypothesis of induction, f defines uniquely a derivation sequence of (G_ℓ, W_ℓ) for ψ. By hypothesis of induction, property 1. holds trivially. Finally, using the hypothesis of induction and the definition of (G_ℓ, W_ℓ) we have:

   $$W_\ell(f(\psi)) = W_\ell(\langle X^{\lceil d \rceil} \rangle \to \langle X^{[e]} \rangle) \cdot W_\ell(f(\psi')) = W^{\lceil k \rceil}(X^{\lceil d \rceil} \to X^{[e]}) \cdot W^{\lceil k \rceil}(\psi') = W^{\lceil k \rceil}(\psi) \ .$$

- Finally, assume w.l.o.g, that ψ has the form:

$$\begin{aligned}
\psi = X^{[d]} &\Rightarrow^{\pi}_{ldf} u_0 Z_1 u_1 \dots u_{n-2} Z_{n-1} u_{n-1} Z_n u_n \\
&\Rightarrow^*_{ldf} u_0 Z_1 u_1 \dots u_{n-2} Z_{n-1} u_{n-1} \psi'_n u_n \\
&\Rightarrow^*_{ldf} u_0 Z_1 u_1 \dots u_{n-2} \psi'_{n-1} u_{n-1} w_n u_n \Rightarrow^*_{ldf} \dots \\
&\Rightarrow^*_{ldf} u_0 w_1 u_1 w_2 u_2 \dots u_{n-1} w_n u_n \ ,
\end{aligned}$$

where $\pi \in r_2 \cup r_3$, and for each $i \in \{1, \dots, n\}$ $\psi'_i = Z_i \Rightarrow^*_{ldf} w_i$ with $w_i \in \Sigma^*$ for all i. Then, define

$$\begin{aligned}
f(\psi) &\stackrel{\text{def}}{=} \langle X^{[d]} \rangle \Rightarrow^{\pi^{\varepsilon}} u_0 u_1 u_2 \dots u_n \langle \mathcal{LDF}(Z_1 \dots Z_n) \rangle \\
&\Rightarrow^* u_0 u_1 u_2 \dots u_n f(\psi'_n) \langle \mathcal{LDF}(Z_1 \dots Z_{n-1}) \rangle \\
&\Rightarrow^* u_0 u_1 u_2 \dots u_n w'_n f(\psi'_{n-1}) \langle \mathcal{LDF}(Z_1 \dots Z_{n-2}) \rangle \Rightarrow^* \dots \\
&\Rightarrow^* u_0 u_1 u_2 \dots u_n w'_n w'_{n-1} \dots w'_1 \ ,
\end{aligned}$$

where each $w'_i \in \Sigma^*$ corresponds to the word generated by each $f(\psi'_i)$ inductively. Note that the first step in the derivation $f(\psi)$ uses a rule of the form $(X^{[d]} \to u_0 u_1 u_2 \dots u_n \langle \mathcal{LDF}(Z_1 \dots Z_n) \rangle)$ which according to the definition of (G_ℓ, W_ℓ) defines a rule in R_ℓ. Relying on this and the hypothesis of induction, f defines uniquely a derivation sequence of (G_ℓ, W_ℓ) for ψ. It is easy to see property 1. holds since, by hypothesis of induction, each w'_i satisfies $\wr w'_i \wr = \wr w_i \wr$. Finally, using the hypothesis of induction and the definition of (G_ℓ, W_ℓ) we have:

$$W_\ell(f(\psi)) = W_\ell(X^{[d]} \to u_0 u_1 \dots u_n \langle \mathcal{LDF}(Z_1 \dots Z_n) \rangle) \cdot \prod_{i=1}^{n} W_\ell(f(\psi'_i))$$

$$= W^{\lceil k \rceil}(X^{[d]} \to u_0 Z_1 u_1 \dots u_{n-1} Z_n u_n) \cdot \prod_{i=1}^{n} W^{\lceil k \rceil}(\psi'_i) = W^{\lceil k \rceil}(\psi) \ .$$

Finally, the fact that f is a surjective function follows from its construction. First, note that each rule in R_ℓ is intended to simulate a LDF derivation step of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$, while each variable in V_ℓ represents a derivation sequence of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$. On the other hand, the reader can check that, in each case, the definition of f intends to simulate a LDF derivation of $(G^{\lceil k \rceil}, W^{\lceil k \rceil})$ using the convenient definition of rules and variables of (G_ℓ, W_ℓ). It follows that every derivation sequence of (G_ℓ, W_ℓ) is covered by the image of f.

Relying on the definition of f and its properties, the following equalities hold. For each $v \in \Sigma^{\oplus}$:

$$\begin{aligned}
Pk[\![G^{\lceil k \rceil}]\!]_{W^{\lceil k \rceil}}(v) &= \sum_{v = \wr w \wr} [\![G^{\lceil k \rceil}]\!]_{W^{\lceil k \rceil}}(w) && \text{by definition of Parikh image} \\
&= \sum_{v = \wr w \wr} \sum_{\psi \in parse_{G^{\lceil k \rceil}}(w)} W^{\lceil k \rceil}(\psi) && \text{by definition of semantics} \\
&= \sum_{v = \wr w \wr} \sum_{\psi \in parse_{G^{\lceil k \rceil}}(w)} W_\ell(f(\psi)) && f \text{ satisfies prop. 2.} \\
&= \sum_{v = \wr w' \wr} \sum_{\psi' \in parse_{G_\ell}(w')} W_\ell(\psi') && f \text{ is a bijection and satisfies prop. 1.} \\
&= \sum_{v = \wr w' \wr} [\![G_\ell]\!]_{W_\ell}(w') && \text{by definition of semantics} \\
&= Pk[\![G_\ell]\!]_{W_\ell}(v). && \text{by definition of Parikh image}
\end{aligned}$$

◀

B   Counterexample in the unary case

In Example 4 we show an expansive WCFG (G_2, W_2) over the alphabet $\{a, \overline{a}\}$ that satisfies the Parikh property. Now we show, by means of the following example, that there also exists an expansive WCFG (G, W) over the alphabet $\{a\}$ that satisfies the Parikh property.

▶ **Example 28.** The idea behind this example is to use the definition of (G_2, W_2) from Example 4 (a complete definition is given in Example 11) and replace each occurrence of the alphabet symbol $\overline{a}$ in the rules of (G_2, W_2) by a. Thus, define the WCFG (G, W) where $G = (\{X, \overline{D}, D, Y, Z\}, \{a\}, X, R)$, R is given by:

$$X \to D \mid \overline{D} \qquad\qquad \overline{D} \to D\,a\,Y \mid D\,a\,Z \qquad\qquad Z \to D\,a\,Z \mid D \ ,$$
$$D \to a\,D\,a\,D \mid \varepsilon \qquad\qquad Y \to a\,Y \mid \varepsilon$$

and the weight function W is defined over $(\mathbb{N}, +, \cdot, 0, 1)$ and assigns 1 to each production in the grammar except from the rule $Y \to a\,Y$ which is assigned weight 2. Notice that we preferred to assign weight 2 to the later rule instead of adding two copies each of weight 1. Recall that $Pk[\![G_2]\!]_{W_2} = (a + \overline{a})^*$. Now, relying on our construction of (G, W), we have that $Pk[\![G]\!]_W$ is the formal power series that results from replacing each $\overline{a}$ by a in the series $Pk[\![G_2]\!]_{W_2}$. Thus, we obtain that $Pk[\![G]\!]_W = (a + a)^* = (2a)^*$. The reader can check that the formal power series $(2a)^*$ corresponds to the Parikh image of the regular WCFG (G_ℓ, W_ℓ) where G_ℓ is defined as $G_\ell = (\{X\}, \{a\}, X, \{X \to aX, X \to \varepsilon\})$ and the weight function W_ℓ is defined over $(\mathbb{N}, +, \cdot, 0, 1)$ and assigns 2 to the rule $X \to aX$ and 1 to the rule $X \to \varepsilon$.

▶ Remark. Let us check that (G, W) has the Parikh property using the decision procedure presented in Section 4. The algebraic system S corresponding to (G, W) consists of the following equations:

$$X = D + \overline{D} \qquad\qquad \overline{D} = D\,a\,Y + D\,a\,Z \qquad\qquad Z = D\,a\,Z + D \ .$$
$$D = a\,D\,a\,D + 1 \qquad\qquad Y = 2a\,Y + 1$$

Let $\sigma = (r_1, r_2, r_3, r_4, r_5)$ be its strong solution where r_1 corresponds to the solution for the initial variable X.

Now we construct the irreducible polynomial $q(X) \in \mathbb{Q}\langle\{a\}^\oplus\rangle\langle X\rangle$ following the procedure given in Theorem 14. Let $F = \{X - D - \overline{D}, \overline{D} - DaY - DaZ, Z - DaZ - D, D - aDaD - 1, Y - 2aY - 1\}$. The reduced Groebner basis[6] G of F w.r.t. lexicographic ordering is:

$$G = \left\{ X - \frac{1}{1 - 2a},\ D + \overline{D} - \frac{1}{1 - 2a},\ Z - \left(\frac{a}{1 - 2a}\right)\overline{D} - \frac{1 - 3a}{4a^2 - 4a + 1},\ Y - \frac{1}{1 - 2a}, \right.$$
$$\left. \overline{D}^2 + \left(\frac{2a^2 + 2a - 1}{2a^3 - a^2}\right)\overline{D} + \frac{5a - 2}{4a^3 - 4a^2 + a} \right\} \ .$$

Clearly, the polynomial $g \in G$ such that $g \in K\langle X\rangle$ where K is the fraction field of $\mathbb{Q}\langle\{a\}^\oplus\rangle$, and $g(r_1) \equiv 0$ is:

$$g(X) = X - \frac{1}{1 - 2a} \ .$$

---

[6] The Groebner basis G was computed using the `groebner_basis` method of the open-source mathematics software system SageMath.

This polynomial cannot be reduced into factors in the fraction field of $\mathbb{Q}\langle\{a\}^{\oplus}\rangle$. Now we multiply g by $(1-2a)$ and thus obtain $q(X) = (1-2a)X - 1$ in $\mathbb{Q}\langle\{a\}^{\oplus}\rangle\langle X\rangle$. We have that $q(X) = (1-2a)X - 1$ is the irreducible polynomial described by Theorem 9 (up to a factor in $\mathbb{Q}\langle\{a\}^{\oplus}\rangle$). Now we apply the decision procedure given in Theorem 15. We observe that q is linear in X and can be written as:

$$q(X) = (1-s)X - t = (1-2a)X - 1 \ ,$$

with $(s, \varepsilon) = 0$. Then, we conclude that (G, W) satisfies the Parikh property as expected. Let us give the regular Parikh-equivalent WCFG (G_ℓ, W_ℓ). We know that the algebraic system:

$$X = 2aX + 1 \tag{6}$$

has r_1 as solution. Then the WCFG (G_ℓ, W_ℓ) corresponding to the regular system (6) is given by $G_\ell = (\{X\}, \{a\}, R_\ell, X)$ with R_ℓ defined as:

$$\pi_1 = X \to aX$$
$$\pi_2 = X \to \varepsilon$$

and W_ℓ defined over $(\mathbb{Q}, +, \cdot, 0, 1)$ as:

$$W_\ell(\pi) = \begin{cases} 2 & \text{if } \pi = \pi_1 \\ 1 & \text{if } \pi = \pi_2 \end{cases} \ .$$

Note that (G_ℓ, W_ℓ) coincides with the regular WCFG given in the example.

C A decision procedure for the Parikh property over the rationals

Theorem 5. Let (G, W) be a cycle-free WCFG and let S be the algebraic system in commuting variables corresponding to (G, W). Then, the strong solution r of S exists and the first component of r corresponds to $Pk[\![G]\!]_W$.

Proof. First, we prove that if a WCFG (G, W) defined over a commutative and partially ordered semiring A is cycle-free, then the strong solution of the algebraic system in commuting variables corresponding to (G, W) exists. Second, we show that the first component of the strong solution corresponds to $Pk[\![G]\!]_W$.

We give a proof of the first statement by contraposition by showing the following statement: let S be the algebraic system corresponding to a WCFG (G, W) and let $\sigma^0, \sigma^1, \ldots, \sigma^j, \ldots$ be the approximation sequence associated to S. If $\lim_{j\to\infty} \sigma^j$ does not exist, then (G, W) is not cycle-free. Note that $\lim_{j\to\infty} \sigma^j$ does not exist iff either the approximation sequence oscillates between a finite number of states, or there exists a length $k \geq 0$ such that the coefficient of some monomial $v \in \Sigma^{\oplus}$, with $|v| \leq k$, increases (w.r.t. the partial ordering of A) unboundedly at every step in the approximation sequence. Formally, there exists $k \geq 0$ such that, for every $m \geq 0$, $R_k(\sigma^m) \leq R_k(\sigma^{m+j})$, for some $j > 0$, where $\leq$ is the partial ordering of A. The first case cannot hold because every approximation sequence is monotonic [15, Lemma 14.4]. Now we see that, if the second case holds, then necessarily the corresponding WCFG (G, W) is not cycle-free.

To give some intuition, let us consider the following simple scenario. Consider that the set of variables V of G contains only 1 variable, say X, and assume that the limit of the

approximation sequence of its corresponding algebraic system does not exist. Intuitively, for each $j \geq 0$, the monomials occurring in the finite series σ^j in the approximation sequence of the corresponding system S correspond to the monomials that can be produced by G in at most j derivation steps, and the coefficient of each monomial corresponds to its weight if only derivations of at most j steps are considered. By hypothesis, there exists a length $k \geq 0$ such that the coefficient of some monomial $v \in \Sigma^{\oplus}$, with $|v| \leq k$, increases unboundedly at every step in the approximation sequence. It means that, for each m-step derivation sequence π_m of G with $m \geq 1$ generating $w \in \Sigma^*$ with $\lceil w \rfloor = v$, there is another l-step derivation sequence π_l with $l > m$ such that π_l generates $w' \in \Sigma^*$ with $\lceil w' \rfloor = v$. In other words, there exist derivation sequences in G of arbitrary length. Because the number of rules of G is finite and so is the number of words $w \in \Sigma^*$ such that $\lceil w \rfloor = v$, then either G contains a rule of the form $X \to X$, or G contains a rule of the form $X \to \gamma$ with $\gamma \in \{X\}^+$ and $|\gamma| > 1$, and a rule of the form $X \to \varepsilon$. It follows that there exists a derivation sequence in G of the form $X \Rightarrow^+ X$ and thus, G is not cycle-free.

The proof of the statement for every WCFG (G, W) with an arbitrary number of variables goes in a similar fashion. By hypothesis, there exists a length $k \geq 0$ such that for some monomial $v \in \Sigma^{\oplus}$, with $|v| \leq k$, for every m-step derivation sequence π_m of G $(m \geq 1)$ generating $w \in \Sigma^*$ with $\lceil w \rfloor = v$, there is another l-step derivation sequence π_l with $l > m$ such that π_l generates $w' \in \Sigma^*$ with $\lceil w' \rfloor = v$. That is, there are arbitrarily large derivation sequences in G using rules that do not add alphabet symbols. Since the number of grammar rules of G is finite and so is the number of words $w \in \Sigma^*$ such that $\lceil w \rfloor = v$, there must exist a cycle in G, i.e., a derivation sequence of the form $X_i \Rightarrow^+ X_i$, where X_i is a variable of G. Then, we conclude that G is not cycle-free.

We have shown that the strong solution r of S exists. Now we prove that the first component of r corresponds to $Pk[\![G]\!]_W$. First, consider $\widetilde{S}$ as the algebraic system in *noncommuting* variables corresponding to a cycle-free (G, W) that is built as follows:

$$X_i = \sum_{\substack{\pi \in R \\ \pi=(X_i \to \gamma)}} W(\pi)\, \gamma \ . \tag{7}$$

Note that (7) now is of the form:

$$X_i = p_i \ , \ \text{with } p_i \in A\langle (\Sigma \cup V)^* \rangle \ .$$

Salomaa et al. prove that, if $\widetilde{r}_1$ is the first component of the strong solution of $\widetilde{S}$, then $\widetilde{r}_1(w) = [\![G]\!]_W(w)$ for every $w \in \Sigma^*$ when the weight function W is defined over $(\mathbb{N}, +, \cdot, 0, 1)$ and assigns 1 to each rule in G [20, Theorem 1.5]. In the proof they denote $[\![G]\!]_W(w)$ by $amb(G, w)$ as it corresponds to the ambiguity of w according to G. The proof for the more general case where W is any arbitrary weight function defined over a commutative semiring reduces to replacing $[\![G]\!]_W(w)$ by $amb(G, w)$ and using the corresponding semiring operations.

Now consider S as the algebraic system in commuting variables corresponding to (G, W) and built as in (1) (page 7). Let r_1 be the first component of its solution. It is known that r_1 and $\widetilde{r}_1$ verify the following equality [15]. For each $v \in \Sigma^{\oplus}$:

$$r_1(v) = \sum_{\substack{v = \lceil w \rfloor \\ w \in \Sigma^*}} \widetilde{r}_1(w) \ .$$

Then for each $v \in \Sigma^{\oplus}$:

$$Pk[\![G]\!]_W(v) = \sum_{\substack{v = \lfloor w \rceil \\ w \in \Sigma^*}} [\![G]\!]_W(w) = \sum_{\substack{v = \lfloor w \rceil \\ w \in \Sigma^*}} \widetilde{r_1}(w) = r_1(v) \ \ .$$

◀

Lemma 7.   Let (G, W) be a WCFG. Then $Pk[\![G]\!]_W \in A^{rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$ iff (G, W) satisfies the Parikh property.

Proof.   Let (G, W) be a WCFG with $Pk[\![G]\!]_W \in A^{rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$. Then $Pk[\![G]\!]_W$ is the first component of the solution of a regular algebraic system S. Hence, the WCFG (G_ℓ, W_ℓ) corresponding to S is regular and $Pk[\![G]\!]_W = Pk[\![G_\ell]\!]_{W_\ell}$.

Now, let (G, W) be a WCFG with the Parikh property. Then there exists a regular WCFG (G_ℓ, W_ℓ) such that $Pk[\![G]\!]_W = Pk[\![G_\ell]\!]_{W_\ell}$. Let S be the regular algebraic system corresponding to (G_ℓ, W_ℓ). The first component of its solution vector is $Pk[\![G_\ell]\!]_{W_\ell}$ and thus it is in $A^{rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$. As $Pk[\![G]\!]_W = Pk[\![G_\ell]\!]_{W_\ell}$ then $Pk[\![G]\!]_W$ is also in $A^{rat}\langle\langle\Sigma^{\oplus}\rangle\rangle$.   ◀

D   Unary polynomially ambiguous WCFGs are nonexpansive

Bhattiprolu et al. [2] consider the class of *polynomially ambiguous* WCFGs over the unary alphabet. They show that every WCFG in this class satisfies the Parikh property. Now we show that in the unary case the class of nonexpansive CFGs strictly contains the class of polynomially ambiguous grammars.

First, we introduce the definitions we will use in this section (some of them as given in [2]). For convenience, we will adopt the standard way to write parse trees as labeled trees where nodes are either variables or terminals (see the definition of parse tree and yield of a parse tree in Chapter 5 of [10]). Note that, for parse trees defined as in [10], the yield may contain variables (as opposed to the yield of a parse tree as defined in Section 2 which is always a (possibly empty) sequence alphabet symbols of Σ). We denote by $\mathcal{D}_G$ the set of parse trees of a grammar G when they are defined as in [10]. We denote by $\mathcal{D}_G(X)$ the set of all parse trees τ in $\mathcal{D}_G$ of the form $\tau = X(\ldots)$, i.e., rooted at variable X. A parse tree $\tau \in \mathcal{D}_G(X)$ is said to be a X-*pumping tree* if $\mathcal{Y}(\tau)\!\downarrow_V = X$. The set of all X-pumping trees is $\mathcal{D}_G^P(X) \stackrel{\text{def}}{=} \{\tau \in \mathcal{D}_G(X) \mid \mathcal{Y}(\tau)\!\downarrow_V = X\}$. The set of all pumping trees of G is given by $\mathcal{D}_G^P \stackrel{\text{def}}{=} \{\tau \in \mathcal{D}_G(X) \mid X \in V\}$. We define the concatenation of two parse trees $\tau_1, \tau_2 \in \mathcal{D}_G$ with $\mathcal{Y}(\tau_1)\!\downarrow_V \neq \varepsilon$, denoted by $\tau_1 \circ \tau_2$, by identifying the root of τ_2 with the first variable of $\mathcal{Y}(\tau_1)$. A set of trees T is *ambiguous* if there are two distinct trees τ and τ' such that $\mathcal{Y}(\tau) = \mathcal{Y}(\tau')$. Otherwise, T is *unambiguous*. They define the *ambiguity function* of a CFG G as a function $\mu_G : \mathbb{N} \mapsto \mathbb{N}$ such that $\mu_G(n) \stackrel{\text{def}}{=} max\{|parse_G(w)| : w \in \Sigma^*, |w| = n\}$, where $parse_G(w)$ denotes the set of all derivations of G that generate w. A grammar is *polynomially ambiguous* iff its ambiguity function $\mu_G(n)$ is bounded by a polynomial $p(n)$. Finally, it is known that a CFG G is polynomially ambiguous iff $\mathcal{D}_G^P$ is unambiguous [23].

Now we show the main result of this section.

▶ **Theorem 29.**   *Every unary polynomially ambiguous CFG is nonexpansive.*

Proof.   The proof goes by contradiction. Let $G = (V, \{a\}, S, R)$ be a polynomially ambiguous grammar and assume that G is expansive. Then, there is a derivation sequence of the form $X \Rightarrow^* w_0 X w_1 X w_2$ with $X \in V$ and $w_i \in (V \cup \Sigma)^*$. Assuming that every derivation sequence in G can produce a word of terminals (i.e., G does not contain useless rules),

there exist necessarily at least two distinct parse trees $\tau_1 = X(\ldots)$ and $\tau_2 = X(\ldots)$ with $\mathcal{Y}(\tau_1), \mathcal{Y}(\tau_2) \in \Sigma^*$ (not necessarily $\mathcal{Y}(\tau_1) \neq \mathcal{Y}(\tau_2)$). Let τ be the parse tree that corresponds to the derivation sequence $X \Rightarrow^* w_0\, X\, w_1\, X\, w_2$ and consider the pumping trees $\tau \circ \tau_1$ and $\tau \circ \tau_2$ with

$$\mathcal{Y}(\tau \circ \tau_1) = w_0\, \mathcal{Y}(\tau_1)\, w_1\, X\, w_2$$

$$\mathcal{Y}(\tau \circ \tau_2) = w_0\, \mathcal{Y}(\tau_2)\, w_1\, X\, w_2 \ .$$

Now define the X-pumping trees τ' and τ'' as follows:

$$\tau' = (\tau \circ \tau_1) \circ (\tau \circ \tau_2)$$

$$\tau'' = (\tau \circ \tau_2) \circ (\tau \circ \tau_1) \ .$$

Then,

$$\mathcal{Y}(\tau') = w_0\, \mathcal{Y}(\tau_1)\, w_1\, w_0\ \mathcal{Y}(\tau_2)\, w_1\, X\, w_2\, w_2$$

$$\mathcal{Y}(\tau'') = w_0\, \mathcal{Y}(\tau_2)\, w_1\, w_0\ \mathcal{Y}(\tau_1)\, w_1\, X\, w_2\, w_2 \ .$$

Because the alphabet is unary,

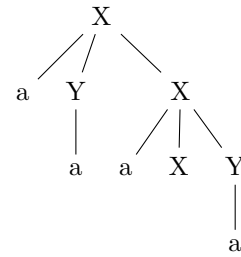$$\mathcal{Y}(\tau') = \mathcal{Y}(\tau'') \ .$$

However, $\tau' \neq \tau''$. As there exist two distinct trees in $\mathcal{D}_G^P$ with the same yield, $\mathcal{D}_G^P$ is ambiguous (contradiction). ◀

We show that the converse is not true with the following counterexample.

▶ **Example 30.** Let $G = (\{X, Y\}, \{a\}, X, \{X \to aXY,\ X \to aYX,\ X \to a,\ Y \to a\})$. Note that X produces derivation sequences with at most one occurence of itself, and Y only produces one terminal symbol. Thus, G is nonexpansive. However, there are two distinct X-pumping trees τ_1 and τ_2 with $\mathcal{Y}(\tau_1) = \mathcal{Y}(\tau_2)$ (Figures 1 and 2). Then, G is not polynomially ambiguous.
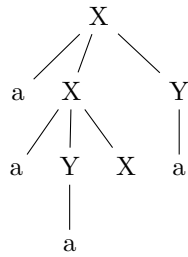
Figure 1 $\tau_1 = X(a, X(a, Y(a), X), Y(a))$

Figure 2 $\tau_2 = X(a, Y(a), X(a, X, Y(a)))$

◀