Weighted HOM-Problem for Nonnegative Integers

Andreas Maletti
Institute of Computer Science, Leipzig University, Germany

Andreea-Teodora Nász
Institute of Computer Science, Leipzig University, Germany

Erik Paul
Institute of Computer Science, Leipzig University, Germany

Abstract

The HOM-problem asks whether the image of a regular tree language under a given tree homomorphism is again regular. It was recently shown to be decidable by Godoy, Giménez, Ramos, and Àlvarez. In this paper, the ℕ-weighted version of this problem is considered and its decidability is proved. More precisely, it is decidable in polynomial time whether the image of a regular ℕ-weighted tree language under a nondeleting, nonerasing tree homomorphism is regular.

2012 ACM Subject Classification Theory of computation → Quantitative automata; Theory of computation → Computability; Theory of computation → Tree languages; Theory of computation → Grammars and context-free languages

Keywords and phrases Weighted Tree Automaton, Decision Problem, Subtree Equality Constraint, Tree Homomorphism, HOM-Problem, Weighted Tree Grammar, Weighted HOM-Problem

Digital Object Identifier 10.4230/LIPIcs.STACS.2024.51

Related Version *Previous Version*: `https://arxiv.org/abs/2311.11067`

1 Introduction

The prominent model of finite-state string automata has seen a variety of extensions in the past few decades. Notably, their qualitative evaluation was generalized to a quantitative one to yield the weighted automata of [26]. These automata are able to neatly represent process factors such as costs, consumption of resources or time, and probabilities related to the input, and have been extensively studied [25]. Semirings [17, 18] present themselves as a well suited algebraic structure for evaluating the weights because of their generality as well as their reasonable computational efficiency that is derived from distributivity.

Parallel to this development, finite-state string automata have been generalized to process other forms of inputs such as infinite words [23], graphs [3] and trees [5]. Finite-state tree automata and the *regular tree languages* they generate have been widely researched since their introduction in [7, 27, 28]. These models prove to be useful in a variety of application areas including natural language processing [19], image generation [8], and compiler construction [29]. Many applications require both features: trees as more expressive input structure and quantitative evaluation. This led to the development of several weighted tree automata (WTA) models. An extensive overview can be found in [12, Chapter 9].

Finite-state tree automata have serious limitations; notably, they cannot guarantee that two specific subtrees are always equal in the accepted trees provided that those subtrees can be arbitrarily large. Similarly finite-state string automata cannot ensure that the number of a's and b's in the accepted words is equal. These restrictions are well-known [13], and the mentioned drawback was addressed in [21], where an extension was proposed that can explicitly require certain subtrees to be equal or different. This extension is very convenient in the study of tree transformations [12] that can duplicate subtrees, and it is also the primary tool used in the seminal paper [15] to prove the decidability of the *HOM-problem*.

41st International Symposium on Theoretical Aspects of Computer Science (STACS 2024).
Editors: Olaf Beyersdorff, Mamadou Moustapha Kanté, Orna Kupferman, and Daniel Lokshtanov;
Article No. 51; pp. 51:1–51:17
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The HOM-problem, a previously long-standing open question in the study of tree languages, asks whether the image of a regular tree language under a given tree homomorphism is also regular. The image need not be regular since tree homomorphisms can also duplicate subtrees. Indeed, if this duplication ability is removed from the tree homomorphism (e.g., linear tree homomorphisms), then the image is always regular [13]. The HOM-problem was recently solved in [15, 16], where the image is represented by a tree automaton with constraints, for which it is then determined whether it generates a regular tree language. Later the HOM-problem was shown to be EXPTIME-complete [6].

In the weighted case, decidability of the HOM-problem remains wide open. Previous research on the preservation of regularity in the weighted setting [4, 9, 10, 11] focuses on cases that explicitly exclude the duplication power of the homomorphism. Recently, the weighted HOM-problem over zero-sum free semirings was addressed, but only solved for significantly restricted inputs [22]. In the present work, we prove that the HOM-problem for regular weighted tree languages over the semiring $\mathbb{N}$ of nonnegative integers can be decided in polynomial time. The proof outline is as follows: Consider such a regular $\mathbb{N}$-weighted tree language and a nondeleting, nonerasing tree homomorphism. First, we represent this image efficiently using an extension (WTGh) of weighted tree automata that permits constraints [20]. Next, we ask whether this WTGh generates a regular weighted tree language. This semantic problem is reduced to an easier, **essentially** syntactic property: the *large duplication property*. In turn, this allows us to prove decidability of the weighted HOM-problem in polynomial time by directly proving it for the large duplication property. If the WTGh for the homomorphic image does not have this property, then we give an effective construction of an equivalent $\mathbb{N}$-weighted tree automaton without constraints (albeit in exponential time), thus proving its regularity. Otherwise, we use a pumping lemma presented in [20] and isolate a strictly non-regular part from the WTGh. The most challenging part of our proof and our main technical contribution is showing that the remaining trees in the homomorphic image cannot compensate for the non-regular behavior of this isolated part. For this, we employ Ramsey's theorem [24] to identify a witness for the non-regularity of the whole weighted tree language.

Compared to the unweighted case where the HOM-problem is EXPTIME-complete [6], the $\mathbb{N}$-weighted HOM-problem can be decided in polynomial time. Both proofs reduce the (non)regularity of the homomorphic image in question to a decidable property of the tree grammar with constraints representing it; however, the unweighted regularity notion is very different from the corresponding notion for weighted tree languages over $\mathbb{N}$. Unlike the Boolean semiring $\mathbb{B}$, which corresponds to the unweighted case, the semiring $\mathbb{N}$ can be embedded into a field, which allows us to apply methods of linear algebra. The *large duplication property*, to which we successfully reduce the $\mathbb{N}$-weighted HOM-problem, is certainly necessary, but insufficient in the unweighted case. This is due to the fact that the Boolean semiring is idempotent, which permits covering an irregular tree language with the help of a regular one (e.g., the union $L \cup T_\Sigma = T_\Sigma$ of an irregular tree language L with the regular tree language T_Σ of all trees is again regular). We will prove that such covers cannot happen in the semiring $\mathbb{N}$ of nonnegative integers. As a consequence, the *large duplication property* is necessary and sufficient for non-regularity of weighted tree languages over $\mathbb{N}$. In summary, our overall strategy for approaching the HOM-problem is similar to [15], but the required notions and details of the proofs significantly differ.

Tree automata and grammars with constraints are applied in domains such as automated deduction [2] or security verification [1]. In this context, studying quantitative extensions of these models is naturally relevant. Tree structures are also central in XML, and homomorphic transformations on trees allow us to modify the codes while preserving the hierarchical structure. Moreover, the HOM-problem plays a role in the context of term rewriting [14]: For a term rewrite system, the set of normal forms (i.e., terms to which no rule can be applied) can be expressed as the complement of a homomorphic image; a better understanding of these images can help generalize known results in this field.

2 Preliminaries

We denote the set of nonnegative integers by $\mathbb{N}$. For $i, j \in \mathbb{N}$ we let $[i, j] = \{k \in \mathbb{N} \mid i \leq k \leq j\}$ and $[j] = [1, j]$. Let Z be an arbitrary set. The cardinality of Z is denoted by $|Z|$, and the set of words over Z (i.e., the set of ordered finite sequences of elements of Z) is denoted by Z^*.

Trees, Substitutions, and Contexts

A *ranked alphabet* (Σ, rk) consists of a finite set Σ and a mapping $\mathrm{rk}\colon \Sigma \to \mathbb{N}$ that assigns a rank to each symbol of Σ. If there is no risk of confusion, then we denote the ranked alphabet (Σ, rk) by Σ alone. We write $\sigma^{(k)}$ to indicate that $\mathrm{rk}(\sigma) = k$. Moreover, for every $k \in \mathbb{N}$ we let $\Sigma_k = \mathrm{rk}^{-1}(k)$ and $\mathrm{rk}(\Sigma) = \max \{k \in \mathbb{N} \mid \Sigma_k \neq \emptyset\}$ be the maximal rank of symbols of Σ. Let $X = \{x_i \mid i \in \mathbb{N}\}$ be a countable set of (formal) variables. For every $n \in \mathbb{N}$, we let $X_n = \{x_i \mid i \in [n]\}$. Given a ranked alphabet Σ and a set Z, the set $T_\Sigma(Z)$ of *Σ-trees indexed by Z* is the smallest set such that $Z \subseteq T_\Sigma(Z)$ and $\sigma(t_1, \ldots, t_k) \in T_\Sigma(Z)$ for every $k \in \mathbb{N}$, $\sigma \in \Sigma_k$, and $t_1, \ldots, t_k \in T_\Sigma(Z)$. We abbreviate $T_\Sigma(\emptyset)$ simply by T_Σ, and any subset $L \subseteq T_\Sigma$ is called a *tree language*.

Let Σ be a ranked alphabet, Z a set, and consider a tree $t \in T_\Sigma(Z)$. The set $\mathrm{pos}(t)$ of *positions of t* is defined by $\mathrm{pos}(z) = \{\varepsilon\}$ for all $z \in Z$ and by $\mathrm{pos}(\sigma(t_1, \ldots, t_k)) = \{\varepsilon\} \cup \{iw \mid i \in [k], w \in \mathrm{pos}(t_i)\}$ for all $k \in \mathbb{N}$, $\sigma \in \Sigma_k$, and $t_1, \ldots, t_k \in T_\Sigma(Z)$. With their help, we define the *size* 'size(t)' and *height* 'ht(t)' of t as $\mathrm{size}(t) = |\mathrm{pos}(t)|$ and $\mathrm{ht}(t) = \max_{w \in \mathrm{pos}(t)} |w|$. Positions are partially ordered by the standard prefix order $\leq$ on $[\mathrm{rk}(\Sigma)]^*$, and they are totally ordered by the ascending lexicographic order $\preceq$ on $[\mathrm{rk}(\Sigma)]^*$, in which prefixes are larger; i.e., ε is the largest element. More precisely, for $v, w \in \mathrm{pos}(t)$ if there exists $u \in [\mathrm{rk}(\Sigma)]^*$ with $vu = w$, then we write $v \leq w$, call v a *prefix* of w, and let $v^{-1}w = u$ because u is uniquely determined if it exists. Provided that $u = u_1 \cdots u_n$ with $u_1, \ldots, u_n \in [\mathrm{rk}(\Sigma)]$ we also define the *path* $[v, \ldots, w]$ *from v to w* as the sequence $(v, vu_1, vu_1u_2, \ldots, w)$ of positions. Any two positions that are $\leq$-incomparable are called *parallel*.

Given $t, t' \in T_\Sigma(Z)$ and $w \in \mathrm{pos}(t)$, the *label* $t(w)$ of t at w, the *subtree* $t|_w$ of t at w, and the *substitution* $t[t']_w$ of t' into t at w are defined by $z(\varepsilon) = z|_\varepsilon = z$ and $z[t']_\varepsilon = t'$ for all $z \in Z$ and by $t(\varepsilon) = \sigma$, $t(iw') = t_i(w')$, $t|_\varepsilon = t$, $t|_{iw'} = t_i|_{w'}$, $t[t']_\varepsilon = t'$, and $t[t']_{iw'} = \sigma\big(t_1, \ldots, t_{i-1}, t_i[t']_{w'}, t_{i+1}, \ldots, t_k\big)$ for all trees $t = \sigma(t_1, \ldots, t_k)$ with $k \in \mathbb{N}$, $\sigma \in \Sigma_k$, $t_1, \ldots, t_k \in T_\Sigma(Z)$, all $i \in [k]$, and all $w' \in \mathrm{pos}(t_i)$. For all sets $S \subseteq \Sigma \cup Z$ of symbols, we let $\mathrm{pos}_S(t) = \{w \in \mathrm{pos}(t) \mid t(w) \in S\}$, and we write $\mathrm{pos}_s(t)$ instead of $\mathrm{pos}_{\{s\}}(t)$ for every $s \in \Sigma \cup Z$. The set of variables occuring in t is $\mathrm{var}(t) = \{x \in X \mid \mathrm{pos}_x(t) \neq \emptyset\}$. Finally, consider $n \in \mathbb{N}$ and a mapping $\theta'\colon X_n \to T_\Sigma(Z)$. Then by substitution, θ' induces a mapping $\theta\colon T_\Sigma(Z) \to T_\Sigma(Z)$ defined by $\theta(x) = \theta'(x)$ for every $x \in X_n$, $\theta(z) = z$ for every $z \in Z \setminus X_n$, and $\theta(\sigma(t_1, \ldots, t_k)) = \sigma(\theta(t_1), \ldots, \theta(t_k))$ for all $k \in \mathbb{N}$, $\sigma \in \Sigma_k$, and $t_1, \ldots, t_k \in T_\Sigma(Z)$. For $t \in T_\Sigma(Z)$, we denote $\theta(t)$ by $t\theta$ or, more commonly, by $t[x_1 \leftarrow \theta'(x_1), \ldots, x_n \leftarrow \theta'(x_n)]$.

Let $\square \notin \Sigma$. A *context* is a tree $C \in T_\Sigma(\square)$ with $\mathrm{pos}_\square(C) \neq \emptyset$. More specifically, we call C an *n-context* if $n = |\mathrm{pos}_\square(C)|$. For an n-context C and $t_1, \ldots, t_n \in T_\Sigma$, we define the substitution $C[t_1, \ldots, t_n]$ as follows. Let $\mathrm{pos}_\square(C) = \{w_1, \ldots, w_n\}$ be the occurrences of $\square$ in C in lexicographic order $w_1 \prec \cdots \prec w_n$. Then we let $C[t_1, \ldots, t_n] = C[t_1]_{w_1} \cdots [t_n]_{w_n}$.

Tree Homomorphisms and Weighted Tree Grammars

Given ranked alphabets Σ and Γ, let $h' \colon \Sigma \to T_\Gamma(X)$ be a mapping with $h'(\sigma) \in T_\Gamma(X_k)$ for all $k \in \mathbb{N}$ and $\sigma \in \Sigma_k$. We extend h' to $h \colon T_\Sigma \to T_\Gamma$ by $h(\alpha) = h'(\alpha) \in T_\Gamma(X_0) = T_\Gamma$ for all $\alpha \in \Sigma_0$ and $h(\sigma(t_1, \ldots, t_k)) = h'(\sigma)[x_1 \leftarrow h(t_1), \ldots, x_k \leftarrow h(t_k)]$ for all $k \in \mathbb{N}$, $\sigma \in \Sigma_k$, and $t_1, \ldots, t_k \in T_\Sigma$. The mapping h is called the *tree homomorphism induced by* h', and we identify h' and its induced tree homomorphism h. For the complexity analysis of our decision procedure, we define the size of h as $\mathrm{size}(h) = \sum_{\sigma \in \Sigma} \mathrm{size}(h(\sigma))$. We call h *nonerasing* (respectively, *nondeleting*) if $h'(\sigma) \notin X$ (respectively, $\mathrm{var}(h'(\sigma)) = X_k$) for all $k \in \mathbb{N}$ and $\sigma \in \Sigma_k$. In this contribution, we will only consider nonerasing and nondeleting tree homomorphisms $h \colon T_\Sigma \to T_\Gamma$, which are therefore *input finitary*; i.e., the preimage $h^{-1}(u)$ is finite for every $u \in T_\Gamma$ since $|t| \leq |u|$ for every $t \in h^{-1}(u)$. Any mapping $A \colon T_\Sigma \to \mathbb{N}$ is called $\mathbb{N}$-*weighted tree language*, and we define the weighted tree language $h_A \colon T_\Gamma \to \mathbb{N}$ for every $u \in T_\Gamma$ by $h_A(u) = \sum_{t \in h^{-1}(u)} A(t)$ and call it the *image of A under h*. This definition relies on the tree homomorphism to be input-finitary; otherwise the defining sum is not finite, so the value $h_A(u)$ is not necessarily well-defined.

A *weighted tree grammar with equality constraints* (WTGc) [20] is a tuple $(Q, \Sigma, F, P, \mathrm{wt})$, in which Q is a finite set of *states*, Σ is a ranked alphabet of *input symbols*, $F \colon Q \to \mathbb{N}$ assigns a *final weight* to every state, P is a finite set of *productions* of the form (ℓ, q, E) with $\ell \in T_\Sigma(Q) \setminus Q$, $q \in Q$, and finite subset $E \subseteq \mathbb{N}^* \times \mathbb{N}^*$, and $\mathrm{wt} \colon P \to \mathbb{N}$ assigns a *weight* to every production. A production $p = (\ell, q, E) \in P$ is usually written $p = \ell \xrightarrow{E} q$ or $p = \ell \xrightarrow{E}_{\mathrm{wt}(p)} q$, and the tree ℓ is called its *left-hand side*, q is its *target state*, and E are its *equality constraints*, respectively. Equality constraints $(v, v') \in E$ are also written as $v = v'$. A state $q \in Q$ is *final* if $F(q) \neq 0$.

Next, we recall the *derivation semantics* of WTGc from [20]. Let $(v, v') \in \mathbb{N}^* \times \mathbb{N}^*$ be an equality constraint and $t \in T_\Sigma$. The tree t satisfies (v, v') if and only if $v, v' \in \mathrm{pos}(t)$ and $t|_v = t|_{v'}$, and for a finite set $C \subseteq \mathbb{N}^* \times \mathbb{N}^*$ of equality constraints, we write $t \models C$ if t satisfies all $(v, v') \in C$. Let $G = (Q, \Sigma, F, P, \mathrm{wt})$ be a WTGc. A *sentential form (for G)* is a tree $\xi \in T_\Sigma(Q)$. Given an input tree $t \in T_\Sigma$, sentential forms $\xi, \zeta \in T_\Sigma(Q)$, a production $p = \ell \xrightarrow{E} q \in P$, and a position $w \in \mathrm{pos}(\xi)$, we write $\xi \Rightarrow_{G,t}^{p,w} \zeta$ if $\xi|_w = \ell$, $\zeta = \xi[q]_w$, and $t|_w \models E$; i.e., the equality constraints E are fulfilled on $t|_w$. A sequence $d = (p_1, w_1) \cdots (p_n, w_n) \in (P \times \mathbb{N}^*)^*$ is a *derivation (of G) for t* if there exist $\xi_0, \ldots, \xi_n \in T_\Sigma(Q)$ such that $\xi_0 = t$ and $\xi_{i-1} \Rightarrow_{G,t}^{p_i,w_i} \xi_i$ for all $i \in [n]$. We call d *left-most* if additionally $w_1 \prec w_2 \prec \cdots \prec w_n$. Note that the sentential forms $\xi_0, \ldots, \xi_n$ are uniquely determined if they exist, and for any derivation d for t there exists a unique permutation of d that is a left-most derivation for t. We call d *complete* if $\xi_n \in Q$, and in this case we also call it a derivation *to ξ_n*. The set of all complete left-most derivations for t to $q \in Q$ is denoted by $D_G^q(t)$. A complete derivation to some final state is called *accepting*. If for every $p \in P$, there exists a tree $t \in T_\Sigma$, a final state q and a derivation $(p_1, w_1) \cdots (p_m, w_m) \in D_G^q(t)$ such that $F(q) \cdot \prod_{i=1}^n \mathrm{wt}(p_i) \neq 0$ and $p \in \{p_1, \ldots, p_m\}$; i.e. if every production is used in an accepting derivation with nonzero weights, then G is *trim*.

Let $d = (p_1, w_1) \cdots (p_n, w_n) \in D_G^q(t)$ for some $t \in T_\Sigma$ and $i \in [n]$. Moreover, let $\{j_1, \ldots, j_\ell\}$ be the set $\{j \in [n] \mid w_i \leq w_j\}$ with the indices $j_1 < \cdots < j_\ell$ of those positions of which w_i is a prefix. We refer to $(p_{j_1}, w_i^{-1} w_{j_1}), \ldots, (p_{j_\ell}, w_i^{-1} w_{j_\ell})$ as the *derivation for $t|_{w_i}$ incorporated in d*. Conversely, for $w \in \mathbb{N}^*$ we abbreviate the derivation $(p_1, ww_1) \cdots (p_n, ww_n)$ by wd.

The *weight* of a derivation $d = (p_1, w_1) \cdots (p_n, w_n)$ is defined as $\mathrm{wt}_G(d) = \prod_{i=1}^n \mathrm{wt}(p_i)$. The weighted tree language generated by G, written $[\![G]\!] \colon T_\Sigma \to \mathbb{N}$, is defined for all $t \in T_\Sigma$ by $[\![G]\!](t) = \sum_{q \in Q,\, d \in D_G^q(t)} F(q) \cdot \mathrm{wt}_G(d)$. For $t \in T_\Sigma$ and $q \in Q$, we will often use the value $\mathrm{wt}_G^q(t)$ defined as $\mathrm{wt}_G^q(t) = \sum_{d \in D_G^q(t)} \mathrm{wt}_G(d)$. Using distributivity, $[\![G]\!](t)$ then simplifies to $[\![G]\!](t) = \sum_{q \in Q} F(q) \cdot \mathrm{wt}_G^q(t)$. We call two WTGc *equivalent* if they generate the same weighted tree language.

We call a WTGc $(Q, \Sigma, F, P, \mathrm{wt})$ a *weighted tree grammar* (WTG) if $E = \emptyset$ for every production $\ell \xrightarrow{E} q \in P$; i.e., no production utilizes equality constraints. Instead of $\ell \xrightarrow{\emptyset} q$ we also simply write $\ell \to q$. Moreover, we call a WTGc a *weighted tree automaton with equality constraints* (WTAc) if $\mathrm{pos}_\Sigma(\ell) = \{\varepsilon\}$ for every production $\ell \xrightarrow{E} q \in P$, and a *weighted tree automaton* (WTA) if it is both a WTG and a WTAc. The classes of WTGc and WTAc are equally expressive, and they are strictly more expressive than the class of WTA [20]. We call a weighted tree language *regular* if it is generated by a WTA and *constraint-regular* if it is generated by a WTGc. Productions with weight 0 are obviously useless, so we may assume that $\mathrm{wt}(p) \neq 0$ for every production p. Finally, we define the size of a WTGc as follows.

▶ **Definition 1.** *Let $G = (Q, \Sigma, F, P, \mathrm{wt})$ be a WTGc and $p = \ell \xrightarrow{E} q \in P$ be a production. We define the* height *of p as $\mathrm{ht}(p) = \mathrm{ht}(\ell)$ and its* size *as $\mathrm{size}(p) = \mathrm{size}(\ell)$, the* height *of P as $\mathrm{ht}(P) = \max_{p \in P} \mathrm{ht}(p)$ and its* size *as $\mathrm{size}(P) = \sum_{p \in P} \mathrm{size}(p)$, and finally the* height *of G as $\mathrm{ht}(G) = |Q| \cdot \mathrm{ht}(P)$ and its* size *as $\mathrm{size}(G) = |Q| + \mathrm{size}(P)$.*

It is known [20] that WTGc of a particular shape can represent homomorphic images of regular weighted tree languages. This subclass of WTGc will be central in our work.

▶ **Definition 2.** *A WTGc $\big(Q, \Sigma, F, P, \mathrm{wt}\big)$ is* classic *if every production $p = \ell \xrightarrow{E} q \in P$ satisfies $E \subseteq \mathrm{pos}_Q(\ell)^2$; i.e., all equality constraints point to the Q-labeled positions of its left-hand side. Without loss of generality, we can assume that every set E of equality constraints is reflexive, symmetric, and transitive, that is, an equivalence relation on a subset $D \subseteq \mathrm{pos}_Q(\ell)$, so not all occurrences of states need to be constrained.*

A classic WTGc is eq-restricted *if it has a so-called* sink state *$\perp \in Q \setminus F$ such that (i) $\sigma(\perp, \dots, \perp) \to_1 \perp$ belongs to P for all $\sigma \in \Sigma$, and no other productions target $\perp$, and (ii) for every production $\ell \xrightarrow{E} q$ with $q \neq \perp$, if $\mathrm{pos}_Q(\ell) = \{p_1, \dots, p_n\}$ and $q_i = \ell(p_i)$ for $i \in [n]$, the following conditions hold:*

1. *For each $i \in [n]$, the set $\{q_j \mid p_j \in [p_i]_{\equiv_E}\} \setminus \{\perp\}$ is a singleton.*
2. *There exists exactly one $p_j \in [p_i]_{\equiv_E}$ such that $q_j \neq \perp$.*

In other words, an eq-restricted WTGc G has a designated nonfinal sink state $\perp \in Q$ such that $F(\perp) = 0$ as well as $p_\gamma = \gamma(\perp, \dots, \perp) \to \perp \in P$ and $\mathrm{wt}(p_\gamma) = 1$ for every $\gamma \in \Gamma$. In addition, every production $p = \ell \xrightarrow{E} q \in P$ satisfies the following two properties. First, $E \subseteq \mathrm{pos}_Q(\ell)^2$; i.e., all equality constraints point to the Q-labeled positions of its left-hand side. Second, $\ell(v) = \perp$ and $\ell(w) \neq \perp$ for every $v \in [w']_E \setminus \{w\}$ and w' occurring in E, where $w = \min_\preceq [w']_E$; i.e., all but the lexicographically least position in each equivalence class of E are guarded by state $\perp$. Essentially, an eq-restricted WTGc G performs its checks (and charges weights) exclusively on the lexicographically least occurrences of equality-constrained subtrees. All the other subtrees, which by means of the constraint are forced to coincide with another subtree, are simply ignored by the WTGc, which formally means that they are processed in the designated sink state $\perp$. In the following, we will use $\perp$ to denote such a sink state, and write $Q \cup \{\perp\}$ to explicitly indicate its presence.

To simplify our terminology, we will refer to eq-restricted WTGc simply as *WTGh*.

▶ **Theorem 3** (see [20, Theorem 5])**.** *Let $G = (Q, \Sigma, F, P, \mathrm{wt})$ be a trim WTA and $h \colon T_\Sigma \to T_\Gamma$ be a nondeleting and nonerasing tree homomorphism. Then there exists a trim WTGh G' with $[\![G']\!] = h_{[\![G]\!]}$. Moreover, $\mathrm{size}(G') \in \mathcal{O}\big(\mathrm{size}(G) \cdot \mathrm{size}(h)\big)$ and $\mathrm{ht}(G') \in \mathcal{O}\big(\mathrm{size}(h)\big)$.*

▶ **Example 4.** Let $G = (Q \cup \{\bot\}, \Gamma, F, P, \mathrm{wt})$ with $Q = \{q, q_f\}$, $\Gamma = \{\alpha^{(0)}, \gamma^{(1)}, \delta^{(3)}\}$, $F(q) = F(\bot) = 0$ and $F(q_f) = 1$, and the following set P of productions.

$$\Big\{\alpha \to_1 q,\ \gamma(q) \to_2 q,\ \delta\big(q, \gamma(\bot), q\big) \xrightarrow{1=21}_1 q_f, \quad \alpha \to_1 \bot,\ \gamma(\bot) \to_1 \bot,\ \delta(\bot, \bot, \bot) \to_1 \bot\Big\}$$

The WTGc G is a WTGh. It generates the homomorphic image $[\![G]\!] = h_A$ for the tree homomorphism h induced by the mapping $\alpha \mapsto \alpha$, $\gamma \mapsto \gamma(x_1)$, and $\sigma \mapsto \delta\big(x_2, \gamma(x_2), x_1\big)$ applied to the regular weighted tree language $A \colon T_\Sigma \to \mathbb{N}$ given by $A(t) = 2^{|\mathrm{pos}_\gamma(t)|}$ for every $t \in T_\Sigma$ with $\Sigma = \{\alpha^{(0)}, \gamma^{(1)}, \sigma^{(2)}\}$. The weighted tree language $[\![G]\!]$ is itself not regular because its support is clearly not a regular tree language.

The restrictions in the definition of a WTGh allow us to trim it effectively.

▶ **Lemma 5.** *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a WTGh. An equivalent, trim WTGh G' can be constructed in polynomial time.*

Proof. First, recall that we may assume $\mathrm{wt}(p) \neq 0$ for every $p \in P$ because $\mathrm{wt}_G(d) = 0$ for every derivation d of G that contains a production p with $\mathrm{wt}(p) = 0$. For the proof, we employ a simple reachability algorithm. For every $n \in \mathbb{N}$ and $U \subseteq Q$, let

$$Q_0 = \emptyset \qquad Q_{n+1} = Q_n \cup \bigcup_{\substack{(\ell \xrightarrow{E} q) \in P \\ \ell \in T_\Sigma(Q_n)}} \{q\} \qquad \Pi_U = \bigcup_{\substack{(\ell \xrightarrow{E} q) \in P \\ \ell \in T_\Sigma(U)}} \big\{(q, q') \in U^2 \mid \mathrm{pos}_{q'}(\ell) \neq \emptyset\big\} \ .$$

Since Q is finite, there exists N with $Q_N = Q_{N+1}$. Let $Q' = Q_N$. A straightforward proof shows that $q \in Q'$ if and only if for some $t \in T_\Sigma$ there exists $d \in D^q_G(t)$ with $\mathrm{wt}_G(d) \neq 0$. To ensure the reachability of a final state, we let $\lhd$ be the smallest reflexive and transitive relation on Q' that contains $\Pi_{Q'}$. Then $P' = \{\ell \xrightarrow{E} q \in P \mid q \in Q',\ \exists q_{\mathrm{f}} \in Q' \colon F(q_{\mathrm{f}}) \neq 0,\ q_{\mathrm{f}} \lhd q\}$, and the desired WTGh is simply $G' = (Q \cup \{\bot\}, \Sigma, F, P', \mathrm{wt}|_{P'})$. ◀

3 Substitutions in the Presence of Equality Constraints

This short section recalls from [20] some definitions together with a pumping lemma for WTGh, which will be essential for deciding the integer-weighted HOM-problem. First, we need to refine the substitution of trees such that it complies with existing constraints.

▶ **Definition 6** (see [20] and cf. [15]). *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a WTGh. Moreover, let $q, q' \in Q$, $t, t' \in T_\Sigma$, and $d \in D^q_G(t)$ as well as $d' \in D^{q'}_G(t')$ such that $q \neq \bot \neq q'$ and $d = \underline{d}(p, \varepsilon)$ uses $p = c[q_1, \ldots, q_k] \xrightarrow{E,\emptyset} q \in P$ as its final production. For every $i \in [k]$ let $w_i = \mathrm{pos}_{x_i}(c)$ and d_i be the unique left-most derivation for $t_i = t|_{\mathrm{pos}_{x_i}(c)}$ incorporated in d. Finally, for every $u \in T_\Sigma$ let $d^\bot_u$ be the unique left-most derivation for u to $\bot$. For every $w \in \mathrm{pos}(t)$ at which the production used in d targets q', we recursively define the derivation substitution $d[\![d']\!]_w$ of d' into d at w and the resulting tree $t[\![t']\!]^d_w$ as follows. If $w = \varepsilon$, then $d[\![d']\!]_\varepsilon = d'$ and $t[\![t']\!]^d_\varepsilon = t'$. Otherwise $w = w_j\underline{w}$ for some $j \in [k]$ and we have*

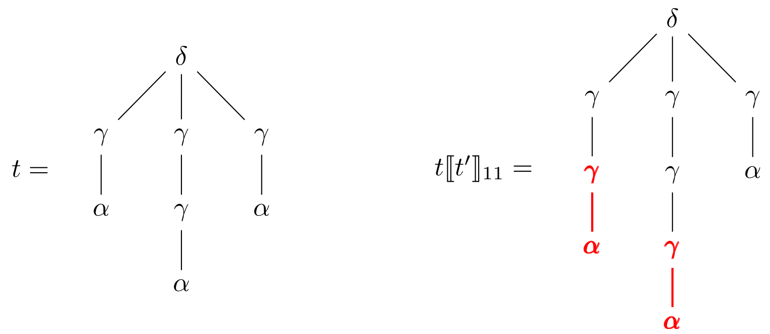
$$d[\![d']\!]_w = d'_1 \cdots d'_k(p, \varepsilon) \qquad \text{and} \qquad t[\![t']\!]^d_w = c[t'_1, \ldots, t'_k] \ ,$$

where for each $i \in [k]$ we have

- *if $i = j$ (i.e., w_i is a prefix of w), then $d'_i = w_i(d_i[\![d']\!]_{\underline{w}})$ and $t'_i = t_i[\![t']\!]^{d'_i}_{\underline{w}}$,*
- *if $q_i = \bot$ and $w_i \in [w_j]_{\equiv_E}$ (i.e., it is a position that is equality restricted to w_j), then $d'_i = w_i d^\bot_u$ and $t'_i = u$ with $u = t_j[\![t']\!]^{d'_j}_{\underline{w}}$, and*
- *otherwise $d'_i = w_i d_i$ and $t'_i = t_i$ (i.e., derivation and tree remain unchanged).*

It is straightforward to verify that $d[\![d']\!]_w$ is a complete left-most derivation for $t[\![t']\!]^d_w$ to q.

▶ **Example 7.** Consider the WTGh G of Example 4 and the following tree t it generates into which we want to substitute the tree $t' = \gamma(\alpha)$ at position $w = 11$.



We consider the following complete left-most derivation for t to q_f.

$$d = \Big(\alpha \to q, 11\Big)\Big(\gamma(q) \to q, 1\Big) \quad \Big(\alpha \to \bot, 211\Big)\Big(\gamma(\bot) \to \bot, 21\Big)$$
$$\Big(\alpha \to q, 31\Big)\Big(\gamma(q) \to q, 3\Big)\Big(\delta\big(q, \gamma(\bot), q\big) \xrightarrow{1=21} q_f, \varepsilon\Big)$$

Moreover, let $d' = (\alpha \to q, 1)\,\big(\gamma(q) \to q, \varepsilon\big)$ and $d'_\bot = (\alpha \to \bot, 1)\,\big(\gamma(\bot) \to \bot, \varepsilon\big)$. With the notation of Definition 6, in the first step we have $v_1 = 1$, $v_2 = 21$, $v_3 = 3$, $d_1 = d_3 = d'$, $d_2 = d'_\bot$, and $\hat{w} = v_1^{-1}w = 1$. Respecting the only constraint $1 = 21$, we set $d'_1 = d_1[\![d']\!]_{\hat{w}} = d'[\![d']\!]_1$, $d'_2 = d_2[\![d'_\bot]\!]_{\hat{w}} = d'_\bot[\![d'_\bot]\!]_1$, and $d'_3 = d_3 = d'$. Eventually, $d'_1 = (\alpha \to q, 11)(\gamma(q) \to q, 1)(\gamma(q) \to q, \varepsilon)$ and $d'_2 = (\alpha \to \bot, 11)(\gamma(\bot) \to \bot, 1)(\gamma(\bot) \to \bot, \varepsilon)$. Hence, we obtain the following derivation $d[\![d']\!]_{11}$ for our new tree $t[\![t']\!]_{11}$.

$$d[\![d']\!]_{11} = \Big(\alpha \to q, 111\Big)\Big(\gamma(q) \to q, 11\Big)\Big(\gamma(q) \to q, 1\Big)\Big(\alpha \to \bot, 2111\Big)\Big(\gamma(\bot) \to \bot, 211\Big)$$
$$\Big(\gamma(\bot) \to \bot, 21\Big) \quad \Big(\alpha \to q, 31\Big)\Big(\gamma(q) \to q, 3\Big)\Big(\delta\big(q, \gamma(\bot), q\big) \xrightarrow{1=21} q_f, \varepsilon\Big)$$

Although $t|_{31} = \alpha$ also coincides with the subtree $t|_{11} = \alpha$ we replaced, these two subtrees are not equality-constrained, so the simultaneous substitution does not affect $t|_{31}$.

The substitution of Definition 6 allows us to prove a pumping lemma for the class of WTGh: If d is an accepting derivation of a WTGh $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ for a tree t with $\mathrm{ht}(t) > \mathrm{ht}(G)$, then there exist at least $|Q \setminus \{\bot\}| + 1$ positions $w_1 > \cdots > w_{|Q|+1}$ in t at which d applies productions with non-sink target states. By the pigeonhole principle, there thus exist two positions $w_i > w_j$ in t at which d applies productions with the same non-sink target state. Employing the substitution we just defined, we can substitute $t|_{w_j}$ into w_i and obtain a derivation of G for $t[\![t|_{w_j}]\!]_{w_i}$. This process can be repeated to obtain an infinite sequence of trees strictly increasing in size. Formally, the following lemma was proved in [20].

▶ **Lemma 8** ([20, Lemma 4]). *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a WTGh. Consider some tree $t \in T_\Sigma$ and non-sink state $q \in Q \setminus \{\bot\}$ such that $\mathrm{ht}(t) > \mathrm{ht}(G)$ and $D_G^q(t) \neq \emptyset$. Then there are infinitely many pairwise distinct trees $t_0, t_1, \ldots$ such that $D_G^q(t_i) \neq \emptyset$ for all $i \in \mathbb{N}$.*

▶ **Example 9.** Recall the WTGh G of Example 4. We have $\mathrm{ht}(P) = 2$ and $\mathrm{ht}(G) = 4$, but for simplicity, we choose the smaller tree $t = \delta(\gamma(\alpha), \gamma(\gamma(\alpha)), \gamma(\alpha))$, which we also considered in Example 7, since it also allows pumping. The derivation d presented in Example 7 for t applies the productions $(\alpha \to q)$ at 11 and $\gamma(q) \to q$ at 1, so we substitute $t|_1 = \gamma(\alpha)$ at 11 to obtain $t[\![\gamma(\alpha)]\!]_{11}$. In fact, this is exactly the substitution we illustrated in Example 7.

4 The Decision Procedure

Let us now turn to the $\mathbb{N}$-weighted version of the HOM-problem. In the following, we show that the regularity of the homomorphic image of a regular $\mathbb{N}$-weighted tree language is decidable in polynomial time. More precisely, we prove the following theorem.

▶ **Theorem 10.** *The weighted HOM-problem over $\mathbb{N}$ is polynomial; i.e. for fixed ranked alphabets Γ and Σ, given a trim WTA A over Γ, and a nondeleting, nonerasing tree homomorphism $h\colon T_\Gamma \to T_\Sigma$, it is decidable in polynomial time whether $h_{[\![A]\!]}$ is regular.*

In the beginning, the proof of Theorem 10 resembles the unweighted case [15]: Given a regular weighted tree language A (represented by a trim WTA) and a tree homomorphism h, we first construct a trim WTGh G for its image $[\![G]\!] = h_A$ applying Theorem 3. We then show that $[\![G]\!]$ is regular if and only if the equality constraints used in G only act on subtrees of height at most $\operatorname{ht}(G)$. In other words, if there exists a production $\ell \xrightarrow{E} q$ in G such that for some equality constraint $(u, v) \in E$ with non-sink state $q = \ell(u)$ there exists a tree $t \in T_\Sigma$ with $\operatorname{ht}(t) > \operatorname{ht}(G)$ and $D^q_G(t) \neq \emptyset$, then $[\![G]\!]$ is not regular, and if no such production exists, then $[\![G]\!]$ is regular. There are thus three parts to our proof. First, we show that the existence of such a production is decidable in polynomial time. Then we show that $[\![G]\!]$ is regular if no such production exists. Finally, we show that $[\![G]\!]$ is not regular if such a production exists. The latter part employs Ramsey's theorem [24] and is the most significant technical contribution in our paper. For convenience, we attach a name to the property described here.

▶ **Definition 11.** *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a trim WTGh. We say that G has the* large duplication property *if there exist a production $\ell \xrightarrow{E} q \in P$, an equality constraint $(u, v) \in E$ with $\ell(u) \neq \bot = \ell(v)$, and a tree $t \in T_\Sigma$ such that $\operatorname{ht}(t) > \operatorname{ht}(G)$ and $D^{\ell(u)}_G(t) \neq \emptyset$.*

We start with the decidability of the large duplication property.

▶ **Lemma 12.** *Consider a fixed ranked alphabet Σ. The following is decidable in polynomial time: Given a trim WTGh G, does it satisfy the large duplication property?*

Proof. Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ and construct the directed graph $G = (Q, E)$ with edges $E = \bigcup_{\ell \xrightarrow{E} q \in P} \{(q', q) \mid q' \in Q, \operatorname{pos}_{q'}(\ell) \neq \emptyset\}$. Clearly, the large duplication property is equivalent to the condition that there exists a production $\ell \xrightarrow{E} q \in P$, an equality constraint $(u, v) \in E$ with $\ell(u) \neq \bot = \ell(v)$, and a state $q' \in Q \setminus \{\bot\}$ such that there exists a cycle from q' to q' in G and a path from q' to $\ell(u)$ in G. This equivalent condition can be checked in polynomial time. The equivalence of the two statements is easy to establish. If the large duplication property holds, then the pumping lemma [20, Lemma 4] exhibits the required cycle and path. Conversely, if the cycle and path exist, then the pumping lemma [20, Lemma 4] can be used to derive arbitrarily tall trees for which a derivation exists. ◀

Next, we show that if a WTGh G does not satisfy the large duplication property, then its generated weighted tree language $[\![G]\!]$ is regular. To this end, we construct the *linearization* of G. The linearization of a WTGh G is a WTG that simulates all derivations of G which only ensure the equality of subtrees of height at most $\operatorname{ht}(G)$. For this, we replace every production $\ell \xrightarrow{E} q$ in G by the collection of all productions $\ell' \to q$ which can be obtained by instantiating E, i.e., substituting each position constrained by E with a compatible tree of height at most $\operatorname{ht}(G)$ that satisfies E. Note that positions in ℓ that are unconstrained by E are unaffected by these substitutions. Formally, we define the linearization following [15, Definition 7.1].

▶ **Definition 13.** *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a WTGh. The linearization* $\mathrm{lin}(G)$ *of G is the WTG $\mathrm{lin}(G) = (Q \cup \{\bot\}, \Sigma, F, P_{\mathrm{lin}}, \mathrm{wt}_{\mathrm{lin}})$, where P_{lin} and $\mathrm{wt}_{\mathrm{lin}}$ are defined as follows. For $\ell' \in T_\Sigma(Q) \setminus Q$ and $q \in Q$, we let $(\ell' \to q) \in P_{\mathrm{lin}}$ if and only if there exist a production $(\ell \xrightarrow{E} q) \in P$, positions $w_1, \dots, w_k \in \mathrm{pos}_{Q \cup \{\bot\}}(\ell)$, and trees $t_1, \dots, t_k \in T_\Sigma$ with*

- *$\{w_1, \dots, w_k\} = \bigcup_{w \in \mathrm{pos}_\bot(\ell)} [w]_E$; i.e., E constrains exactly the positions $w_1, \dots, w_k$,*
- *$t_i = t_j$ if $(w_i, w_j) \in E$ for all $i, j \in [k]$,*
- *$\ell' = \ell[t_1]_{w_1} \cdots [t_k]_{w_k}$, and*
- *$D_G^{\ell(w_i)}(t_i) \neq \emptyset$ and $\mathrm{ht}(t_i) \leq \mathrm{ht}(G)$ for all $i \in [k]$.*

For every such production $\ell' \to q$ we define $\mathrm{wt}_{\mathrm{lin}}(\ell' \to q)$ as the sum over all weights

$$\mathrm{wt}(\ell \xrightarrow{E} q) \cdot \prod_{i \in [k]} \mathrm{wt}_G^{\ell(w_i)}(t_i)$$

for all $(\ell \xrightarrow{E} q) \in P$, $w_1, \dots, w_k \in \mathrm{pos}_{Q \cup \{\bot\}}(\ell)$, and $t_1, \dots, t_k \in T_\Sigma$ as above.

If a trim WTGh G does not satisfy the large duplication property, then every equality constraint in every derivation of G only ensures the equality of subtrees of height at most $\mathrm{ht}(G)$. Thus, $\mathrm{lin}(G)$ and G generate the same weighted tree language $[\![G]\!] = [\![\mathrm{lin}(G)]\!]$, which is then regular because $\mathrm{lin}(G)$ is a WTG. Thus we summarize:

▶ **Proposition 14.** *Let G be a trim WTGh and suppose that G does not satisfy the large duplication property. Then $[\![G]\!]$ is a regular weighted tree language.*

Finally, we show that if a WTGh $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ satisfies the large duplication property, then $[\![G]\!]$ is not regular. For this, we first show that if G satisfies the large duplication property, then we can decompose it into two WTGh G_1 and G_2 such that $[\![G]\!] = [\![G_1]\!] + [\![G_2]\!]$ and at least one of $[\![G_1]\!]$ and $[\![G_2]\!]$ is not regular. To conclude the desired statement, we then show that the sum $[\![G]\!] = [\![G_1]\!] + [\![G_2]\!]$ is also not regular. For the decomposition, consider the following idea. Assume that there exists a production $p = (\ell \xrightarrow{E} q) \in P$ as in the large duplication property such that $F(q) \neq 0$. Then we create two copies G_1 and G_2 of G as follows. In G_1 we set all final weights to 0, add a new state f with final weight $F(q)$, and add the new production $(\ell \xrightarrow{E} f)$ with the same weight as p. On the other hand, in G_2 we set the final weight of q to 0, add a new state f with final weight $F(q)$, and for every production $p' = (\ell' \xrightarrow{E'} q) \in P$ except p, we add the new production $\ell' \xrightarrow{E'} f$ to G_2 with the same weight as p'. Then $[\![G]\!] = [\![G_1]\!] + [\![G_2]\!]$ because every derivation of G whose last production is p is now a derivation of G_1 to f, and every other derivation is either directly a derivation of G_2 or, in case of other derivations to q, is a derivation of G_2 to f.

By our assumption on the production $p = (\ell \xrightarrow{E} q)$, there exist a tall tree $t \in T_\Sigma$ with $\mathrm{ht}(t) > \mathrm{ht}(G)$ and a constraint $(u, v) \in E$ with $\ell(u) \neq \bot = \ell(v)$ and $D_G^{\ell(u)}(t) \neq \emptyset$. Thus, every tree t' generated by G_1 satisfies $t'|_u = t'|_v$, and by Lemma 8, there exist infinitely many pairwise distinct trees with a derivation to $\ell(u)$. The support (i.e., set of nonzero weighted trees) of $[\![G_1]\!]$ is therefore not a regular tree language. This implies that $[\![G_1]\!]$ is not regular, as the support of every regular weighted tree language over $\mathbb{N}$ is a regular tree language [12].

In general, we cannot expect that a production $\ell \xrightarrow{E} q$ as in the large duplication property exists that already targets a final state. We therefore “grow” productions from the top, beginning with a production whose target state is final, by substituting Q-labeled positions with left-hand sides of other productions until we have “synthesized” a production which satisfies the large duplication property. We then construct G_1 by adding this newly formed production as a production to a new state f. We construct G_2 simply to ensure that it simulates all derivations of G that are not already accounted for by G_1. Formally, we show the following lemma.

▶ **Lemma 15.** *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a trim WTGh that satisfies the large duplication property. Then there exist two trim WTGh $G_1 = (Q_1 \cup \{\bot\}, \Sigma, F_1, P_1, \mathrm{wt}_1)$ and $G_2 = (Q_2 \cup \{\bot\}, \Sigma, F_2, P_2, \mathrm{wt}_2)$ such that $[\![G]\!] = [\![G_1]\!] + [\![G_2]\!]$ and for some $f \in Q_1$ we have*
- *$F_1(f) \neq 0$ and $F_1(q) = 0$ for all $q \in Q_1 \setminus \{f\}$, and*
- *there exists exactly one production $p_\mathrm{f} = (\ell_\mathrm{f} \xrightarrow{E_\mathrm{f}} f) \in P_1$ with target state f, and for this production there exists $(u, v) \in E_\mathrm{f}$ with $\ell_\mathrm{f}(u) \neq \ell_\mathrm{f}(v) = \bot$ and an infinite sequence of pairwise distinct trees $t_0, t_1, t_2, \ldots \in T_\Sigma$ such that $D_{G_1}^{\ell_\mathrm{f}(u)}(t_i) \neq \emptyset$ for all $i \in \mathbb{N}$.*

Proof. Let $p = (\ell \xrightarrow{E} q) \in P$ be a production as in the large duplication property. Since G is trim, there exist a tree $t' \in T_\Sigma$, a final state $q_\mathrm{f} \in Q$ with $F(q_\mathrm{f}) \neq 0$, a derivation $d = (p_1, w_1) \cdots (p_m, w_m) \in D_G^{q_\mathrm{f}}(t')$, and $i \in [m]$ such that $p_i = p$. In other words, there is a derivation utilizing production p. We let $p_j = \ell_j \xrightarrow{E_j} q_j$ for every $j \in [m]$, and let $w_{i_1} > \cdots > w_{i_k}$ be the sequence of prefixes of w_i among the positions $\{w_1, \ldots, w_m\}$ in strictly descending order with respect to the prefix order. In particular, we have $w_{i_1} = w_i$ and $w_{i_k} = \varepsilon$.

For a position w and a set E' of constraints, we define $wE' = \{(wu, wv) \mid (u, v) \in E'\}$. We want to join the left-hand sides of the productions $p_{i_1}, \ldots, p_{i_k}$ to a new production $\ell_{i_k}[\ell_{i_{k-1}}]_{w_{i_{k-1}}} \cdots [\ell_{i_1}]_{w_{i_1}} \xrightarrow{E_\mathrm{f}} q_\mathrm{f}$ with $E_\mathrm{f} = \bigcup_{j \in [k]} w_{i_j} E_{i_j}$. However, we need to ensure that $w_{i_1}, \ldots, w_{i_k}$ do not occur in E_f. Therefore, we assume that p, t', q_f, d, and i above are chosen such that w_i is of minimal length among all possible choices. Then we see as follows that $w_{i_1}, \ldots, w_{i_k}$ do not occur in E_f.

Let $(u, v) \in E$ with $\ell(u) \neq \ell(v) = \bot$ and $t \in T_\Sigma$ with $\mathrm{ht}(t) > \mathrm{ht}(G)$ and $D_G^{\ell(u)}(t) \neq \emptyset$. Suppose there exists $j \in [k]$ such that w_{i_j} occurs in E_f. Then there exists $(u', v') \in E_{i_{j+1}}$ with $w_{i_j} = w_{i_{j+1}} u'$. Then the tree $t'[\![t]\!]_{w_i u}|_{w_{i_j}}$ shows us that $p_{i_{j+1}}$ is also a production as in the large duplication property, but $|w_{i_{j+1}}| < |w_i|$, so w_i is not of minimal length.

We define $G_1 = (Q_1 \cup \{\bot\}, \Sigma, F_1, P_1, \mathrm{wt}_1)$ as follows. Let $f \notin Q \cup \{\bot\}$ be a new state. We set $Q_1 = Q \cup \{f\}$, $F_1(f) = F(q_\mathrm{f})$, and $F_1(q') = 0$ for all $q' \in Q$. For the production $p_\mathrm{f} = (\ell_{i_k}[\ell_{i_{k-1}}]_{w_{i_{k-1}}} \cdots [\ell_{i_1}]_{w_{i_1}} \xrightarrow{E_\mathrm{f}} f)$ with $E_\mathrm{f} = \bigcup_{j \in [k]} w_{i_j} E_{i_j}$, we let $P_1 = P \cup \{p_\mathrm{f}\}$, $\mathrm{wt}_1(p_\mathrm{f}) = \prod_{j \in [k]} \mathrm{wt}(p_{i_j})$, and $\mathrm{wt}_1(p') = \mathrm{wt}(p')$ for all $p' \in P$. Then G_1 simulates all derivations of G with productions $p_{i_1}, \ldots, p_{i_k}$ at the positions $w_{i_1}, \ldots, w_{i_k}$, respectively. For the existence of the infinite sequence of trees, let $(u, v) \in E$ with $\ell(u) \neq \ell(v) = \bot$ and $t \in T_\Sigma$ with $\mathrm{ht}(t) > \mathrm{ht}(G)$ and $D_G^{\ell(u)}(t) \neq \emptyset$. By Lemma 8, there exists an infinite sequence $t_0, t_1, t_2, \ldots \in T_\Sigma$ of pairwise distinct trees with $D_G^{\ell(u)}(t_i) \neq \emptyset$ for all $i \in \mathbb{N}$. Since $D_G^{\ell(u)}(t_i) \subseteq D_{G_1}^{\ell(u)}(t_i)$ for all $i \in \mathbb{N}$, this is the desired sequence. We conclude the definition of G_1 by noting that $(w_i u, w_i v) \in E_\mathrm{f}$ and that the left-hand side ℓ_f of p_f satisfies $\ell_\mathrm{f}(w_i u) = \ell(u)$.

Next, we construct G_2 such that it simulates all remaining derivations of G in the following sense. If d is a derivation of G to a state different from q_f, then it is a derivation of G_2 to that same state. If d is a derivation of G to q_f but its last production is not p_{i_k}, then it is simulated by a derivation of G_2 to a new state f. If d is a derivation of G and its last production is p_{i_k} but the production at $w_{i_{k-1}}$ is not $p_{i_{k-1}}$, then it again is simulated by a derivation of G_2 to f, and so on. To have a more compact definition for G_2, we use the symbol $\Box$ to denote a tree of height 0 and a term $\Box[\ell_{i_k}]_{w_{i_k}} \cdots [\ell_{i_{j+1}}]_{w_{i_{j+1}}}[\ell']_{w_{i_j}}$ for $j = k$ is to be read as $\Box[\ell']_{w_{i_j}}$. We let $f \notin Q \cup \{\bot\}$ be a new state and define $G_2 = (Q_2 \cup \{\bot\}, \Sigma, F_2, P_2, \mathrm{wt}_2)$ by $Q_2 = Q \cup \{f\}$, $F_2(q_\mathrm{f}) = 0$, $F_2(f) = F(q_\mathrm{f})$, and $F_2(q') = F(q')$ for all $q' \in Q \setminus \{q_\mathrm{f}\}$.

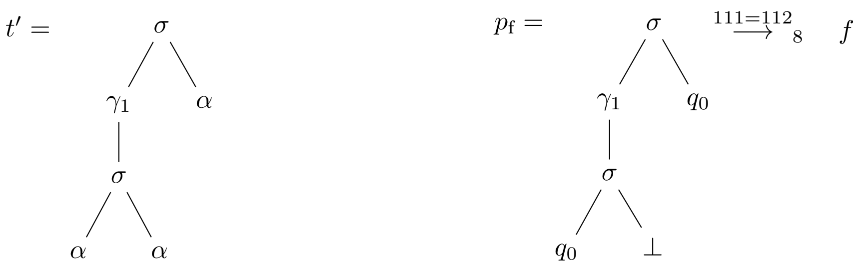


Figure 1 The tree t' and the new production p_f.

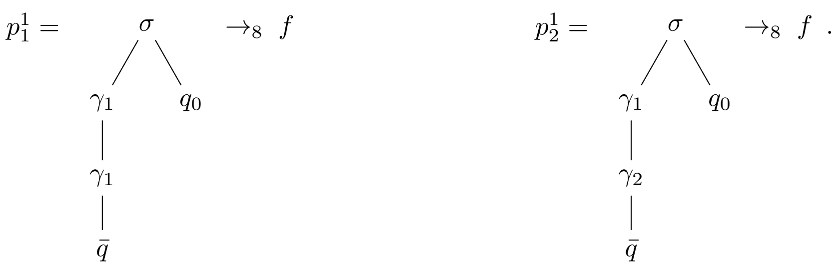
For the set P_2 of productions, we let

$$P_2 = P \cup \bigcup_{j \in [k]} \Big\{ \Box[\ell_{i_k}]_{w_{i_k}} \cdots [\ell_{i_{j+1}}]_{w_{i_{j+1}}} [\ell']_{w_{i_j}} \xrightarrow{E_f} f \;\Big|\; p' = (\ell' \xrightarrow{E'} q_{i_j}) \in P \setminus \{p_{i_j}\},$$
$$E_f = w_{i_j} E' \cup \bigcup_{j'=j+1}^{k} w_{i_{j'}} E_{i_{j'}} \Big\} .$$

For a production $p_f = \Box[\ell_{i_k}]_{w_{i_k}} \cdots [\ell_{i_{j+1}}]_{w_{i_{j+1}}} [\ell']_{w_{i_j}} \xrightarrow{E_f} f$ constructed from p' as above we let $\mathrm{wt}_2(p_f) = \mathrm{wt}(p') \cdot \prod_{j'=j+1}^{k} \mathrm{wt}(p_{i_{j'}})$ and for every $p' \in P$ we let $\mathrm{wt}_2(p') = \mathrm{wt}(p')$. Then we have $[\![G]\!](t) = [\![G_1]\!](t) + [\![G_2]\!](t)$ for every $t \in T_\Sigma$. Note that trimming G_1 and G_2 will not remove any of the newly added productions under the assumption that G is trim. ◀

▶ **Example 16.** We present an example for the decomposition in Lemma 15. Consider the trim WTGh $G = (Q \cup \{\perp\}, \Sigma, P, F, \mathrm{wt})$ with $Q = \{q_0, \bar{q}, q_{\mathrm{f}}\}$, $\Sigma = \{\alpha^{(0)}, \gamma^{(1)}, \sigma^{(2)}, \gamma_1^{(1)}, \gamma_2^{(1)}\}$, final weights $F(q_{\mathrm{f}}) = 1$ and $F(q_0) = F(\bar{q}) = F(\perp) = 0$, and the set $P = P_\perp \cup P'$ defined by $P' = \big\{ \alpha \to_1 q_0,\ \gamma(q_0) \to_1 q_0,\ \sigma(q_0, \perp) \xrightarrow{1=2}_2 \bar{q},\ \gamma_1(\bar{q}) \to_2 \bar{q},\ \gamma_2(\bar{q}) \to_2 \bar{q},\ \sigma(\bar{q}, q_0) \to_2 q_{\mathrm{f}} \big\}$ and the usual productions targeting $\perp$ in $P_\perp$. Trees of the form $\gamma(\cdots(\gamma(\alpha))\cdots)$ of arbitrary height are subject to the constraint $1 = 2$, so G satisfies the large duplication property.

We consider t' as in Figure 1 and use its (unique) derivation in G. Following the approach sketched above, we choose a new state f and define $G_1 = (Q \cup \{f\} \cup \{\perp\}, \Sigma, F_1, P_1, \mathrm{wt}_1)$, where $F_1(f) = 1$ and $F_1(q) = 0$ for every $q \in Q \cup \{\perp\}$, and $P_1 = P \cup \{p_{\mathrm{f}}\}$ with the new production p_{f} depicted in Figure 1, which joins all the productions of G used to derive t', from the one evoking the large duplication property to the one targeting a final state. It remains to construct a WTGh G_2 such that $[\![G]\!] = [\![G_1]\!] + [\![G_2]\!]$. All productions of G still occur in G_2, but q_{f} is not final anymore. Instead, we add a state f with $F_2(f) = F(q_{\mathrm{f}}) = 1$ and make sure that this state adopts all other accepting derivations that formerly led to q_{f}. For this, we handle first the derivations that coincide with the derivation for t' at the juncture positions ε and 1, but not at 2. This leads to the following new productions p_1^1 and p_2^1:



Next we cover the derivations that differ from the derivation for t' at the position 1 but coincide with it at the root. This leads to the new productions

$$p_1^2 = \sigma(\sigma(q_0, \bot), q_0) \xrightarrow{11=12}_4 f \qquad\qquad p_2^2 = \sigma(\gamma_2(q_0), q_0) \to_4 f \ .$$

Apart from the production incorporated at the root of p_f, no other production of G targets q_f directly, so no more productions are added to P_2.

Finally, we define the WTGh $G_2 = (Q \cup \{f\} \cup \{\bot\}, \Sigma, F_2, P_2, \mathrm{wt}_2)$ with $F_2(f) = F(q_\mathrm{f}) = 1$, $F_2(q_\mathrm{f}) = F_2(q_0) = F_2(\bar{q}) = F_2(\bot) = 0$, and $P_2 = P \cup \{p_1^1, p_2^1\} \cup \{p_1^2, p_2^2\}$.

It remains to show that the existence of a decomposition $[\![G]\!] = [\![G_1]\!] + [\![G_2]\!]$ as in Lemma 15 implies the non-regularity of $[\![G]\!]$. For this, we employ the following idea. Consider a ranked alphabet Σ containing a letter σ of rank 2, a WTA $G' = (Q, \Sigma, F, P, \mathrm{wt})$ over Σ (which exemplifies G_2), and a sequence $t_0, t_1, t_2, \ldots \in T_\Sigma$ of pairwise distinct trees. At this point, we assume that P contains all possible productions, but we may have $\mathrm{wt}(p) = 0$ for $p \in P$. Using the initial algebra semantics [12], we can find a matrix representation for the weights assigned by G' to trees of the form $\sigma(t_i, t_j)$ as follows. We enumerate the states $Q = \{q_1, \ldots, q_n\}$ and for every $i \in \mathbb{N}$ define a (column) vector $\nu_i \in \mathbb{N}^n$ by $(\nu_i)_k = \mathrm{wt}_{G'}^{q_k}(t_i)$ for $k \in [n]$. Furthermore, we define a matrix $N \in \mathbb{N}^{n \times n}$ by $N_{kh} = \sum_{q \in Q} F(q) \cdot \mathrm{wt}(\sigma(q_k, q_h) \to q)$ for $k, h \in [n]$. Then $[\![G']\!](\sigma(t_i, t_j)) = \nu_i^\mathrm{T} N \nu_j$ for all $i, j \in \mathbb{N}$, where ν_i^T is the transpose of ν_i.

We employ this matrix representation to show that the sum of $[\![G']\!]$ and the (non-regular) characteristic function 1_L of the tree language $L = \{\sigma(t_i, t_i) \mid i \in \mathbb{N}\}$ is not regular. We proceed by contradiction and assume that $[\![G']\!] + 1_L$ is regular. Thus we can find an analogous matrix representation using a matrix N' and vectors ν_i' for $[\![G']\!] + 1_L$. Since the trees $t_0, t_1, t_2, \ldots$ are pairwise distinct, we can write

$$\big([\![G']\!] + 1_L\big)\big(\sigma(t_i, t_j)\big) = (\nu_i')^\mathrm{T} N' \nu_j' = [\![G']\!]\big(\sigma(t_i, t_j)\big) + \delta_{ij} = \nu_i^\mathrm{T} N \nu_j + \delta_{ij}$$

for all $i, j \in \mathbb{N}$, where δ_{ij} denotes the Kronecker delta. The vectors ν_i' and ν_i contain nonnegative integers, so we may consider the concatenated vectors $\langle \nu_i', \nu_i \rangle$ as vectors of $\mathbb{Q}^m$ where $m \in \mathbb{N}$ is the sum of number of states of G' and of the WTA we assumed recognizes $[\![G']\!] + 1_L$. Since $\mathbb{Q}^m$ is a finite dimensional $\mathbb{Q}$-vector space, the $\mathbb{Q}$-vector space spanned by the family $(\langle \nu_i', \nu_i \rangle)_{i \in \mathbb{N}}$ is also finite dimensional. We may thus select a finite generating set from $(\langle \nu_i', \nu_i \rangle)_{i \in \mathbb{N}}$. For simplicity, we assume that $\langle \nu_1', \nu_1 \rangle, \ldots, \langle \nu_K', \nu_K \rangle$ form such a generating set. Thus there exist $a_1, \ldots, a_K \in \mathbb{Q}$ with $\langle \nu_{K+1}', \nu_{K+1} \rangle = \sum_{i \in [K]} a_i \langle \nu_i', \nu_i \rangle$. Applying the usual distributivity laws for matrix multiplication, we reach a contradiction as follows.

$$\begin{aligned}\big([\![G']\!] + 1_L\big)\big(\sigma(t_{K+1}, t_{K+1})\big) &= (\nu_{K+1}')^\mathrm{T} N' \nu_{K+1}' = \sum_{i \in [K]} a_i (\nu_i')^\mathrm{T} N' \nu_{K+1}' \\ &= \sum_{i \in [K]} a_i \nu_i^\mathrm{T} N \nu_{K+1} = \nu_{K+1}^\mathrm{T} N \nu_{K+1} = [\![G']\!]\big(\sigma(t_{K+1}, t_{K+1})\big)\end{aligned}$$

For the general case, we do not want to assume that $[\![G_2]\!]$ is regular, so we cannot assume to have a matrix representation as we had for $[\![G']\!]$ above. In order to make our idea work, we identify a set of trees for which the behavior of $[\![G_1]\!] + [\![G_2]\!]$ resembles that of $[\![G']\!] + 1_L$; more precisely, we construct a context C and a sequence $t_0, t_1, t_2, \ldots$ of pairwise distinct trees such that $([\![G_1]\!] + [\![G_2]\!])(C(t_i, t_j)) = \nu_i^{(1)} N \nu_j^{(2)} + \delta_{ij}\mu_i$ for all $i, j \in \mathbb{N}$ and additionally, $\mu_i > 0$ for all $i \in \mathbb{N}$. This representation then allows us to perform linear algebra computations in order to prove that $[\![G_1]\!] + [\![G_2]\!]$ is non-regular. Unfortunately, working with a 2-context C may be insufficient if G_1 uses constraints of the form $\{v = v', v' = v''\}$, where more than two positions are constrained to be pairwise equivalent. Therefore, we have to consider more general n-contexts C and then identify a sequence of trees such that the equation above is satisfied on $C(t_i, t_j, t_j, \dots, t_j)$.

Isolating this desired sequence of trees is the most technically involved proof in our paper. We illustrate the effect of this selection in Example 19 below. Along the way, we will use the following version of Ramsey's theorem [24]. For a set X, we denote by $\binom{X}{2}$ the set of all subsets of X of size 2.

▶ **Theorem 17.** *Let $k \geq 1$ be an integer and $f \colon \binom{\mathbb{N}}{2} \to [k]$ a mapping. There exists an infinite subset $E \subseteq \mathbb{N}$ such that $f|_{\binom{E}{2}} \equiv i$ for some $i \in [k]$.*

▶ **Lemma 18.** *Let $G = (Q \cup \{\bot\}, \Sigma, F, P, \mathrm{wt})$ be a trim WTGh. If G satisfies the large duplication property, then there exists an integer $r \geq 2$, an r-context $C \in T_\Sigma(\Box)$, trees $(t_i)_{i\in\mathbb{N}} \subseteq T_\Sigma$, an integer $m \in \mathbb{N}$, row vectors $(\nu_i^{(1)})_{n\in\mathbb{N}} \subseteq \mathbb{N}^m$, column vectors $(\nu_i^{(2)})_{n\in\mathbb{N}} \subseteq \mathbb{N}^m$, a matrix $N \in \mathbb{N}^{m\times m}$, and weights $(\mu_i)_{i\in\mathbb{N}} \subseteq \mathbb{N}\setminus\{0\}$ with $[\![G]\!](C(t_k, t_h, t_h, \dots, t_h)) = \nu_k^{(1)} N \nu_h^{(2)} + \delta_{kh}\mu_k$ for all $k, h \in \mathbb{N}$.*

Proof. By Lemma 15 there exist two trim WTGh $G_1 = (Q_1 \cup \{\bot\}, \Sigma, F_1, P_1, \mathrm{wt}_1)$ and $G_2 = (Q_2 \cup \{\bot\}, \Sigma, F_2, P_2, \mathrm{wt}_2)$ with $[\![G]\!](t) = [\![G_1]\!](t) + [\![G_2]\!](t)$ for all $t \in T_\Sigma$. Additionally, there exists $f \in Q_1$ with $F_1(f) \neq 0$ and $F_1(q) = 0$ for all $q \in Q_1 \setminus \{f\}$ and there exists exactly one production $p_{\mathrm{f}} = (\ell_{\mathrm{f}} \xrightarrow{E_{\mathrm{f}}} f) \in P_1$ whose target state is f. Finally, for this production p_{f} there exists $(u^{(1)}, v^{(1)}) \in E_{\mathrm{f}}$ with $\ell_{\mathrm{f}}(u^{(1)}) \neq \ell_{\mathrm{f}}(v^{(1)}) = \bot$ and an infinite sequence $t_0, t_1, t_2, \ldots \in T_\Sigma$ of pairwise distinct trees with $D_{G_1}^{\ell_{\mathrm{f}}(u^{(1)})}(t_i) \neq \emptyset$ for all $i \in \mathbb{N}$.

Let $t \in T_\Sigma$ be such that $D_{G_1}^{f}(t) \neq \emptyset$, and let $u_1^{(1)}, \dots, u_r^{(1)}$ be an enumeration of all positions that are equality-constrained to $u^{(1)}$ via E_{f}, where we assume that $u_1^{(1)} = u^{(1)}$. We define a context $C = t[\Box]_{u_1^{(1)}} \cdots [\Box]_{u_r^{(1)}}$. Then $[\![G_1]\!](C(t_i, t_j, t_j, \dots, t_j)) > 0$ iff $i = j$.

Let us establish some additional notations. Let $k, h \in \mathbb{N}$ and assume there is $q \in Q_2$ with $F_2(q) \neq 0$ and $d = (p_1, w_1) \cdots (p_m, w_m) \in D_{G_2}^{q}(C(t_k, t_h, t_h, \dots, t_h))$. Let $p_i = \ell_i \xrightarrow{E_i} q_i$ for every $i \in [m]$, and for a set $X \subseteq \mathrm{pos}(C(t_k, t_h, t_h, \dots, t_h))$, we let $i_1 < \cdots < i_n$ be such that $w_{i_1}, \dots, w_{i_n}$ is an enumeration of $\{w_1, \dots, w_m\} \cap X$; i.e., all positions in X to which d applies productions. We set $d|_X = (p_{i_1}, w_{i_1}) \cdots (p_{i_n}, w_{i_n})$, $\mathrm{wt}_2(d|_X) = \prod_{j\in[n]} \mathrm{wt}_2(p_{i_j})$, and $D_{kh} = \{d'|_{\mathrm{pos}(C)} \mid \exists q' \in Q_2 \colon F_2(q') \neq 0,\, d' \in D_{G_2}^{q'}(C(t_k, t_h, t_h, \dots, t_h))\}$.

We now employ Ramsey's theorem in the following way. For $k, h \in \mathbb{N}$ with $k < h$, we consider the mapping $\{k, h\} \mapsto D_{kh}$. This mapping has a finite range as every D_{kh} is a set of finite words over the alphabet $P_2 \times \mathrm{pos}(C)$ of length at most $\mathrm{size}(C)$. Thus, by Ramsey's theorem, we obtain a subsequence $(t_{i_j})_{j\in\mathbb{N}}$ with $D_{i_k i_h} = D_<$ for all $k, h \in \mathbb{N}$ and some set $D_<$. For simplicity, we assume $D_{kh} = D_<$ for all $k, h \in \mathbb{N}$ with $k < h$. Similarly, we select a further subsequence and assume $D_{kh} = D_>$ for all $k, h \in \mathbb{N}$ with $k > h$. Finally, the mapping $k \mapsto D_{kk}$ also has a finite range, so by the pigeonhole principle, we may select a further subsequence and assume that $D_{kk} = D_=$ for all $k \in \mathbb{N}$ and some set $D_=$. In the following, we show that $D_< = D_> \subseteq D_=$.

For now, we assume $D_< \neq \emptyset$, let $(p_1, w_1) \cdots (p_m, w_m) \in D_<$, and let $p_i = \ell_i \xrightarrow{E_i} q_i$ for every $i \in [m]$. Also, we define $C_{kh} = C(t_k, t_h, t_h, \dots, t_h)$, $C_{k\Box} = C(t_k, \Box, \Box, \dots, \Box)$, and $C_{\Box h} = C(\Box, t_h, t_h, \dots, t_h)$ for $k, h \in \mathbb{N}$. We show that every constraint from every E_i is satisfied on all C_{kh} with $k, h \geq 1$, not just for $k < h$. More precisely, let $i \in [m]$, $(u', v') \in E_i$, and $(u,v) = (w_i u', w_i v')$. We show $C_{kh}|_u = C_{kh}|_v$ for all $k,h \geq 1$. Note that by assumption, $C_{kh}|_u = C_{kh}|_v$ is true for all $k,h \in \mathbb{N}$ with $k < h$. We show our statement by a case distinction depending on the position of u and v in relation to the positions $u_1^{(1)}, \dots, u_r^{(1)}$.

1. If both u and v are parallel to $u^{(1)}{}_1$, then $C_{ij}|_u$ and $C_{ij}|_v$ do not depend on i. Thus, $C_{0j}|_u = C_{0j}|_v$ for all $j \geq 1$ implies the statement.
2. If u is in prefix-relation with $u^{(1)}{}_1$ and v is parallel to $u^{(1)}{}_1$, then $C_{ij}|_v$ does not depend on i. If $u \leq u^{(1)}{}_1$, then by our assumption that $(t_i)_{i \in \mathbb{N}}$ are pairwise distinct, we obtain the contradiction $C_{02}|_v = C_{02}|_u \neq C_{12}|_u = C_{12}|_v$, where $C_{02}|_v = C_{12}|_v$ should hold. Thus, we have $u^{(1)}{}_1 \leq u$ and in particular, $C_{ij}|_u$ does not depend on j. Thus, for all $i,j \geq 1$ we obtain $C_{ij}|_u = C_{i,i+1}|_u = C_{i,i+1}|_v = C_{0,i+1}|_v = C_{0,i+1}|_u = C_{0j}|_u = C_{0j}|_v = C_{ij}|_v$. If v is in prefix-relation with $u^{(1)}{}_1$ and u is parallel to $u^{(1)}{}_1$, then we come to the same conclusion by formally exchanging u and v in this argumentation.
3. If u and v are both in prefix-relation with $u^{(1)}{}_1$, then u and v being parallel to each other implies $u^{(1)}{}_1 \leq u$ and $u^{(1)}{}_1 \leq v$. In particular, both u and v are parallel to all $u_2^{(1)}, \dots, u_r^{(1)}$. Thus, we obtain, as in the first case, that $C_{ij}|_u$ and $C_{ij}|_v$ do not depend on j and the statement follows from $C_{i,i+1}|_u = C_{i,i+1}|_v$ for all $i \in \mathbb{N}$.

Let $k,h \geq 1$ and $d_C \in D_<$, and let $q \in Q_2$, $d_{k,k+1} \in D^q_{G_2}(C_{k,k+1})$, and $d_{h-1,h} \in D^q_{G_2}(C_{h-1,h})$ such that $d_C = d_{k,k+1}|_{\mathrm{pos}(C)} = d_{h-1,h}|_{\mathrm{pos}(C)}$. Then for $d_k = d_{k,k+1}|_{\mathrm{pos}(C_{k,k+1}) \setminus \mathrm{pos}(C_{\Box,k+1})}$ and $d_h = d_{h-1,h}|_{\mathrm{pos}(C_{h-1,h}) \setminus \mathrm{pos}(C_{h-1,\Box})}$, we can reorder $d = d_k d_h d_C$ to a complete left-most derivation of G_2 for C_{kh}, as all equality constraints from d_k are satisfied by the assumption on $d_{k,k+1}$, all equality constraints from d_h are satisfied by the assumption on $d_{h-1,h}$, and all equality constraints from d_C are satisfied by our case distinction. Considering the special cases $k=2$, $h=1$, and $k=h=1$, and the definitions of $D_>$ and $D_=$, we obtain $d_C \in D_{21} = D_>$ and $d_C \in D_{11} = D_=$, and hence, $D_< \subseteq D_>$ and $D_< \subseteq D_=$.

The converse inclusion $D_> \subseteq D_<$ follows with an analogous reasoning. In conclusion, we obtain $D_< = D_> \subseteq D_=$. By the reasoning above, the case $D_< = \emptyset$ we excluded earlier is only possible if also $D_> = \emptyset$, in which case we again have $D_< = D_> \subseteq D_=$.

Let $d_1, \dots, d_n$ be an enumeration of $D_<$, $i \in [n]$, and $k \in \mathbb{N}$. We define the sets

$$D^{(1)}_{i,k} = \big\{ d|_{\mathrm{pos}(C_{k,k+1}) \setminus \mathrm{pos}(C_{\Box,k+1})} \mid d \in D^q_{G_2}(C_{k,k+1}),\ d_i = d|_{\mathrm{pos}(C)},\ q \in Q_2 \big\}$$
$$D^{(2)}_{i,k} = \big\{ d|_{\mathrm{pos}(C_{k+1,k}) \setminus \mathrm{pos}(C_{k+1,\Box})} \mid d \in D^q_{G_2}(C_{k+1,k}),\ d_i = d|_{\mathrm{pos}(C)},\ q \in Q_2 \big\}$$

and the corresponding weights $\nu^{(1)}_{i,k} = \sum_{d \in D^{(1)}_{i,k}} \mathrm{wt}_2(d)$ and $\nu^{(2)}_{i,k} = \sum_{d \in D^{(2)}_{i,k}} \mathrm{wt}_2(d)$. Let q_i be the target state of the last production in d_i and define $\nu_i = F_2(q_i) \cdot \mathrm{wt}_2(d_i)$. Then for all $k,h \in \mathbb{N}$ we have $[\![G_2]\!](C_{kh}) = \sum_{i \in [n]} (\nu^{(1)}_{i,k} \cdot \nu_i \cdot \nu^{(2)}_{i,h}) + \delta_{kh}\mu'_k$ for nonnegative $(\mu'_j)_{j \in \mathbb{N}}$, which stem from the fact that potentially $D_= \setminus D_< \neq \emptyset$. We arrange the weights $\nu^{(1)}_{i,k}$ into a row vector $\nu^{(1)}_k$, and the weights $\nu^{(2)}_{i,h}$ into a column vector $\nu^{(2)}_h$, and the weights ν_i into a diagonal matrix N such that $[\![G_2]\!](C_{kh}) = \nu^{(1)}_k N \nu^{(2)}_h + \delta_{kh}\mu'_k$. Finally, recall that $[\![G_1]\!](C_{kh}) > 0$ iff $k = h$ for all $k,h \in \mathbb{N}$. Thus we set $\mu_k = \mu'_k + [\![G_1]\!](C_{kk})$ and obtain $[\![G]\!](C_{kh}) = [\![G_2]\!](C_{kh}) + [\![G_1]\!](C_{kh}) = \nu^{(1)}_k N \nu^{(2)}_h + \delta_{kh}\mu_k$ with $\mu_k > 0$ for all $k,h \in \mathbb{N}$. ◀

Before concluding the correctness of our decision procedure for the weighted HOM-problem, we want to exemplify how the Lemma 12 acts on a simple weighted tree language.

▶ **Example 19.** Consider the WTGh $G = (\{q, q_f, \bot\}, \{a^{(0)}, g^{(1)}, f^{(2)}\}, F, P, \mathrm{wt})$ with final weights $F(q_f) = 1$, $F(q) = F(\bot) = 0$ and the following productions:

$$P = \big\{\ a \to_1 q,\ \ g(q) \to_2 q,\ \ f(q,\bot) \xrightarrow{1=2}_1 q_f,\ \ f\big(q, g(\bot)\big) \xrightarrow{1=21}_1 q_f\ \big\}\ \cup\ P_\bot$$

where $P_\perp = \{a \to_1 \perp,\ g(\perp) \to_1 \perp,\ f(\perp,\perp) \to_1 \perp\}$. The production $f(q,\perp) \xrightarrow{1=2}_1 q_f$ and the tree $g^{\mathrm{ht}(G)}(a)$ satisfy the conditions in the large duplication property, so let G_1 denote the WTGh constructed according to Lemma 15 which simulates all derivations of G that use this production at ε. Consider the sequence $t_i = g^{i+\mathrm{ht}(G)}(a)$ for $i \in \mathbb{N}$. The context $C = f(\square,\square)$ satisfies $[\![G_1]\!]\big(C(t_i,t_j)\big) \neq 0$ iff $i = j$. In order to reproduce the linear-algebra argument from the special case of $[\![G']\!] + 1_L$ described above, we need a matrix representation for the remaining part $[\![G_2]\!]$, possibly with an additional factor δ_{ij}. In terms of the weights computed by G_2, we can achieve this by the condition that $[\![G_2]\!]\big(C(t_i,t_j)\big) \neq 0$ either for all $i,j \in \mathbb{N}$, or for none, or only if $i = j$. However, because of the production $f\big(q, g(\perp)\big) \xrightarrow{1=21}_1 q_f$, for each i we have $[\![G_2]\!]\big(C(t_i,t_{i+1})\big) \neq 0$ and $[\![G_2]\!]\big(C(t_i,t_j)\big) = 0$ for all $j \neq i+1$. To fix this issue, we may select the subsequence $(t_{2i})_{i\in\mathbb{N}}$: In that case, we have $[\![G_2]\!]\big(C(t_{2i},t_{2j})\big) = 0$ for all $i,j \in \mathbb{N}$, and the matrix representation for $[\![G_2]\!]$ is trivial.

Let us now conclude the decidability of the $\mathbb{N}$-weighted HOM-problem.

▶ **Theorem 20.** *Let $G = (Q \cup \{\perp\}, \Sigma, F, P, \mathrm{wt})$ be a trim WTGh. If G satisfies the large duplication property, then $[\![G]\!]$ is not regular.*

Proof. Let $C \in T_\Sigma(\square)$, $(t_i)_{i\in\mathbb{N}} \subseteq T_\Sigma$, $m \in \mathbb{N}$, $(\nu_i^{(1)})_{n\in\mathbb{N}}, (\nu_i^{(2)})_{n\in\mathbb{N}} \subseteq \mathbb{N}^m$, $N \in \mathbb{N}^{m\times m}$, and $(\mu_i)_{i\in\mathbb{N}} \subseteq \mathbb{N}\setminus\{0\}$ be as in Lemma 18, i.e., $[\![G]\!](C(t_k,t_h,t_h,\dots,t_h)) = \nu_k^{(1)} N \nu_h^{(2)} + \delta_{kh}\mu_k$ for all $k,h \in \mathbb{N}$. If $[\![G]\!]$ is regular, then we can assume a representation for all $k,h \in \mathbb{N}$ as $[\![G]\!](C(t_k,t_h,t_h,\dots,t_h)) = g(\kappa_k,\kappa_h,\kappa_h,\dots,\kappa_h)$, where κ_h is a finite vector of weights over $\mathbb{N}$ where each entry corresponds to the sum of all derivations for t_h to a specific state of a WTA, and g is a multilinear map encoding the weights of the derivations for $C(\square,\square,\dots,\square)$ depending on the specific input states at the $\square$-nodes and the target state at the root ε. We choose K such that the concatenated vectors $\langle\kappa_1,\nu_1^{(1)}\rangle,\dots,\langle\kappa_K,\nu_K^{(1)}\rangle$ form a generating set of the $\mathbb{Q}$-vector space spanned by $(\langle\kappa_i,\nu_i^{(1)}\rangle)_{i\in\mathbb{N}}$. Then there are coefficients $a_1,\dots,a_K \in \mathbb{Q}$ with $\kappa_{K+1} = \sum_{i\in[K]} a_i\kappa_i$ and $\nu_{K+1}^{(1)} = \sum_{i\in[K]} a_i\nu_i^{(1)}$. Thus, we reach our contradiction by

$$\begin{aligned}\nu_{K+1}^{(1)} N \nu_{K+1}^{(2)} + \mu_{K+1} = g(\kappa_{K+1},\kappa_{K+1},\dots,\kappa_{K+1}) &= \sum_{i\in[K]} a_i g(\kappa_i,\kappa_{K+1},\dots,\kappa_{K+1}) \\ &= \sum_{i\in[K]} a_i\nu_i^{(1)} N \nu_{K+1}^{(2)} = \nu_{K+1}^{(1)} N \nu_{K+1}^{(2)}.\end{aligned}$$ ◀

5 Conclusion

In this contribution, we proved that the $\mathbb{N}$-weighted HOM-problem is decidable. Formally, given a regular weighted tree language A over $\mathbb{N}$ and a nondeleting, nonerasing tree homomorphism h as input, it is decidable in polynomial time whether the homomorphic image h_A is again regular. This was achieved by reducing the HOM-problem to the newly introduced large duplication property, which formalizes the non-regular behavior of the investigated weighted tree language h_A, and then showing that this property is decidable.

Initially, h_A is represented by a generalized tree grammar (WTGh) as introduced in [20]. Such a device expresses the duplication of subtrees performed by h by means of explicit equality constraints. This WTGh is trimmed and tested directly for the large duplication property. If it does not satisfy this property, we construct an equivalent weighted tree grammar without constraints, which proves regularity of the generated weighted tree language. However, if the trim WTGh for h_A does satisfy the large duplication property, then no equivalent weighted tree grammar exists. To prove this, we first identify a special sequence of productions, isolate it from the remainder of the WTGc, and then prove that it induces a non-regularity which cannot be compensated by the remaining derivations of the WTGh.

We require h to be nondeleting and nonerasing simply to ensure that h_A is well-defined in general. These properties have no impact on the correctness of the reduction or the computational complexity of the large duplication property, to which we reduce the $\mathbb{N}$-weighted HOM-problem. Indeed, our decision procedure for this problem is polynomial, while the unweighted HOM-problem is EXPTIME-complete [6]. In the $\mathbb{N}$-weighted setting we proved that the large duplication property is sufficient for non-regularity; this is the main technical difficulty and utilizes RAMSEY's theorem to identify a sequence of trees that acts as a witness for the non-regularity of the homomorphic image. A matrix representation that resembles the initial algebra semantics is then utilized to prove non-regularity. In the unweighted case the large duplication property is clearly necessary, but not sufficient. This difference is caused by the different algebraic structures of the underlying semirings. Whereas the semiring $\mathbb{N}$ embeds into a field, the BOOLEAN semiring is idempotent, which can be used to cover non-regular behavior with regular behavior making it irrelevant. Essentially we proved that such covers are impossible in $\mathbb{N}$, which simplifies the execution of the decision procedure and allows us to prove polynomial-time decidability of the $\mathbb{N}$-weighted HOM-problem.

References

1 Adel Bouhoula and Florent Jacquemard. Tree automata, implicit induction and explicit destructors for security protocol verification. Technical report, Research Report LSV-07-10, 2007.

2 Adel Bouhoula and Florent Jacquemard. Automated induction with constrained tree automata. In *International Joint Conference on Automated Reasoning*, pages 539–554. Springer, 2008.

3 Symeon Bozapalidis and Antonios Kalampakas. Graph automata. *Theoret. Comput. Sci.*, 393(1-3):147–165, 2008. doi:10.1016/j.tcs.2007.11.022.

4 Symeon Bozapalidis and George Rahonis. On the closure of recognizable tree series under tree homomorphisms. *J. Autom. Lang. Comb.*, 10(2–3):185–202, 2005. doi:10.25596/jalc-2005-185.

5 H. Comon, M. Dauchet, R. Gilleron, C. Löding, F. Jacquemard, D. Lugiez, S. Tison, and M. Tommasi. Tree Automata — Techniques and Applications. https://jacquema.gitlabpages.inria.fr/files/tata.pdf, 2007.

6 Carles Creus, Adrià Gascón, Guillem Godoy, and Lander Ramos. The HOM problem is EXPTIME-complete. *SIAM J. Comput.*, 45(4):1230–1260, 2016. doi:10.1137/140999104.

7 John Doner. Tree acceptors and some of their applications. *J. Comput. System Sci.*, 4(5):406–451, 1970. doi:10.1016/S0022-0000(70)80041-1.

8 Frank Drewes. *Grammatical picture generation: A tree-based approach.* Springer, 2006. doi:10.1007/3-540-32507-7.

9 Zoltán Ésik and Werner Kuich. Formal tree series. *J. Autom. Lang. Comb.*, 8(2):219–285, 2003. doi:10.25596/jalc-2003-219.

10 Zoltán Fülöp, Andreas Maletti, and Heiko Vogler. Preservation of recognizability for synchronous tree substitution grammars. In *Proc. Workshop Applications of Tree Automata in Natural Language Processing*, pages 1–9. ACL, 2010. URL: https://aclanthology.org/W10-2501.

11 Zoltán Fülöp, Andreas Maletti, and Heiko Vogler. Weighted extended tree transducers. *Fundam. Inform.*, 111(2):163–202, 2011. doi:10.3233/FI-2011-559.

12 Zoltán Fülöp and Heiko Vogler. Weighted tree automata and tree transducers. In *Handbook of Weighted Automata*, chapter 9, pages 313–403. Springer, 2009. doi:10.1007/978-3-642-01492-5_9.

13 Ferenc Gécseg and Magnus Steinby. Tree automata. Technical Report 1509.06233, arXiv, 2015. URL: `https://arxiv.org/pdf/1509.06233.pdf`.

14 Rémy Gilleron and Sophie Tison. Regular tree languages and rewrite systems. *Fundamenta informaticae*, 24(1-2):157–175, 1995.

15 Guillem Godoy and Omer Giménez. The HOM problem is decidable. *J. ACM*, 60(4):1–44, 2013. `doi:10.1145/2508028.2501600`.

16 Guillem Godoy, Omer Giménez, Lander Ramos, and Carme Àlvarez. The HOM problem is decidable. In *Proc. 42nd ACM Symp. Theory of Computing*, pages 485–494. ACM, 2010.

17 Jonathan S. Golan. *Semirings and their Applications*. Kluwer Academic, Dordrecht, 1999. `doi:10.1007/978-94-015-9333-5`.

18 Udo Hebisch and Hanns J. Weinert. *Semirings — Algebraic Theory and Applications in Computer Science*. World Scientific, 1998. `doi:10.1142/3903`.

19 Dan Jurafsky and James H. Martin. *Speech and language processing*. Prentice Hall, 3rd edition, 2023. URL: `https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf`.

20 Andreas Maletti and Andreea-Teodora Nász. Weighted tree automata with constraints. *Theory Comput. Syst.*, 2023. to appear. URL: `https://arxiv.org/pdf/2302.03434.pdf`.

21 J. Mongy-Steen. *Transformation de noyaux reconnaissables d'arbres. Forêts RATEG*. PhD thesis, Université de Lille, 1981.

22 Andreea-Teodora Nász. Solving the weighted HOM-problem with the help of unambiguity. In *Proc. 16th Int. Conf. Automata and Formal Languages*, volume 386 of *EPTCS*, pages 200–214. Open Publishing Association, 2023. `doi:10.4204/EPTCS.386.16`.

23 Dominique Perrin. Recent results on automata and infinite words. In *Proc. 11th Int. Symp. Mathematical Foundations of Computer Science*, volume 176 of *LNCS*, pages 134–148. Springer, 1984. `doi:10.1007/BFb0030294`.

24 F. P. Ramsey. On a problem of formal logic. *Proc. London Math. Soc*, 30, 1930. `doi:10.1112/plms/s2-30.1.264`.

25 Arto Salomaa and Matti Soittola. *Automata-theoretic aspects of formal power series*. Springer, 1978. `doi:10.1007/978-1-4612-6264-0`.

26 Marcel Paul Schützenberger. On the definition of a family of automata. *Inform. and Control*, 4(2–3):245–270, 1961. `doi:10.1016/S0019-9958(61)80020-X`.

27 James W. Thatcher. Characterizing derivation trees of context-free grammars through a generalization of finite automata theory. *J. Comput. Syst. Sci.*, 1(4):317–322, 1967. `doi:10.1016/S0022-0000(67)80022-9`.

28 James W. Thatcher and Jesse B. Wright. Generalized finite automata theory with an application to a decision problem of second-order logic. *Math. Systems Theory*, 2(1):57–81, 1968. `doi:10.1007/BF01691346`.

29 Reinhard Wilhelm, Helmut Seidl, and Sebastian Hack. *Compiler Design*. Springer, 2013. `doi:10.1007/978-3-642-17540-4`.